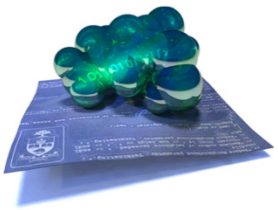
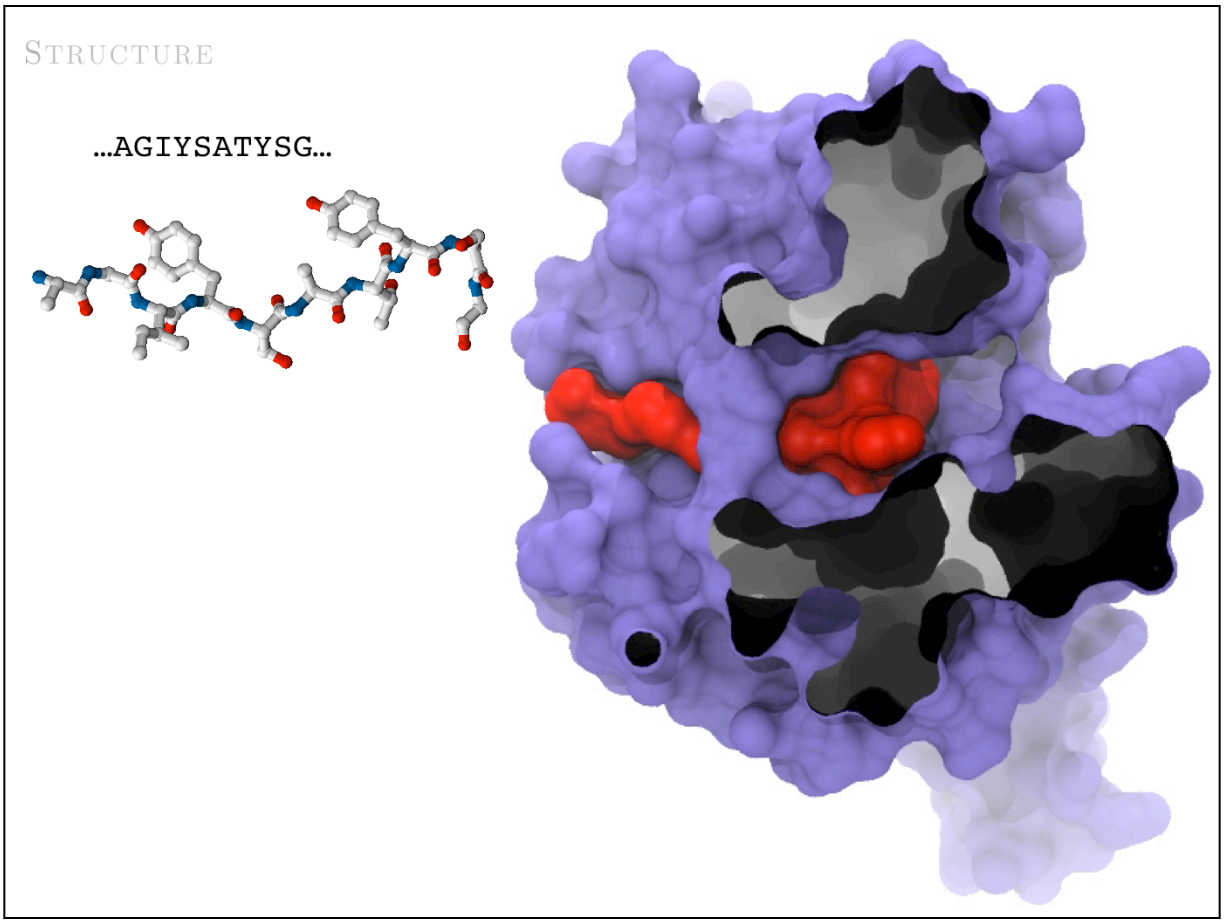
A
Bioinformatics
Course

Protein Structure Concepts



Boris Steipe

department of biochemistry – department of molecular genetics
University of Toronto

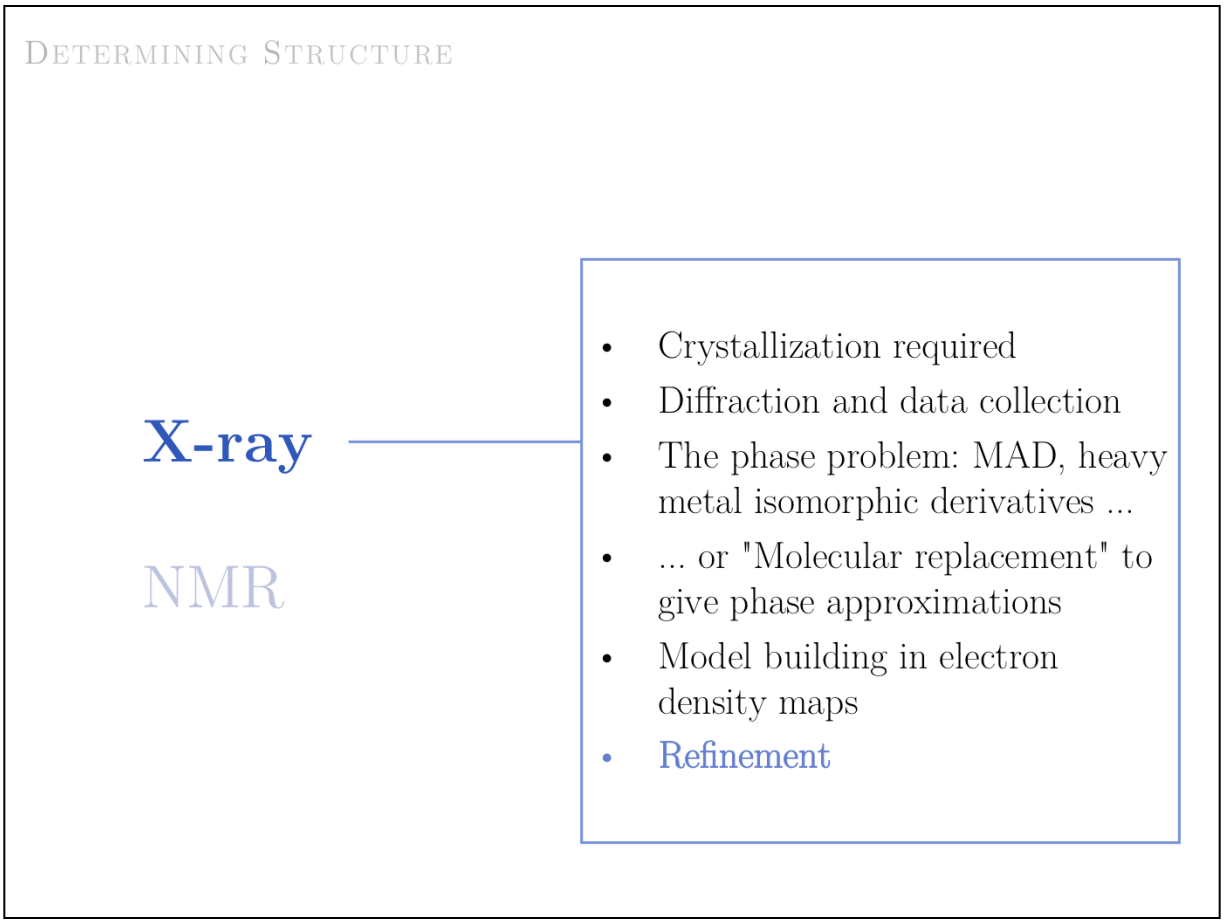


Structure contextualizes sequence.

A sequence provides a description of the molecule, but the *role* of the individual amino acids can only be understood in the context of their three-dimensional context. Such roles can include providing and positioning functional groups in space, providing the building blocks for folding, or generally contributing to the shape of the protein to be complementary to its interaction partners. Molecular function ultimately can be understood in terms of "shape", it relies on the precise "fit" of interacting molecules – their structural complementarity. It is only with reference to the three-dimensional arrangement of atoms (and their dynamics) that mechanistic explanations of biomolecular function becomes possible.

SECTION

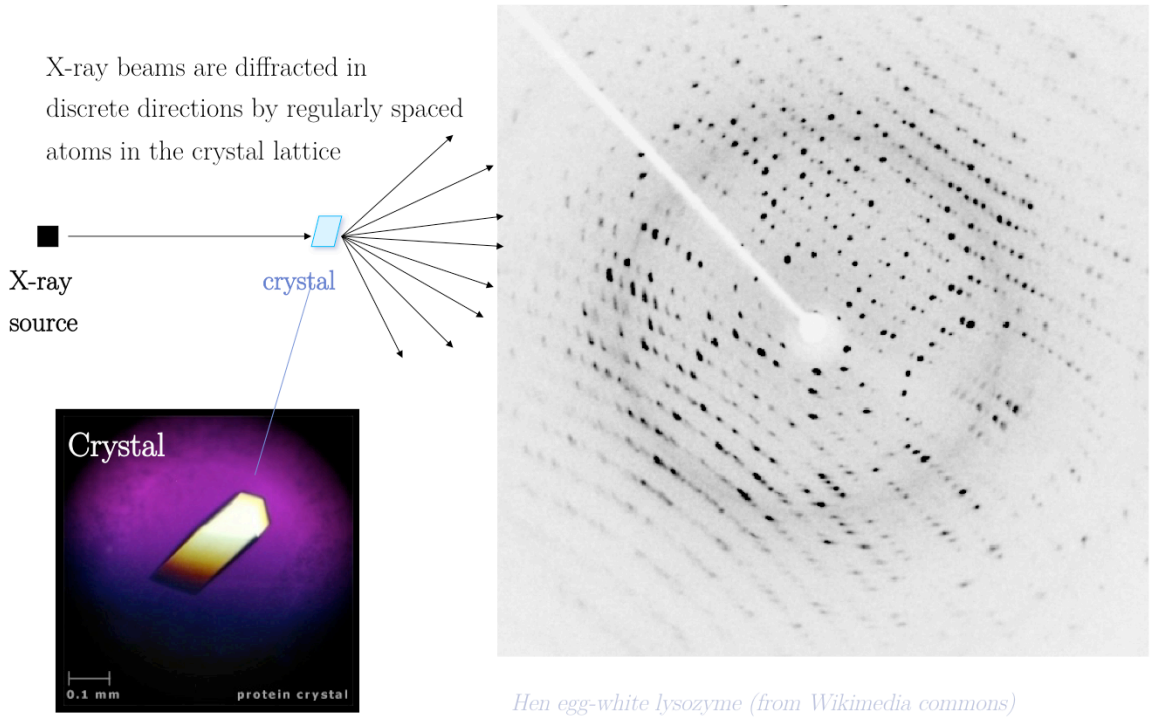
Structure determination



The method of choice for protein structure determination is X-ray crystallography. However, the method requires to crystallize proteins, essentially a controlled process of precipitation from concentrated solution by slow addition of substances that reduce the protein's solubility, and not all proteins can be crystallized.

cf. https://en.wikipedia.org/wiki/X-ray_crystallography

Context



Hen egg-white lysozyme (from Wikimedia commons)

X-ray crystallography is still the method of choice for precise, detailed structures, or structures of complexes that are very large. Collecting X-ray data and "solving" the structure from such data is a process that has become robust, and acessible even for novices in moderately equipped laboratories.

The process involves generating well-ordered crystals of a protein, illuminating them with a small, well collimated, monochromatic X-ray beam, and measuring the diffraction pattern. Diffraction spots are indexed and their positions and intensities can be used to infer the shape of the electron density that reflected the X-rays in the crystal lattice.

However the data is not complete: in order to reconstruct the density exactly, one needs to know not only the intensity and position of each spot, but also the phase of the wave that was diffracted. This phase information can not be measured directly since (*i*) the measuring devices are much thicker than the 1.54Å wavelength of the (typical) X-ray beam, and (*ii*) microheterogeneity of the crystal causes loss of accurate phase information in the first place.

X-ray

The phase problem ...

molecular replacement

direct phasing

MIR phases

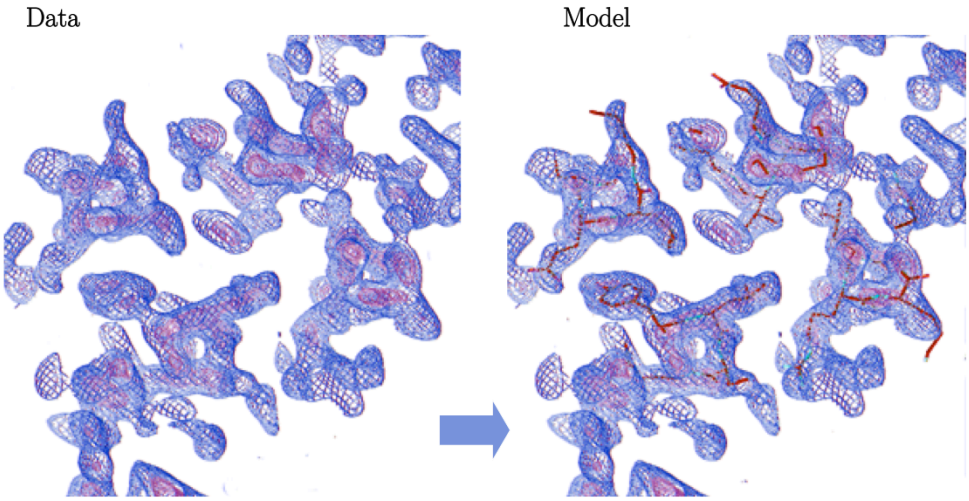
MAD phases

NMR

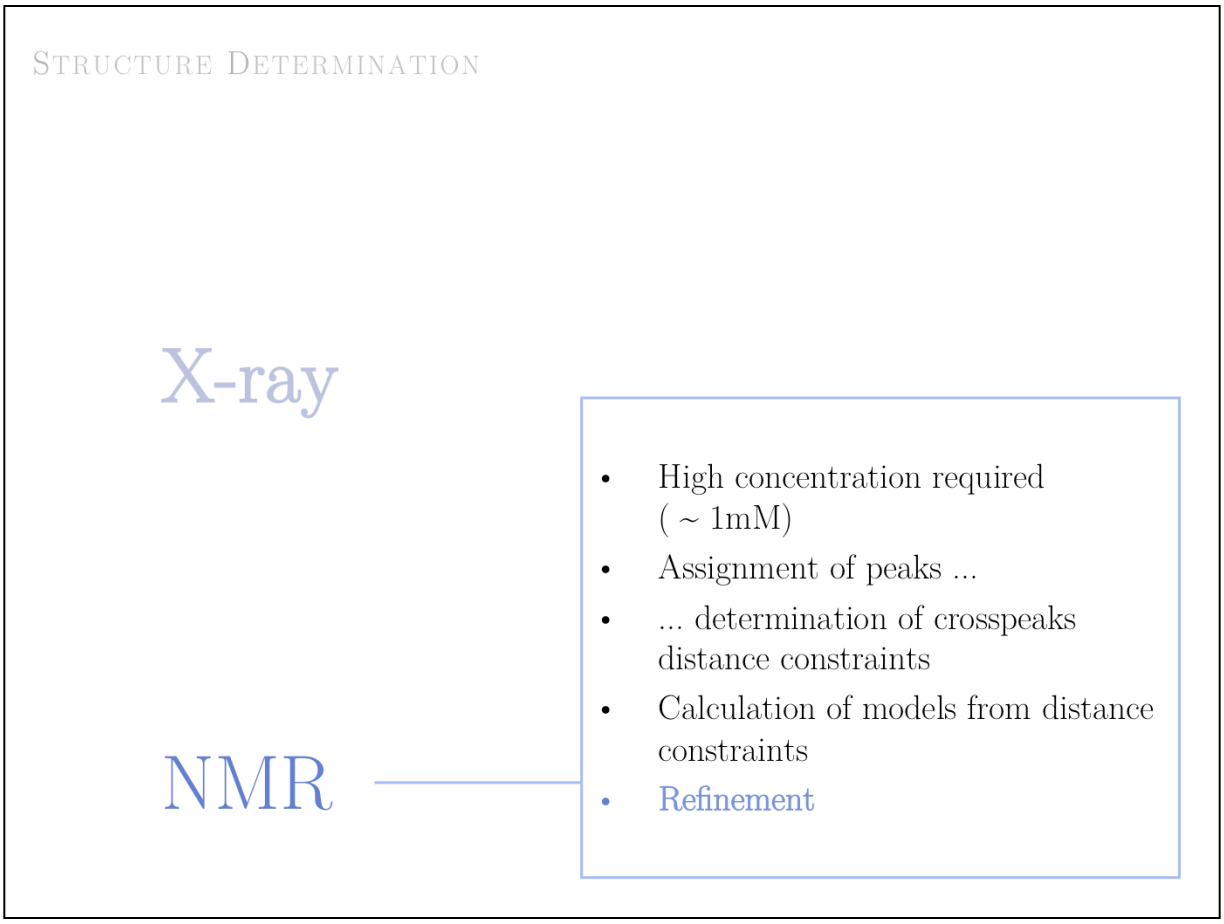
The inability to measure the **'phases** of diffracted photons prevents the reconstruction of the diffracting objects from one set of experimental measurements alone. Additional information must be brought into the process. Usually this makes use of the fact that photons that are **in phase** enhance the measured intensities, whereas photons that are **phase-shifted** by 180° cancel each other's intensities. Thus measuring intensity changes due to pahse-shifts caused by new diffraction centres that are placed into the crystall lattice, allows us to infer relative phases. If several relative phases are known, we can triangulate their absolute values. Experimental error makes this a difficult problem, but under favourable circumstances the electron density map will be interpretable; a structural model can then be built and refined.

Diffraction is not imaging!

Model building and model refinement are required.



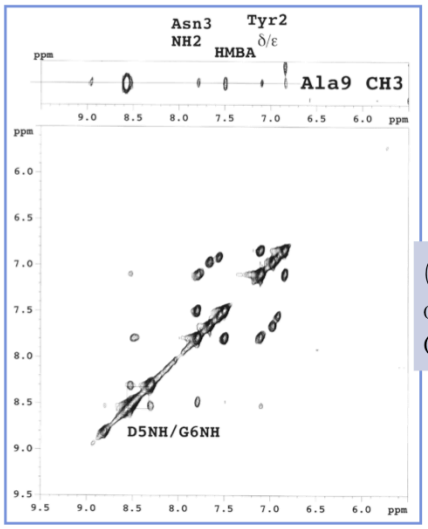
See e.g. Ilari and Savino (2008) Protein structure determination by x-ray crystallography. Methods Mol Biol. 452:63-87. (PMID 18563369)



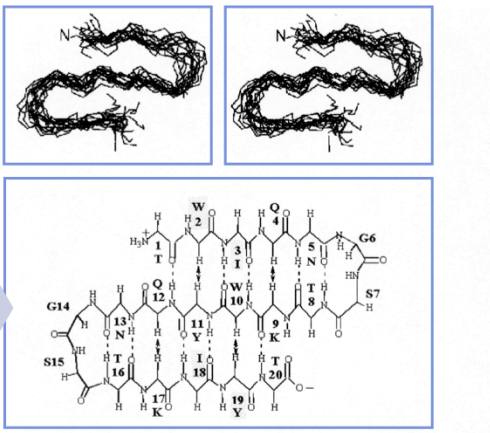
NMR spectroscopy is an important alternative to x-ray crystallographic determination of protein structure.

cf. https://en.wikipedia.org/wiki/Nuclear_magnetic_resonance_spectroscopy_of_proteins

NMR: Distance Constraints



(NOE)
distance constraints
Chem. Shifts, J_{coupl}



(Top) Stereoscopic view of the three-stranded antiparallel β–sheet formed by a 20-residue peptide designed in the lab, the sequence of which is shown at the bottom. The framed positions are the ones selected to be mutated (polar and apolar residues) to form a small combinatorial library in collaboration with participant 6 (Lipotec).

2D proton NMR spectrum of β–hairpin

Atomic model of β–sheet from NMR data

NMR structure determination is based on distance estimates between protons.

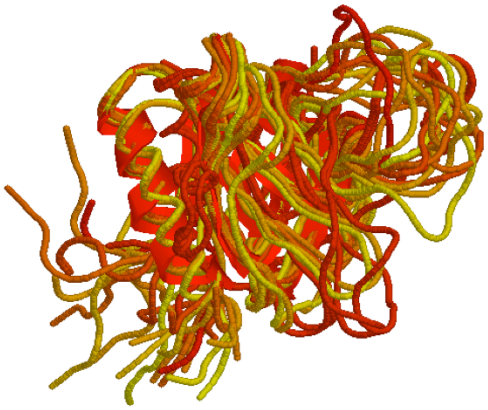
Specific protons are made to precess through applying microwave pulses in strong magnetic fields. When this excitation is stopped, the precessing protons **emit** microwaves.

Which precessing proton resonates with a particular frequency can be determined ("peak assignment"). When the proton is excited at this frequency, its spin-polarization can be transferred to other protons through space via the so-called Nuclear Overhauser Effect[1].

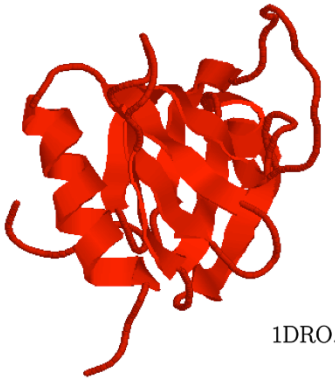
This effect is highly sensitive to spatial separation, therefore it can generate a list of distance constraints between specific protons.

(1) see e.g. http://en.wikipedia.org/wiki/Nuclear_Overhauser_effect

NMR Ensemble



Ensemble of structures that are compatible with experimental distance constraints



1DRO.PDB

Consensus model

Problems and Issues:

- Concentration/Solubility
- Assignment and NOEs
- Refinement

Distance constraints are translated into "pseudo energies" that are used in molecular dynamics simulations to generate structural models. The simulation starts from a random conformation and then runs until the energy of the model is minimzed. Violations of stereochemistry (bond-lengths, angles and steric clashes) are minimized together with violations of experimental distance constraints. This typically generates an **ensemble** of conformational models which is then averaged and further refined to finally arrive at a final consensus model.

The problem with the procedure is that the relationship between average- and final structure is heavily dominated by the force-field that is used for refinement. Thus high-resolution details are unlikely to be obtained.

A great advantage of the method is that it can in principle be used to obtain dynamic information, e.g. distinguish between static disorder and mobility of a part of structure.

cf. Rosato A, Tejero R, and Montelione GT. (2013) Quality assessment of protein NMR structures. Curr Opin Struct Biol. 23(5):715-724. (PMID 24060334)

Other methods:

Modeling
- "Homology models" (see later in the course)
- *ab initio* models

Electron diffraction

Neutron diffraction

} Similar to X-ray crystallography

(Single molecule diffraction)

There are more methods to determine molecular 3D coordinates, but NMR and X-ray crystallography are by far the most prevalent ones.

However: the next game-changer is on the horizon – single-molecule diffraction with X-ray lasers.

cf. Schlichting I. and Miao J. (2012) Emerging opportunities in structural biology with X-ray free-electron lasers. Curr Opin Struct Biol. 22(5):613-626. (PMID 22922042)

outline

SECTION

Abstraction and storage

Structure Concepts

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

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

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

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

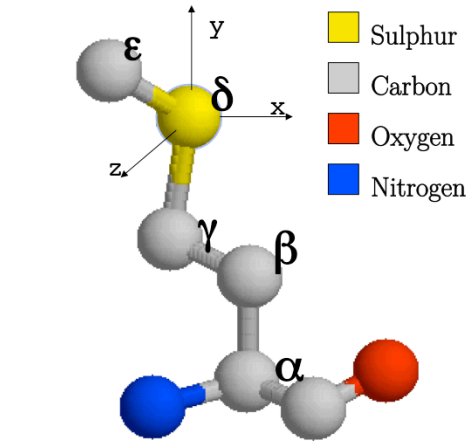
Quality metrics exist, but are not trivial to interpret.

Requirements of a computable abstraction of structure

To store structures we need at least:

- coordinate and
- chemical type

information.



Met

Structure data formats also attach labels to each atom that reflect the topology of the atom in the residue, the position in the protein chain, and the chain identifier, in case the coordinate file contains more than one chain. Following IUPAC rules, atoms are labelled outward from the peptide chain with superscripted greek letters – e.g. C^{α}, C^{β} ... etc – and the PDB file format translates that into uppercase strings like CA, CB ...

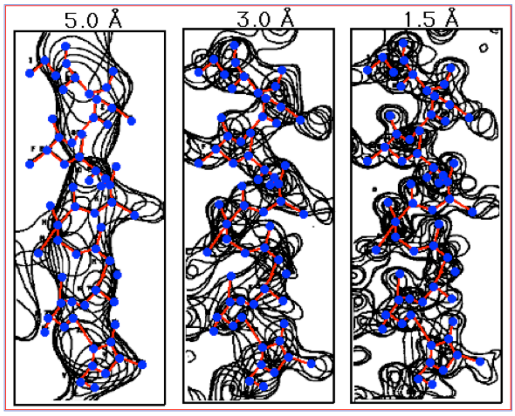
QUALITY METRICS

A number of metrics can be applied to measure the quality of a structural model:

- **Resolution**, R-factor and R-free
- Bond length and angle deviations
- Coordinate error can be estimated from diffraction data

Programs **Whatcheck** and **Procheck** calculate quality metrics:

(Procheck also analyses NMR models)



http://www.sci.sdsu.edu/TFrey/Bio750/Bio750X-Ray.html

Rules of thumb for 'good structures':

Resolution < 2Å, R-factor < 20%, mean coordinate error < 0.2 Å, RMSD bond-lengts: < 0.02Å

CRITERIA FOR CHOOSING MODELS

If multiple models exist for one protein, which one should we choose for analysis? Criteria include ...

- Quality
- Method
- Conditions
- Completeness / coverage
- Similarity
- Complex vs. free, holo- vs. apo-
- Spacegroup and crystal contacts

How to weight these criteria relative to each other depends on our objectives. The quality of the structure is always important, but, for example, a structure with lower resolution but bound ligend or inhibitor may be more informative of the active-site mechanism than an *apo*-structure of the same protein at higher resolution. Here is some context on the criteria:

QUALITY: very generally, the higher, the better. There are two contributing aspects: (*i*) the quality of the dataset, which varies according to the intrinsic order and size of the crystals, or the solubility of the protein and thus its concentration for NMR strcture and (*ii*) the quality of the model-building. The latter has been automated to a large degree in recent years. Look for: low values of nominal resolution, ideally below 2.2 Å, and the R_{free} value, which should really not be more than 25 to 30% larger then the R value, otherwise this may be an indication of "overrefinement", i.e. an overly aggressive fit of coordinates into the electron density map, at the expense of stereochemical plausibility.

METHOD: In general X-ray structures are a more faithful representation of actual conformations wheras NMR structures owe a lot of coordinate details to the force filed used in the refinement. Of course, if the best available crystal structure has poor resolution, e.g. > 2.8Å, than a well determined NMR structure would be preferrable.

CONDITIONS: Structre may be influenced by the (harsh, non-physiological) conditions of crystallization.

COMPLETENESS: The more amino acids have been striucturally determined, the better.

SIMILARITY: The single most important determinant of whetehr inferences can be made between proteins of known and unknown structure is the degree of sequence similarity.

COMPLEX: Structures with bound ligands or inhibitors are most informative regarding active site conformation.

CRYSTAL CONTACTS: ... may induce confronmational changes in the protein.

Structure Data

Population experiments

X-ray: one structure: time and population averaged coordinates

NMR: one "refined" consensus structure, or many plausible MODELs

Incomplete - not all atoms included

Hydrogens are not visible in X-ray diffraction experiments;

segments of the protein with static disorder (alternative conformations) or dynamic disorder (segments in motion) may be missing.

Protein atoms, ligands, solvent are all in one file

multiple chains may be present (homo- or hetero oligomer);

alternate locations may be explicitly modeled in high resolution structures.

Crystallographic space: one asymmetric unit

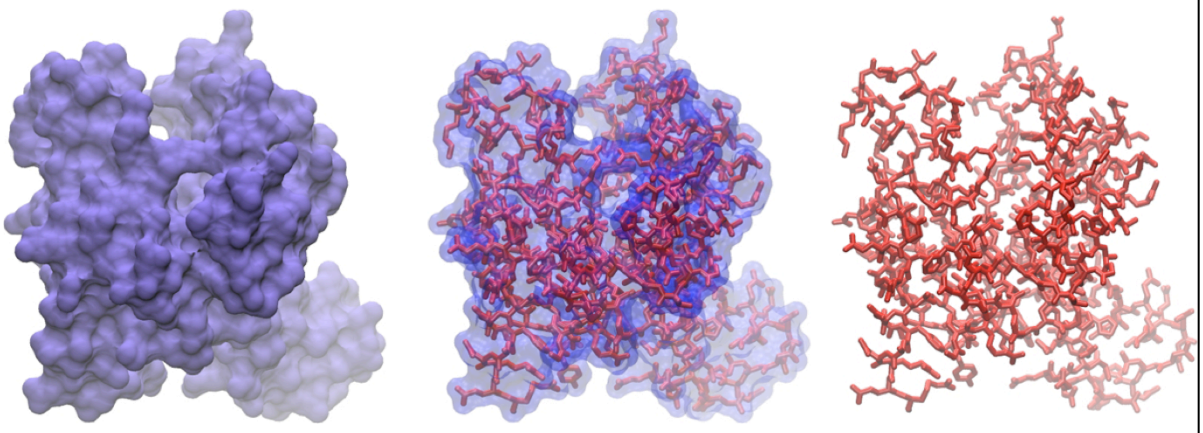
This is not always relevant (you may need to complement the asymmetric unit to a biological unit).

Be aware of the implications and limitations of structure data.

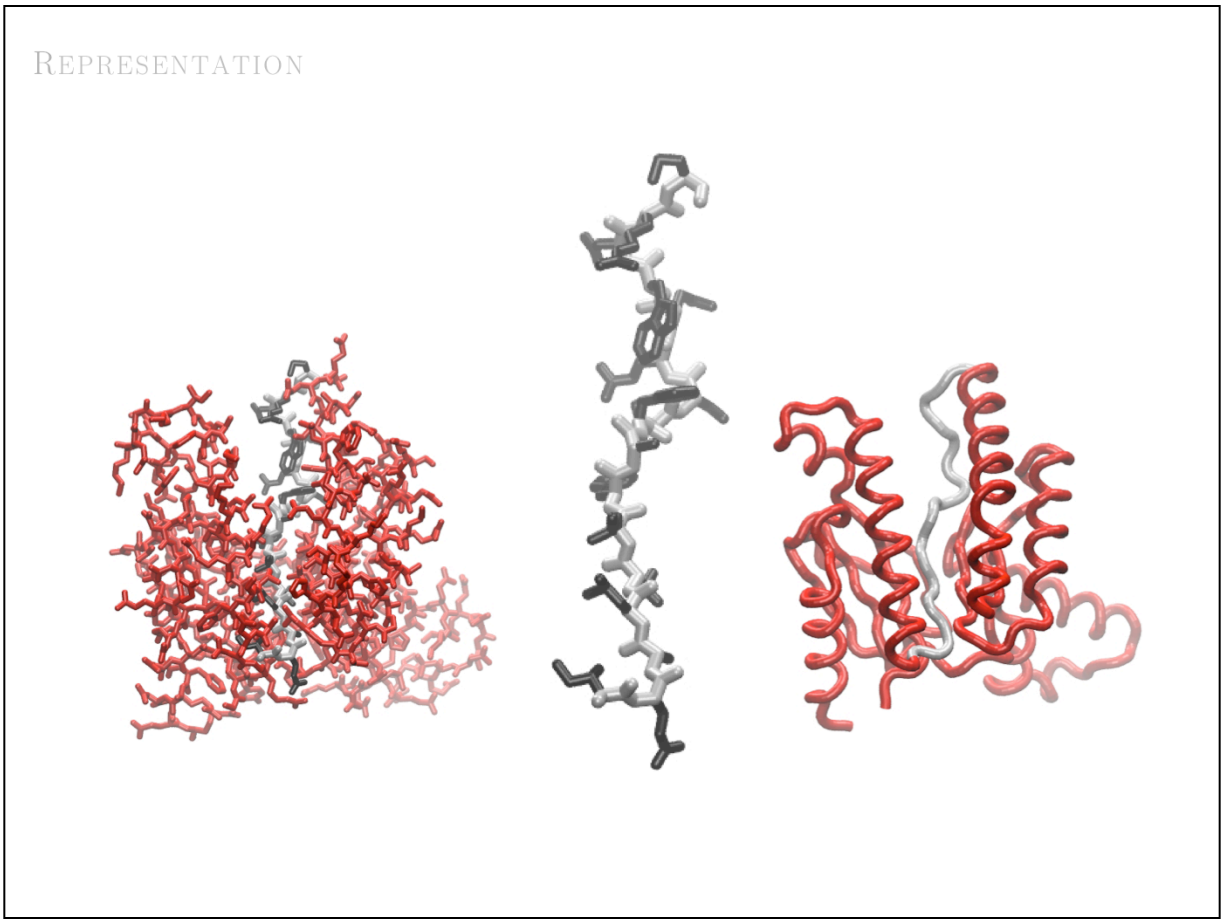
SECTION

Visualization

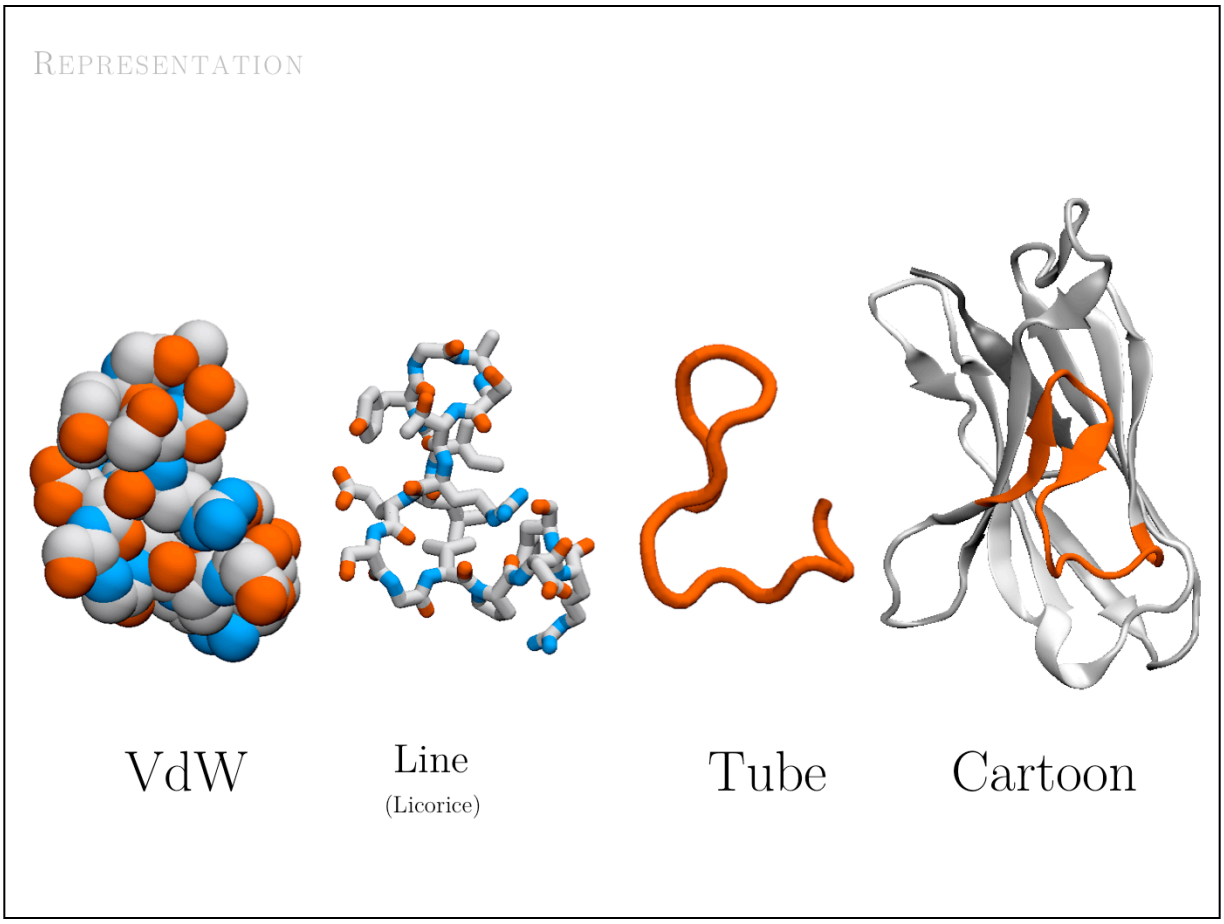
Representation



There are a number of conventions to represent proteins. Surface, atoms and bonds ...

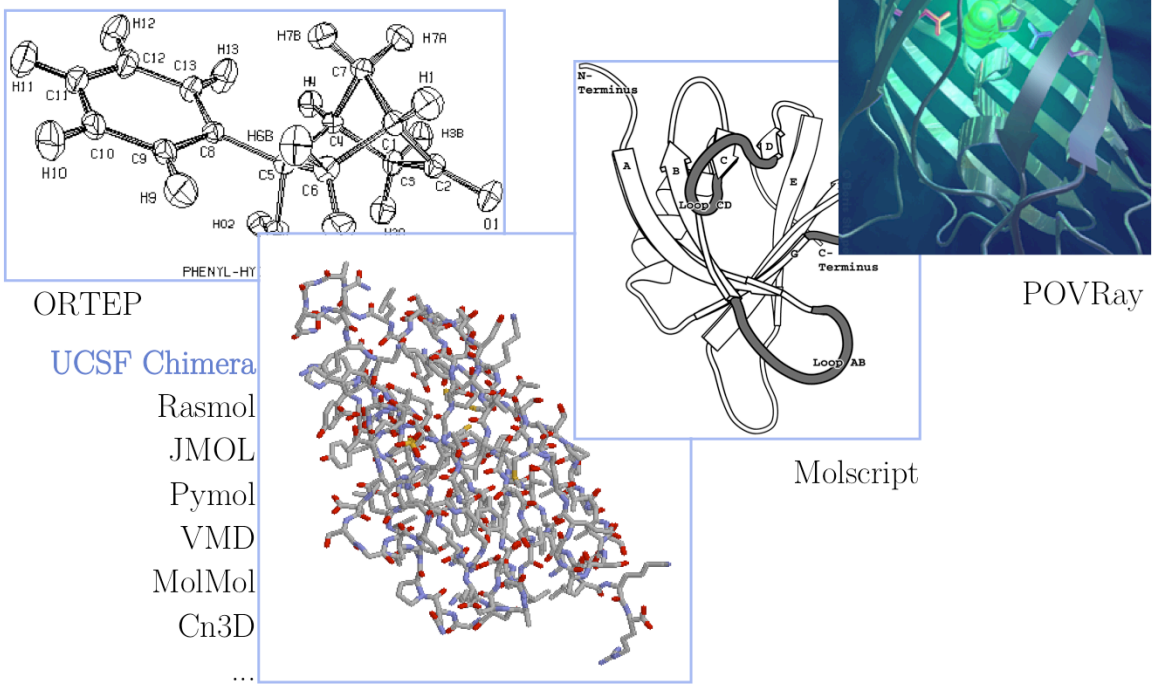


Often it is useful to simplify the representation - a "tube" representation like on the right hand side makes an organisational principle easily visible - proteins are actually long strands



Various options for visualization give different levels of abstraction, from space-filling model, to line drawings that emphasize chemical connectivity, to tube or cartoon models that trace the overall topology of a protein fold.

Visualization



ORTEP

UCSF Chimera

Rasmol

JMOL

Pymol

VMD

MolMol

Cn3D

...

Molscript

POVRay

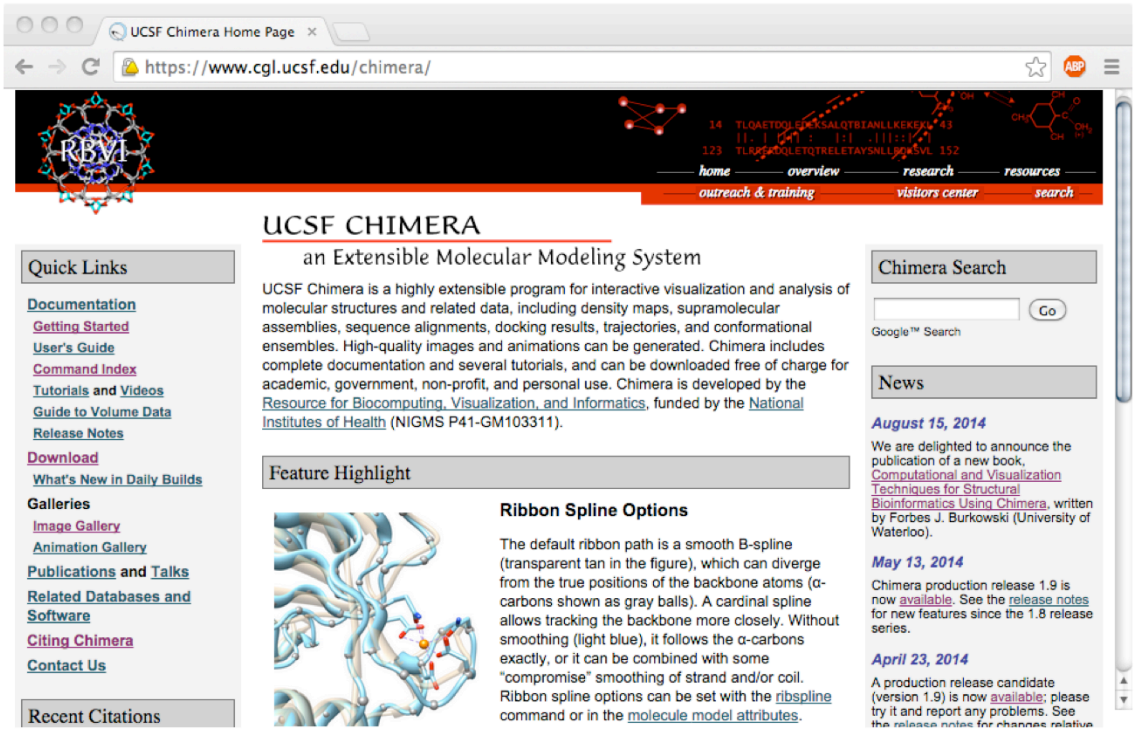
Various tools exist for different visualization tasks.

ORTEP was one of the earliest programs and plots thermal ellipsoids with three degrees of freedom. Most protein structures do not have this data available but many small molecule structures do.

There is a host of 3D, interactive molecular visualization programs available, in this course we use **ChimeraX**. Examples of programs that draw moleculer scenes for publications include Molscript, or the generic ray-tracing program POVray. ChimeraX can generate povray input files, they can be edited to generate for example an internally illuminated view of Green Fluorescent Protein (GFP).

UCSF Chimera

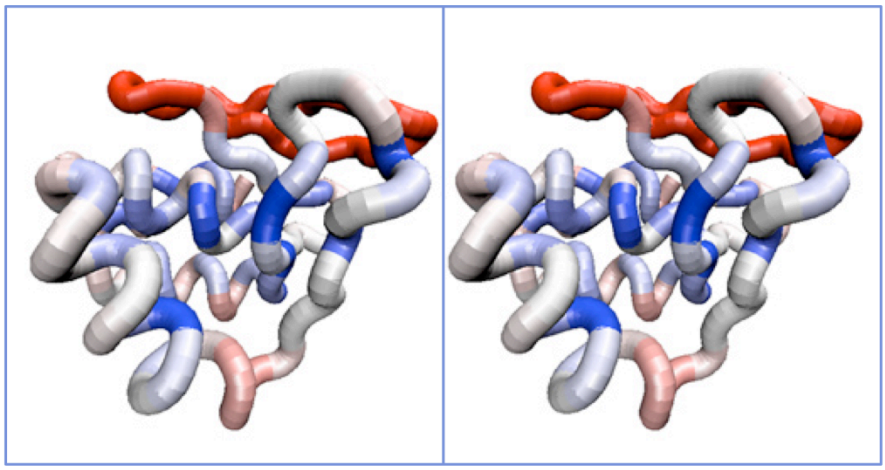
https://www.cgl.ucsf.edu/chimera/



UCSF Chimera is a free, widely used, richly featured and well supported molecular viewer that we will be using throughout the course. On the homepage, you can download the program, find tutorials and handbooks and subscribe to the support mailing list or simply browse the list archives.

Stereo Vision

Stereo vision is a **prerequisite** to understand structure!



"Wall-eyed" stereo view of the fold of 1BM8, the APSES domain of yeast Mbp1. The left- and right eye images are subtly different and allow the visual system to inetrpret the image as having depth.

Stereo viewing is essential to understand structure. All visual aids like shading, shadows and other depth-cues can not replace the entirely different *quality* of a true 3D image. Acquiring this skill opens a whole new, **amazing** way to interact with and understand biomolecular models.

Learning to view molecules in stereo requires to consciously uncouple a reflex in the visual system that normally copules focus and convergence. This can be learned by anyone with (reasonable) binocular vision through some simple exercises over the course of about a week.

This requires the use of a molecular visualization tool, like Chimera.

SECTION

Interpretation:

STRUCTURE AS A MAP OF FEATURES

STRUCTURE INTERPRETATION

Types of questions we might answer with reference tp 3D structures include ...

- Which part of my structure appears to be conserved ?
- Are two functionally important residues possibly in contact ?
- Where is Asn220 relative to the active site ?
- Might the mutation E123A possibly have something to do with protein stability ?
- Could Thr234 be phosphorylated?
- I want to clone my protein into a yeast two-hybrid system: should I fuse the DNA binding domain to the N- or the C-terminus ?

It is useful to view structures as a spatially integrated map of annotations. Spatial relationships provide the context that allows mechanistic, molecular interpretations of observed functions and behaviour.

STRUCTURE → FEATURES

Bonds
Angles, plain- and dihedral
Surfaces
Chemical potential of residues
Static and dynamic disorder
Structural similarity
Electrostatics
Posttranslational modification sites
Conservation patterns (structural and functional)
Quarternary structure
Unexpected homology
[...]

Think of *structure analysis* as creating a spatially integrated feature map

Most features can be derived from coordinates, but structures can also be used for database searches.

STRUCTURE → MECHANISM

Knowledge of structure features (sometimes) can give rise to mechanistic explanations. (But this is not automatic.)

Active site geometry (the textbook example: catalytic triad)

Mechanisms that require a particular local environment for a chemical or other reaction (Lipases, GFP)

Mechanisms that require a particular spacing of ligand binding and active site (eg. Proteasome, Ubiquitin ligases)

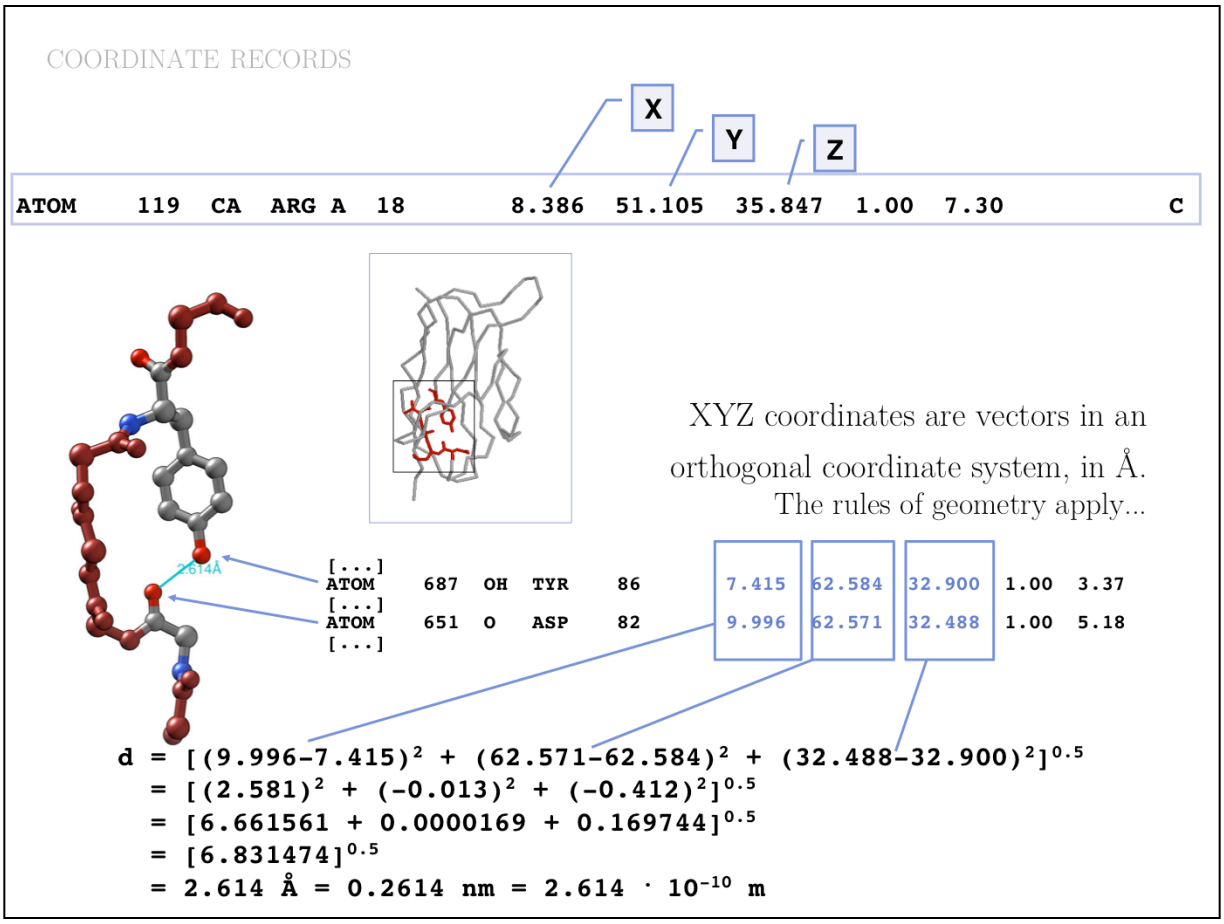
Mechanisms that require the concerted activity of domains (eg. Phospholipase: phosphatase domain + PH-domain)

Mechanisms that require a coupling between catalysis and chemical potential (E.G. membrane pumps)

Mechanisms that require a coupling between catalysis and mechanical motion (ATP synthase, myosin)

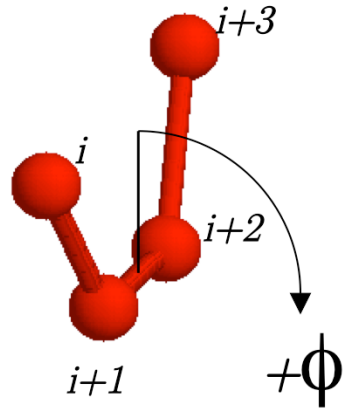
Mechanisms that require a particular molecular shape (Membrane pores, phage/DNA injection assembly, molecular mimicry, antifreeze proteins)

Knowing the protein structure is also one of the ways to begin understanding the mechanism of the protein's function.



This example shows the explicit calculation of a distance from the information in two ATOM records. This is a straightforward application of Pythagoras' theorem in three dimensions. The result shows us that we have an H-bond with an ideal length between the phenolic hydroxyl group of a tyrosine and a backbone carbonyl oxygen.

DIHEDRAL ANGLES



Single bonds:

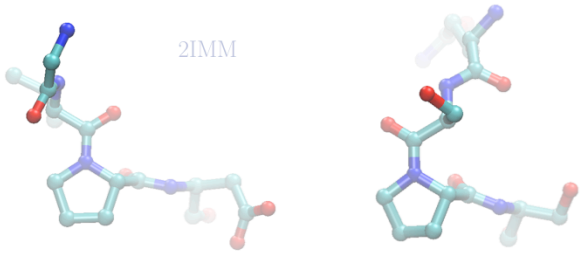
Freely rotatable, but constrained by steric overlap. Small energetic barrier, preference for staggered conformations.

Peptide bonds:

Largely planar geometry. Significant energetic barrier to isomerization. Always synthesized in *trans* - but ~ 10% of x-proline bonds are *cis* in folded structure

Double bonds:

Constrained to planar geometry. Not observed to isomerize.



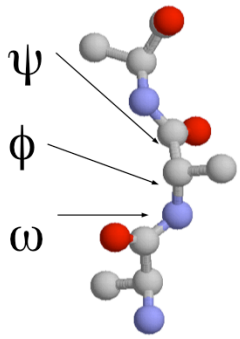
x-pro peptide bond in ...

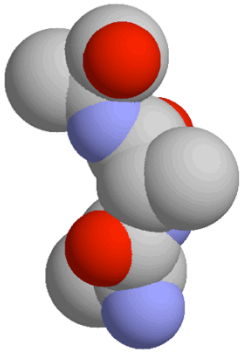
trans *cis*

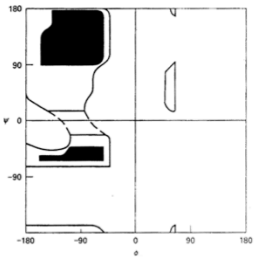
All petide bonds are synthesized in *trans* at the ribosome. Normal peptide bonds have a very high energy difference between cis and *trans* configurations due to steric clashes, but not peptidyl-prolyl bonds. From the view of the peptide bond, the proline's C^{α} and $C^{\delta}T$ atoms look almost identical. Thus, at equilibrium, about 10% of these bonds will be in the disfavoured *cis* form. Interestingly, that is the same fraction that is found to be in *cis* in folded proteins.

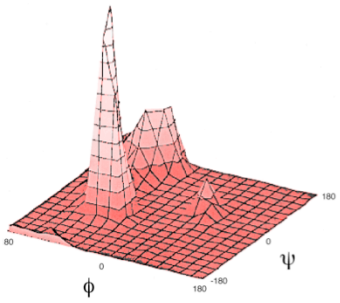
The required isomerization of peptidyl-*cis*-proline bonds is a common rate limiting step for protein folding.

Ramachandran plot





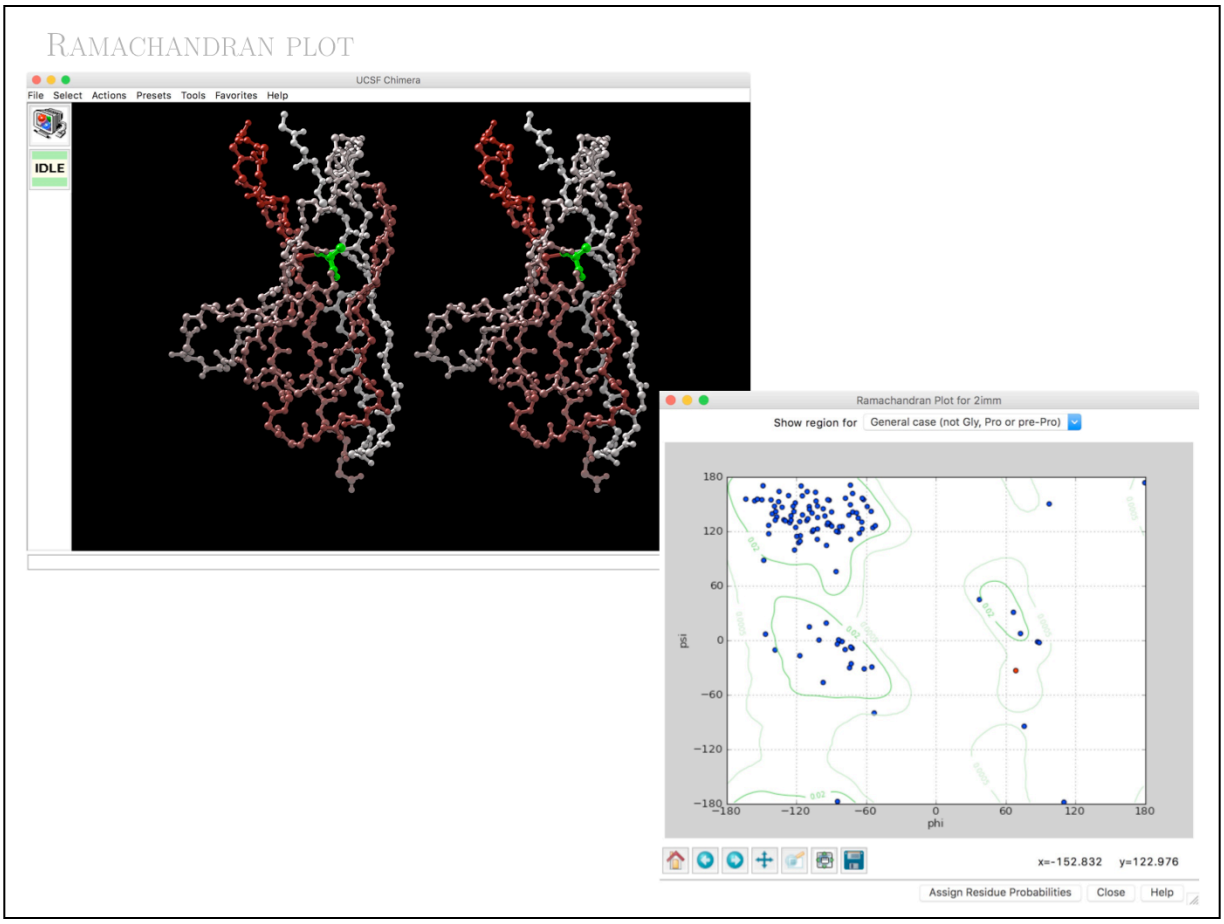




Rotatable bonds in the backbone are named ϕ,ψ and ω.	Due to steric overlap, not all combinations of (ϕ,ψ) are allowed.	Allowed and forbidden regions of (ϕ,ψ) space are shown on the Ramachandran plot.	Observed (ϕ,ψ) values reflect the theoretical boundaries well.

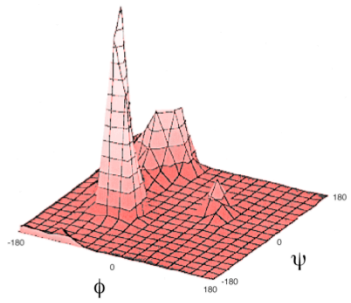
The dsitribution of actually observed backbone rotational angles hints at a relationship between frequency and energy.

Ramachandran plot



Ramachandran plots can be easily displayed via the Chimera Model panel. The image highlights an outlier in the 2IMM structure, an alanine in a region in which only glycine residues would be expected, shown in green. It is positioned at the tip of a tight, three-residue, so-called γ-turn that connects two β-strands. Interestingly, glycine is *not* preferred in this position for that particular structural pattern.

structure obeys Boltzmann's law



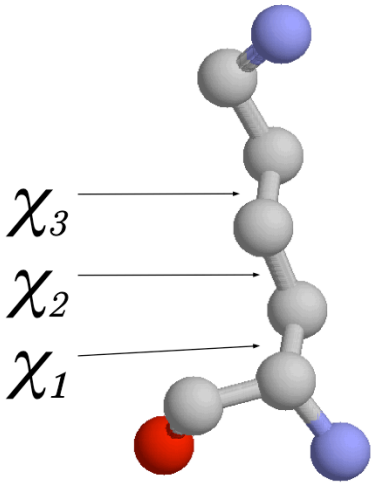
$$\Delta G = -RT \ln K$$

Empirically, we see that measurable features of protein structure are observed in approximately Boltzmann distribution regarding their *free energy*. This is highly non-trivial and provides a conceptual framework for all considerations of protein folding and interaction.

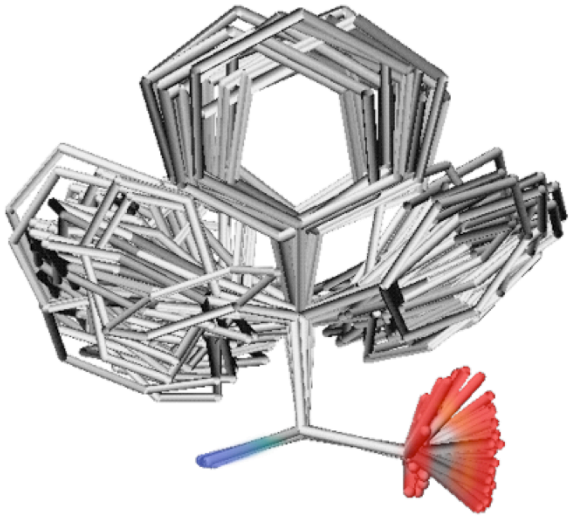
Low energy states are more frequent, but higher energy states are not impossible and often associated with function.

Boltzmann's law becomes apparent in frequency distributions of practically all aspects of protein structure.

SIDE CHAIN ROTAMERS



Ponder & Richards (1987)
J. Mol. Biol. 193, 775-791



The side chains of 100 randomly chosen Phe residues, superimposed on the backbone atoms, clearly cluster into discrete preferred conformations.

Rotamers are low-energy conformations of side-chain dihedral angles. Only a restricted set of rotamer states and combinations are significantly populated in natural proteins. This tremendously simplifies protein structure modelling and prediction problems. This insight is also useful to guide analysis, e.g. in enzyme active sites the rotamers often exist in strained, infrequent conformations.

See the Backbone Dependent Rotamer Library, compiled by Roland Dunbrack (http://dunbrack.fccc.edu/bbdep2010/)

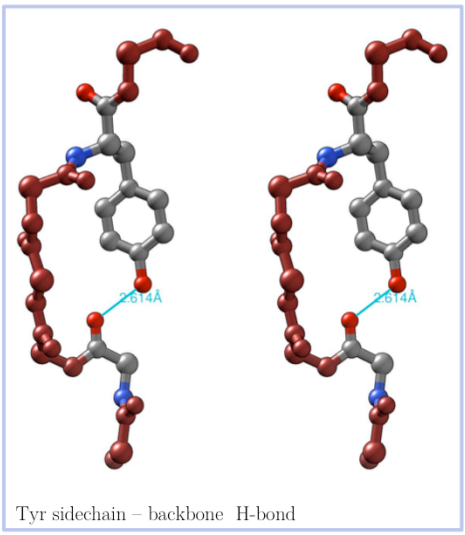
H-BONDS

Example: TYR - Side Chain Donor

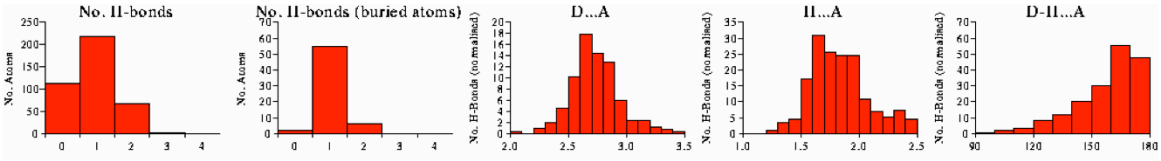
OH can donate a single hydrogen, here in a ~2.7Å H-bond to a Thr OG1. (The OH→H bond is 1.0 Å long and lies in the plane of CE1, CE2, CZ and OH forming an angle of 110 degrees with the CZ-OH bond.)

Assignment of donor/acceptor status in crystal structures may be ambiguous.

Free energy can be up to 20kJ/mol but the contribution to the total free-energy of structure stability may approach zero. Why?



Tyr sidechain – backbone H-bond



Distribution of H-bond counts in all and buried residues, D-A distances, H-A distances and D-H-A angles inTyr sidechains.

McDonald & Thornton (1994) J. Mol. Biol. 238, 777-793

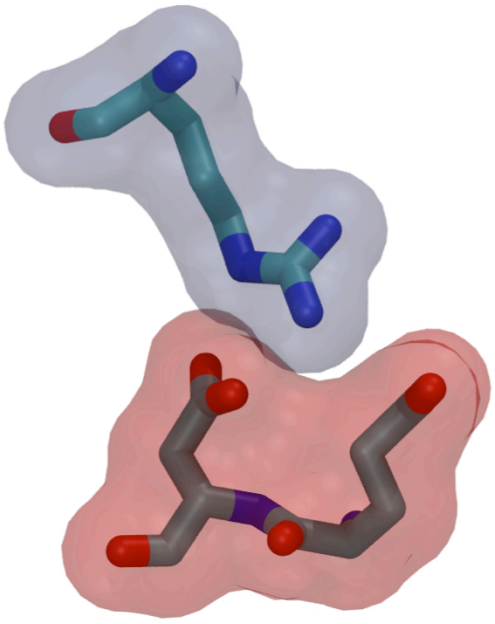
Hydrogen bonds have clear and well defined conformational preferences. The ideal donor-acceptor distance is between 2.7 and at most 3.5 Å, the angles between the non-hydrogen donor and acceptor atoms vary only to a small degree. They are important as a mechanisms to "lock-in" parts of protein structure, even though in and of themselves they don't significantly contribute to the free energy of folding – in the unfolded state, just as many H-bonds are formed with solvent.

Cf. the **Atlas of Hydrogen Bonds** compiled by Ian McDonald and Janet Thornton (http://www.biochem.ucl.ac.uk/bsm/atlas/) – or compile your own statistics in R with the bio3d package.

SALT-BRIDGES

Also called ion-pairs.

Due to desolvation energy requirements, salt bridges usually occur in clusters. They are long range, electrostatic interactions. The distance cutoff for the required minimum approach of charged atoms varies between researchers but 4.5 Å is a reasonable value. The distance should definitely not allow space for an interceding water molecule.

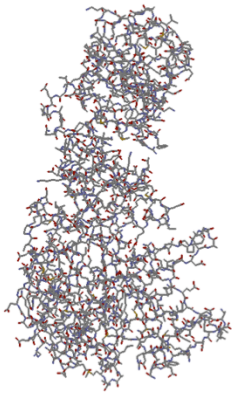


Salt-bridge in 2IMM: R66 – E86/D87

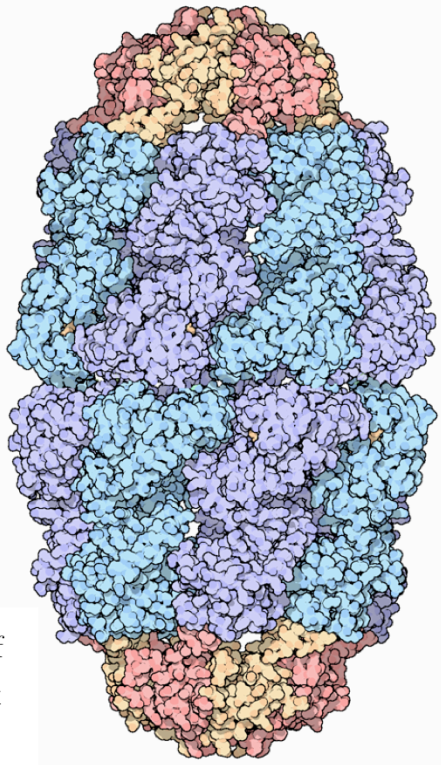
Thornton, JM. (1983) Ion-pairs in Proteins. *JMB*, 168, 867-885

Matsui I, Harata K. (2007) Implication for buried polar contacts and ion pairs in hyperthermostable enzymes. FEBS J. 274(16):4012-22.

MOLECULAR SURFACE



Chain "A" of
1AON.PDB -
GroEL/ES complex

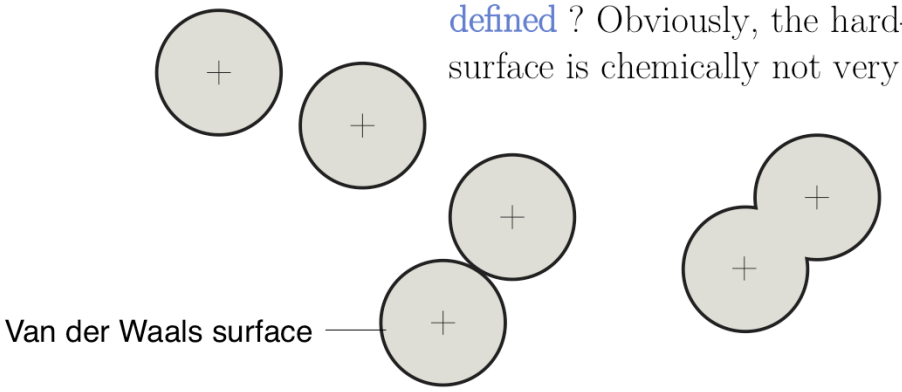


Surface rendering of
GroEL/ES complex
(D. Goodsell)

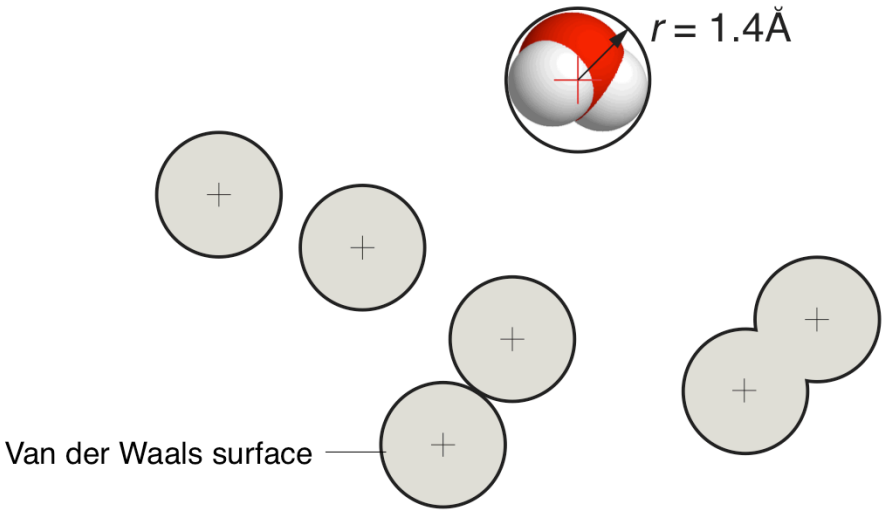
A set of coordinate ''points'' has no surface! Ascribing a surface to coordinates, in order to map coordinate sets to physical molecules, requires to consider the volume of space that is '''occupied''' by atoms, i.e. space that cannot be simultaneously occupied by other atoms.

Surface provides a visual metaphor, and a useful tool to map properties.

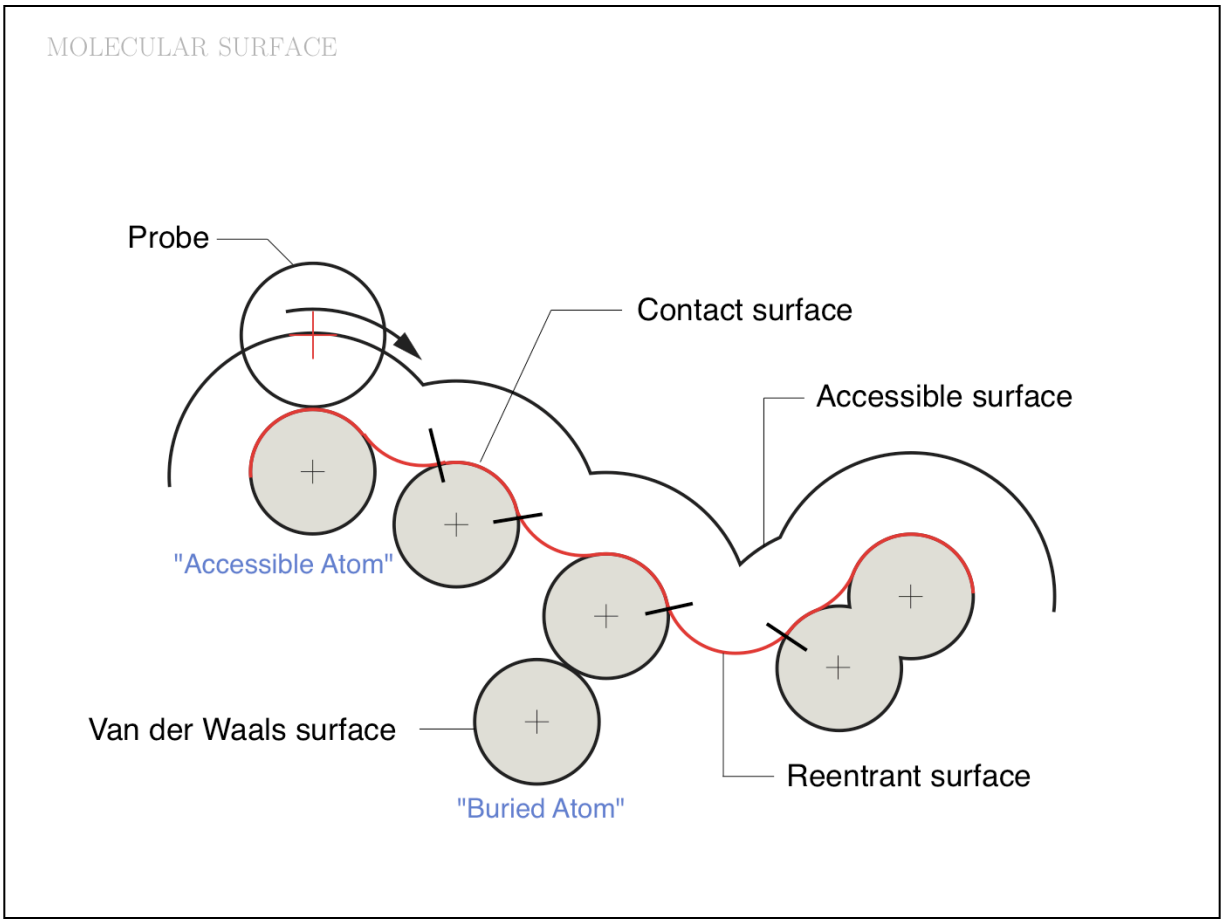
But how can a molecular surface be defined ? Obviously, the hard-sphere surface is chemically not very relevant.



Definition of "surface" via a probe !



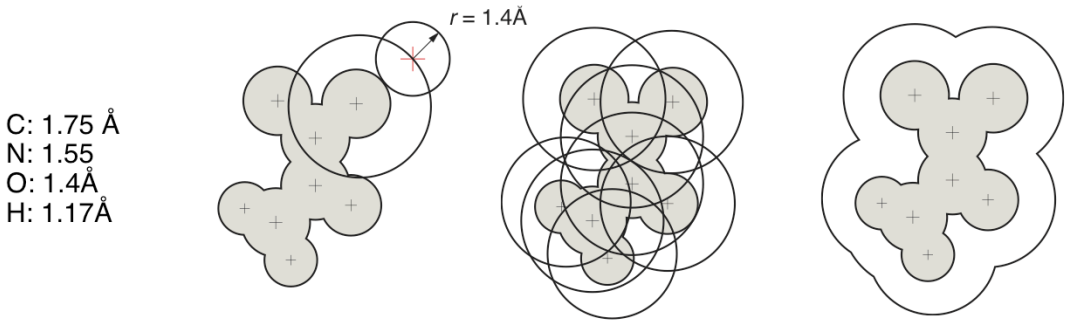
A molecular surface can be defined as the boundary that excludes a probe. Commonly, a probe of 1.4 Å radius is used - this is the spehrical approximation of a water molecule and thus this probe defines a **solvent accessible surface**.



The "Accessible Surface" depends on the probe radius.

SAS algorithm

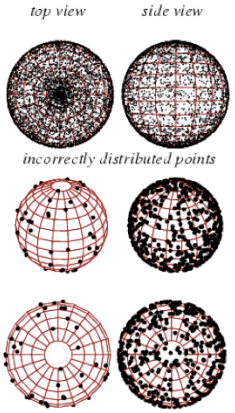
1. Draw a sphere around each atom, with a radius of (VdW + solvent probe).
2. Erase all overlapping sphere surfaces.
3. The remaining area is the accessible surface.



An easy way to calculate SAS (solvent accessible surface) areas is to (1) randomly place a given number of points on a sphere around an atom center, (2) erase points that are overlapped by neighboring atoms' spheres and (3) count the number of points that remain. The fraction of remaining points times the surface area of the original sphere approximates the solvent accessible surface area. Surface areas of individual residues are often expressed as relative values, i.e. the actual accessible surface area of a sidechain divided by the area of a maximally solvent-exposed sidechain. This characterizes the degree to which a sidechain is buried or accessible in a protein structure.

parameters and assumptions

Problem: Analytical solution inefficient.
Solution: Numerical solution with probe points
Problem: Regular placement of n probe points is not anisotropic
Solution: Stochastic placement
Problem: Stochastic placement quite irregular
Solution: Enforce minimum separation
Problem: Efficiency
Solution: Place points only once, translate as needed
Problem: What is a good value for n ?
Solution: Try several n, evaluate standard deviation
Problem: Should n be constant per atom, or per area ?
Solution: dots/area - need to scale dots with r_{VdW}
Problem: Hydrogens - where to get united atom radii ?
Solution: Literature search.
Problem: Reference areas for relative SAA needed
Solution: Model explicitely, as tripeptides
[...]



$\theta \in [0, 2\pi)$

$\phi \in [0, \pi]$

$u,v \in [0,1]$

$\theta = 2p\ u$

$f = \cos^{-1}(2v-1)$

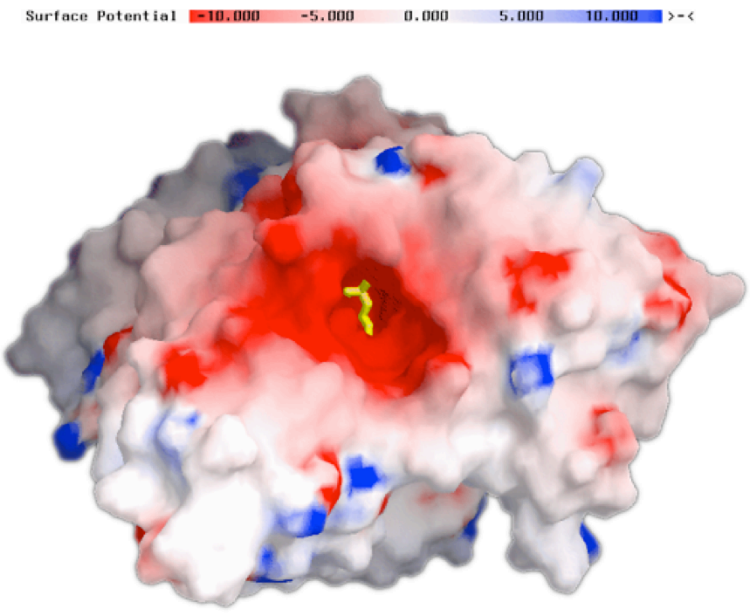
http://mathworld.wolfram.com/SpherePointPicking.html

Even a straightforward algorithm has parameters and assumptions. Results are always only meaningful when we consider which algorithm was used and how the parameters were set..

Results cannot be compared if algorithm and parameters are not specified. This is a common problem, and an important issue for “repreducible reesearch”.

SURFACE PROPERTIES

- Properties of atoms (B-factors)
- Ensemble properties (hydrophobicity, conservation)
- Geometry (local curvature)
- Fields and potentials (isosurfaces, binding potential)

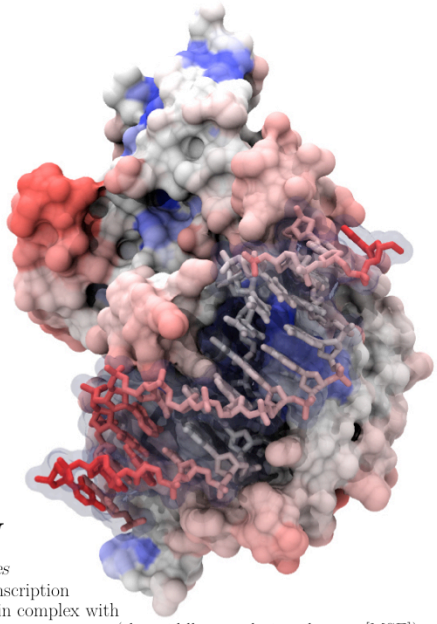


AChE (1ACL.PDB)
surface color coded by electrostatic potential

Surfaces represent the parts of a protein that interact with the surroundings. In this example, the electrostatic potential mapping shows how an electrostatic potential gradient attracts the substrate molecule into Acetylcholine-esterase's active site. Thruogh this, AChE is faster than a diffusion-limited mechanism woudl predict!

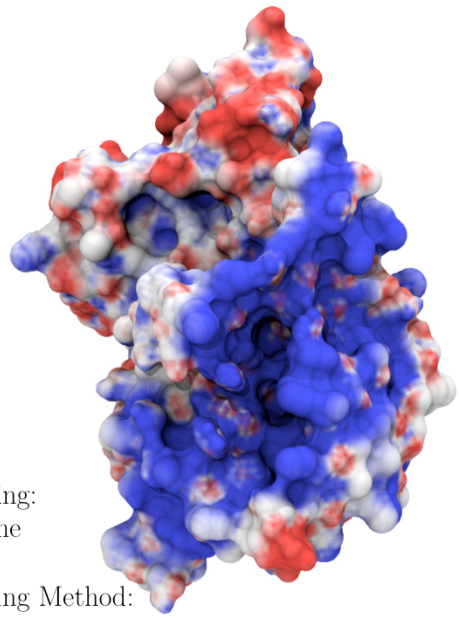
ELECTROSTATICS IN PRACTICE

Chimera has routines for calculating electrostatic potential.



2ETW

Saccharomyces cerevisiae transcription factor Ndt80 in complex with its DNA consensus sequence (the middle sporulation element [MSE]).

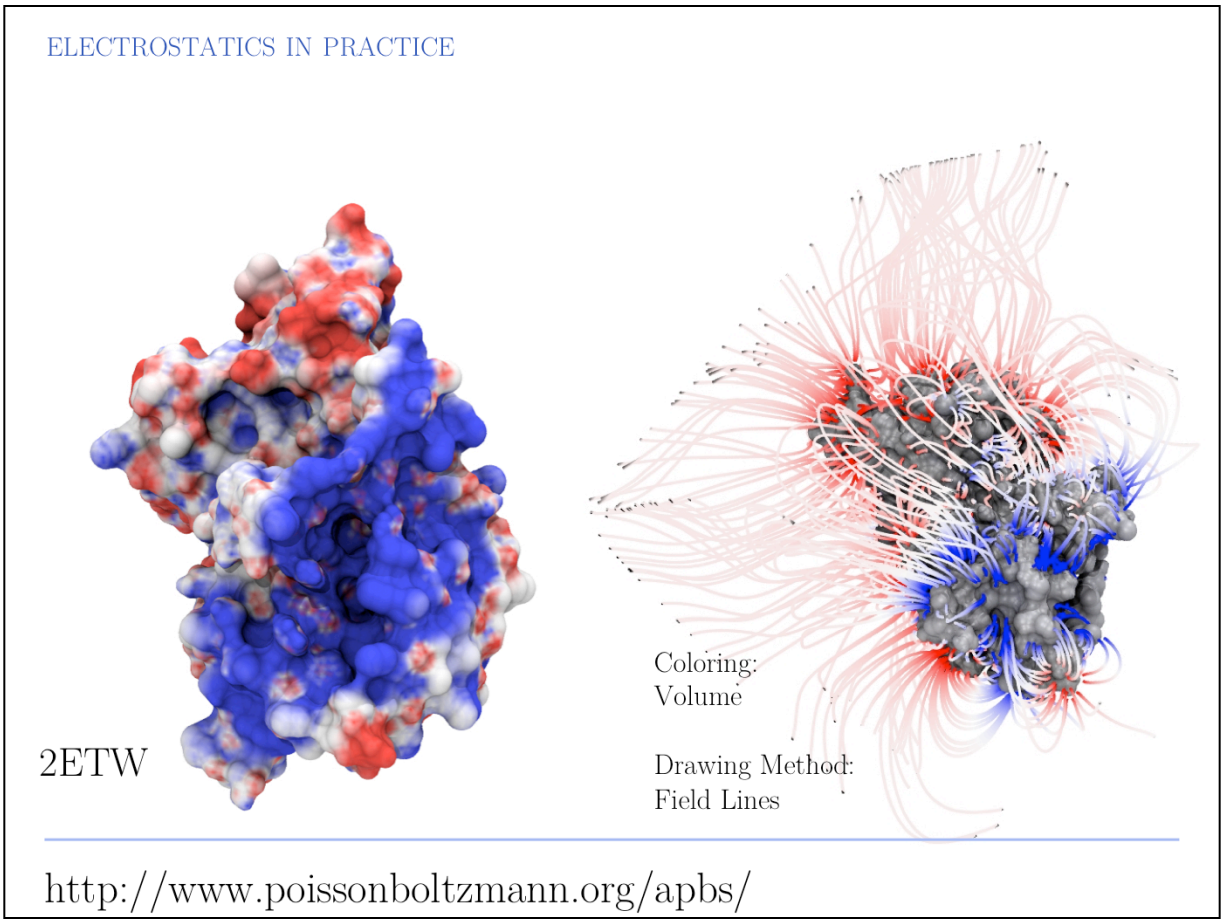


Coloring: Volume

Drawing Method: Isosurface

http://www.poissonboltzmann.org/apbs/

Electrostatic complementarity of a DNA-binding site with the DNA ligand becomes strikingly obvious in this electrostatic potential map of the 2ETW trancription factor.



Field-lines emphasize regions of space from which the ligand is excluded. In this example of the 2ETW transcription factor structure, this provides a mechanistic explanation of how the protein "scans" B-DNA strands for cognate binding sites.

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

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

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