

X-ray Crystallography

Prof. Leonardo Scapozza
Pharmaceutical Biochemistry
School of Pharmaceutical Sciences
University of Geneva, University of Lausanne
E-mail: lenardo.scapozza@pharm.unige.ch



Aim

- Introduce the students to X-ray crystallography
- Give the students the tools to “evaluate” a X-ray structure based scientific paper



Outline

- The History of X-ray
- The Principle of X-ray
- The Steps towards the 3D structure
 - Crystallization
 - X-ray diffraction and data collection
 - From Pattern of Diffraction to Electron Density
 - X-ray structure quality assessment



An extract of a structure paper

2.1. Crystallization

- The hTK1 was cloned as N-terminal thrombin-cleavable His6-tagged fusion protein missing 14 amino acids of the N-terminus and 40 amino acids of the C-terminus of the wild type hTK1 sequence of 234 amino acids (this construct is further on called hTK1). The purified hTK1, consisting of residues 15-194 of the wild type sequence plus an N-terminal extension of 15 residues containing a His6-tag, was eluted from gel filtration column at a concentration of approximately 7 mg/ml with a buffer containing 5 mM Tris at pH 7.2, 10 mM NaCl and 10 mM DTT. For **protein crystallization** we used the **hanging drop method** at 23°C. Initial conditions for crystallization were found using Crystal screen Cryo no. 40 (Hampton Research). The protein solution was mixed in a 1:1 ratio with crystallization buffer (0.095 M tri-sodium citrate pH 5.5, 12% PEG 4000, 10% isopropanol) to set up drops of 6 µl. The reservoir contained 500 µl of crystallization buffer.
- Crystals appeared after 3 days and grew within 3 days to **sizes** of about **300 x 150 x 70 µm³** and diffracted to **1.8 Å resolution**. They belong to **space group C2** with **a= 157.5 Å, b= 122.9 Å, c= 115.3 Å, α = γ = 90° and β= 130°**. **The asymmetric unit contains 8 monomers** forming two tetramers. The solvent content is 48%. For X-ray data collection the crystals were cryo-protected using crystallization buffer with additional 30% glycerol soaking for one minute. The crystals diffract up to **1.8 Å resolution** and are stable under X-radiation from beamline X06SA at Swiss Light Source (Villigen/CH).

2.2. Data collection, phase determination and refinement

- Three datasets were collected from two crystals. A fluorescence scan from one crystal was performed and data were collected at two wavelengths (**λ=1.2776 Å (zinc edge) and 0.9196 Å**). The data processing was performed with the program XDS [11]. These data sets were used for **Multi-wavelength Anomalous Diffraction (MAD) phasing**. The zinc coordinates were determined with program CNS [12]. Initial two wavelength MAD phases to 4 Å were obtained with SHARP [13] and extended to 1.83 Å by eight-fold NCS averaging, solvent flattening and histogram matching implemented in RESOLVE [14] using the third data set (**λ= 0.9793 Å**) collected from the second crystal. **An initial model was obtained with RESOLVE, which was subsequently rebuilt and refined** with XFIT [15] and REFMAC [16], respectively. Model refinement was started using NCS restraints, which were stepwise released and finally omitted. Water molecules were introduced using ARP_WATERS [16] when the **R-factor reached 20%**. Refinement **results were checked with PROCHECK [17] and WHATCHECK [18]**. **Coordinates and structure factors have been deposited in the Brookhaven Protein Data Bank (accession code 1w4r)**. The figures were produced with POVScript+ [19]. For secondary structure assignment DSSP [20] was used. The **B-factor** plot was calculated with BAVEAGE [16] and the density correlation was done with the program OVERLAPMAP [21]. The numbering of the hTK1 [EC 2.7.1.21] sequence corresponds to the hTK1 SwissProt entry P04183.



Birringer, M.S., et al. Scapozza, L. (2005) Structure of a type II thymidine kinase with bound dTTP *FEBS Lett.* 579(6):1376-82.

An extract of a structure paper

TABLE 1: Data collection and refinement statistics.

Data set	dTTP:hTK1
Diffraction data ^a	
X-ray source	X06SA (SLS Villigen/CH)
Unit cell dimensions (Å)	a = 157.5; b = 122.9; c = 115.3; $\alpha = \gamma = 90^\circ$; $\beta = 130.0^\circ$
Resolution range (Å)	20.0–1.83 (1.95–1.83)
Completeness (%)	99.5 (99.8)
Multiplicity	7.1 (7.1)
Unique reflections	147503 (25536)
R _{int} (%)	8.4 (39.4)
I/ σ	17.1 (7.9)
Space group	C2
Protomers per ASU	8
Wavelength (Å)	0.97934
Refinement and final model:	
R _{work} / R _{free} (%)	15.9 / 18.9
Number of reflections in working set	140080 (10345)
Number of reflections in test set	7423 (542)
<u>Number of non-H-atoms:</u>	
Polypeptide	10106
dTTP atoms	232
Water molecules	774
Dithiothreitol atoms	8
<u>Average B-factors (Å²):</u>	
All atoms	29.0
Polypeptide atoms	27.9
Main chain	26.0
Side chain	29.9
dTTP molecules	43.5
Water molecules	39.5
Dithiothreitol molecule	53.1
<u>Rmsd from ideal geometry:</u>	
Bond lengths (Å)	0.016
Angles (°)	1.7
Ramachandran angles:	
Favored regions (%)	93.8
Allowed regions (%)	6.2



^a The data were collected at 100 K. Values in parenthesis are for the outermost shell.

Birringer, M.S., et al. Scapozza, L. (2005) *FEBS Lett.* 579(6):1376–82.



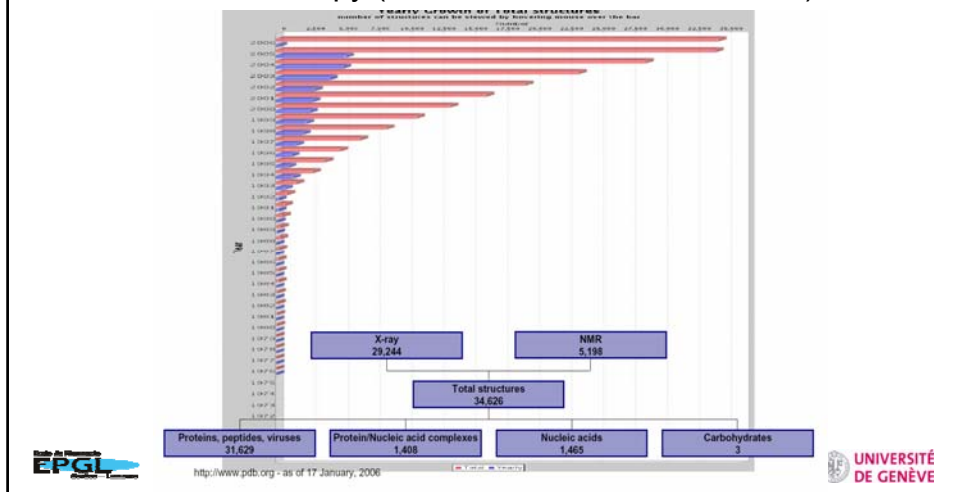
X-ray crystallography timeline

- 1895** W.C.Röntgen Discovery of X-rays (Nobel Prize Physics 1901)
- 1912** W.L.Bragg 1st structure determination (Nobel Prize Physics 1915)
- 1934** J.D.Bernal & D.Crowfoot 1st X-ray photograph of a protein structure
- 1950** L.C.Pauling predicted the existence of alpha-helix and beta-sheet (Nobel Prize Chemistry 1954)
- 1953** F.Crick, J.D.Watson, M.Wilkins, R. Franklin structure of DNA (Nobel Prize Physiology 1962)
- 1954** M.Perutz structure of haemoglobin (Nobel Prize Chemistry 1962)
- 1960** J.C.Kendrew myoglobin structure (Nobel Prize Chemistry 1962)
- 2000** Structure of ribosome's large and small sub-units (Ban, N. et al. *Science* **289**, 905–920; Wimberly et al. *Nature* **407**, 327–339)
- 2001** Structure of Photosystem I at high resolution (Jordan P. et al. *Nature* **411**, 909–917)
- 2003** R.MacKinnon: structure of ion channel (Nobel Prize Chemistry 2003)
P. Agre: structure of aquaporin (Nobel Prize Chemistry 2003)



The majority of the structures is determined by X-ray

- X-ray crystallography (resolution at atomistic level)
- NMR (resolution at atomistic level)
- Electron microscopy (resolution at “molecular” level)



Why Crystallography ?

- The knowledge of accurate molecular structures is a prerequisite for rational drug design and for structure based functional studies.
- Crystallography can reliably provide the answer to many structure related questions, from global folds to atomic details of bonding.

The method's Principle: The Limit of resolution (LR)

Depends on the wavelength you are using: $LR \cong \lambda/2$.

Microscopy ($\lambda = 400$ to 800 nm)

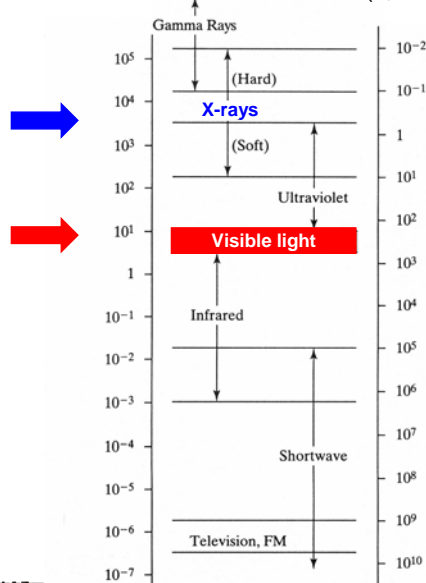
- $LR = 200$ nm = Organelles structures within a cell
- principle of lens/magnifying glass



X-ray ($\lambda = 100$ to 0.1 Å)

- $\lambda < 2.4$ Å
- $LR = 1.2$ Å = distance between two atoms
- Pattern of diffraction are interpreted via mathematics and geometry
- No lens

Photon-Energy Wavelength (λ , nm)



- X-rays are formed by the collision of fast electrons with the matter
- The wavelength depend from the matter
- Example of monochromatic X-ray:



Fe: 1.93 Å

Cu: 1.54 Å



Mo: 0.71 Å

Why Crystallography ?

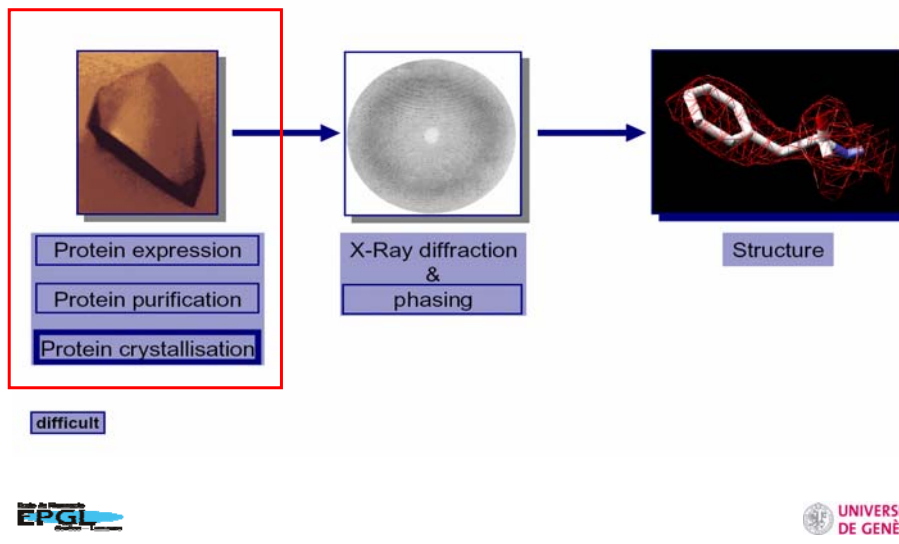
Advantages	Disadvantages
X-ray <ul style="list-style-type: none">• High resolution• No protein mass limit	<ul style="list-style-type: none">• Crystals needed• Possible artifacts due to crystal content and precipitation• Structure is static “average”• Mostly no H seen
<ul style="list-style-type: none">• The price for the high accuracy of crystallographic structures is that a good crystal must be found.	

Determination of 3D structures using X-ray

1. You need a unique crystal. The atoms within this crystal are all well ordered and the molecules have the same conformations
If not a low resolution is expected.
 2. Pattern of diffraction of the unique crystal are collected
 3. The 3D model is constructed in order to interpret the diffraction pattern
-  

The steps towards the 3D structure

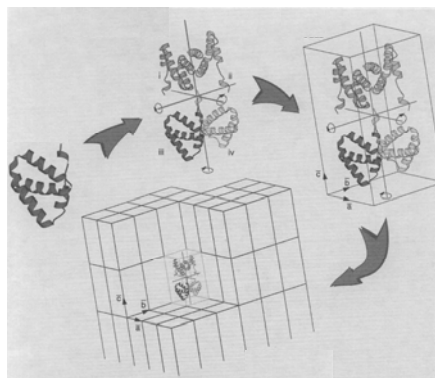


What is a Crystal?

- Is a solid formed by **ordered** atoms and ions
- Ordered means: the same pattern is repeated along a regular lattice



apoHSV1-TK



Why Protein Crystals?

- Crystals, which are three-dimensional arrays of molecules, are required for X-ray diffraction experiments because scattering from individual molecules is far too weak to measure.
- Crystals act as amplifiers by increasing the scattering signal due to the multiple copies of molecules within them.

Strategy for getting crystals

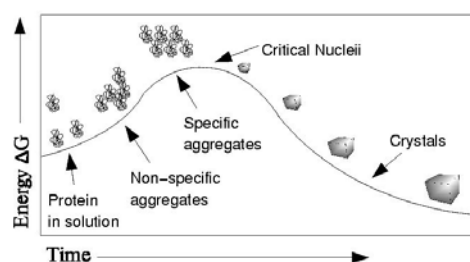
- Isolation of target protein
 - Biological resources (plants, animals)
 - Recombinant protein (*E. coli*, yeast, cell lines, baculovirus)
- Purification
 - Affinity chromatography, Gel filtration ...
- Biochemical characterization
 - Activity, pI, stability, inhibition, pH, salt ...
 - Oligomeric state, secondary structure (CD)
- Crystallization

Crystallization

- A multiparametric process
- Three classical steps
 - Nucleation (formation of first ordered aggregates)
 - Growth
 - Cessation of growth

Energy Diagram

- There is an energy barrier to crystallization
 - Proteins must overcome an energy barrier to crystallize.
 - The critical nucleus corresponds to the higher energy intermediate.
 - The higher the energy barrier, the slower the rate of nucleation.



Energy diagram for crystallization

Crystallization is...

- Macromolecular crystallization is no more (and no less) than forcing a protein to precipitate into a regularly ordered 3D-array... the crystal!

Parameters Affecting Crystallization

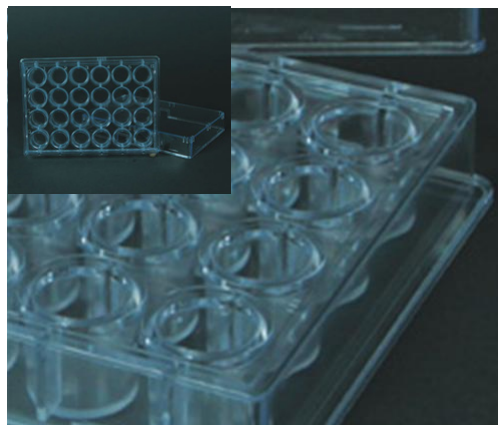
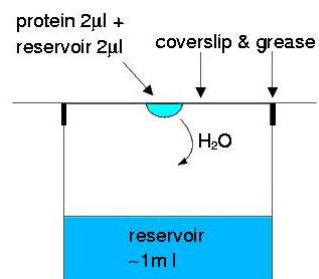
- Intrinsic physico-chemical parameters
 - Temperature
 - pH value (changes)
 - Time (rates of equilibration and of growth)
 - Ionic strength and purity of chemicals
 - Diffusion and convection
 - Volume and geometry of samples and set-ups
 - Dust and other impurities
 - Density and viscosity effects (differences between crystal and mother liquor)
 - Pressure, electric and magnetic field
 - Vibration and sound (acoustic waves)
- Biological parameters
 - Biological origin of the protein
 - Bacterial contaminants
 - Rarity of most biological macromolecules
- Biochemical and biophysical parameters
 - Sensitivity of the protein conformation relative to physical influences (temperature, pH, ionic strength, oxidation,...)
 - Binding of ligands (substrates, co-substrates, inhibitors, metal ions)
 - Additives (reducing agents, non ionic detergents)
 - Ageing of samples (red-ox effects, denaturation, cleavage)

General Rules

- The more you know about your protein, the more likely you get it crystallized!
- Homogenous, compact and globular proteins are more likely to crystallize than heterogenous and floppy ones!

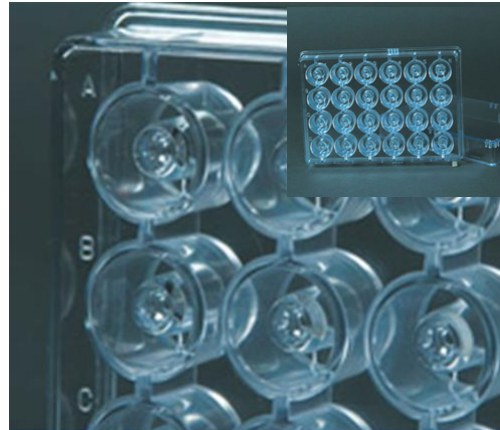
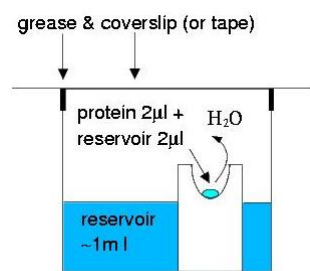
Vapor Diffusion Method 1

Hanging drop



Vapor Diffusion Method 2

Sitting drop

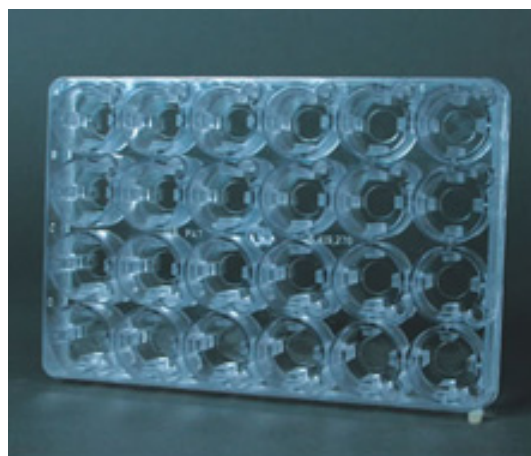
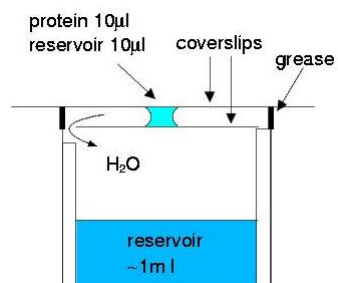


EPGL

UNIVERSITÉ
DE GENÈVE

Vapor Diffusion Method 3

Sandwich drop

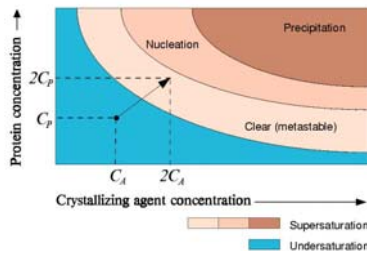


EPGL

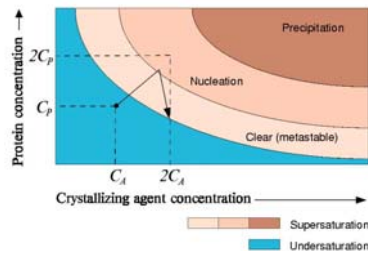
UNIVERSITÉ
DE GENÈVE

What happens within the drops: Phase Diagram

- Phase diagram for vapor diffusion experiment
 - In a vapor diffusion experiment where equal volumes of precipitant and protein are added in the drop, both the precipitant and protein concentration will double during equilibration.



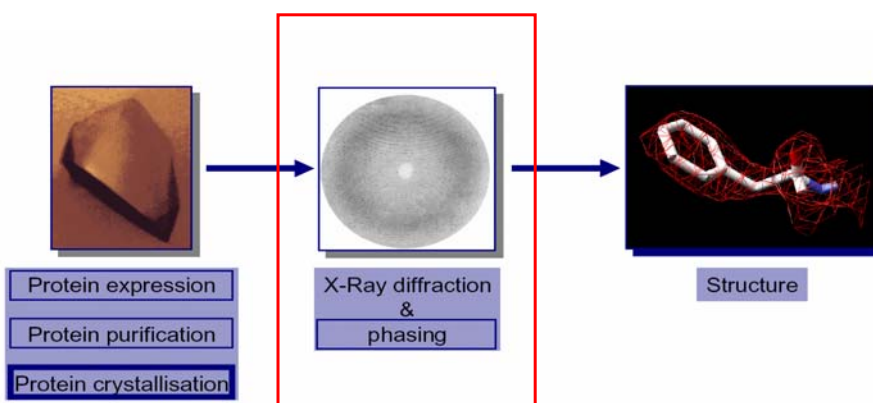
No crystals



Crystals



The steps towards the 3D structure

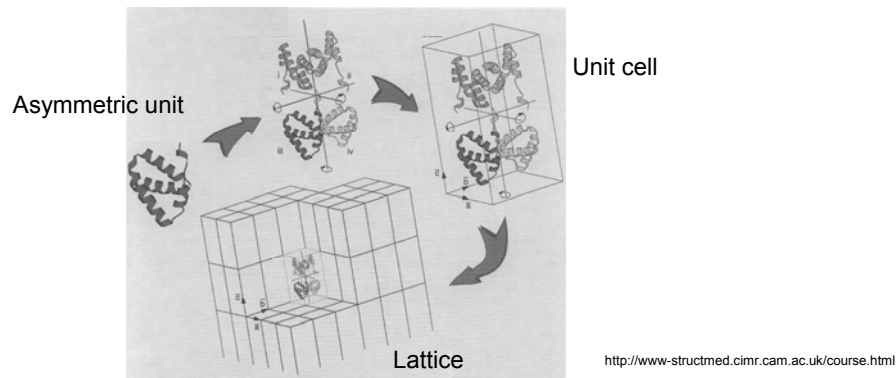


difficult



The single elements forming a crystal

1. Asymmetric unit is the smallest entity (molecule) of the crystal that has no symmetry
2. Applying symmetry operators and translation along the 3 axis (X,Y,Z) the unit cell is built
3. the side of the unit cells form the axis of the crystal (a, b, c, α , β , γ).

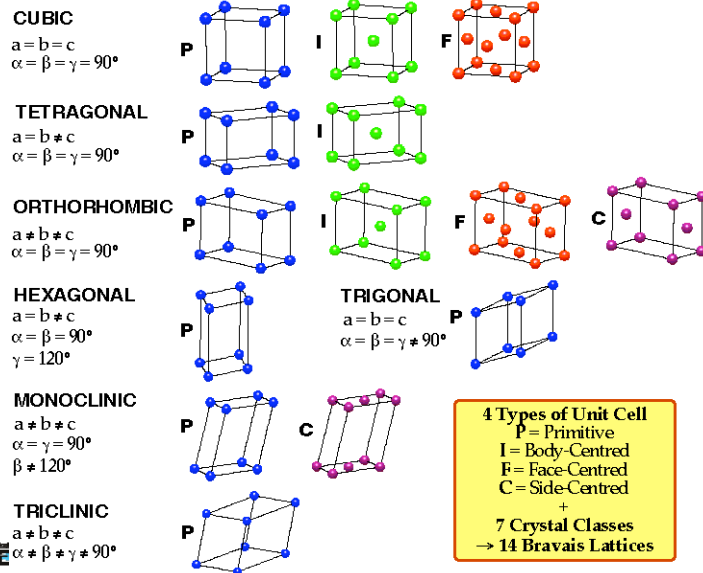


i From the paper: The asymmetric unit contains 8 monomers: What does it mean? ITÉ VE

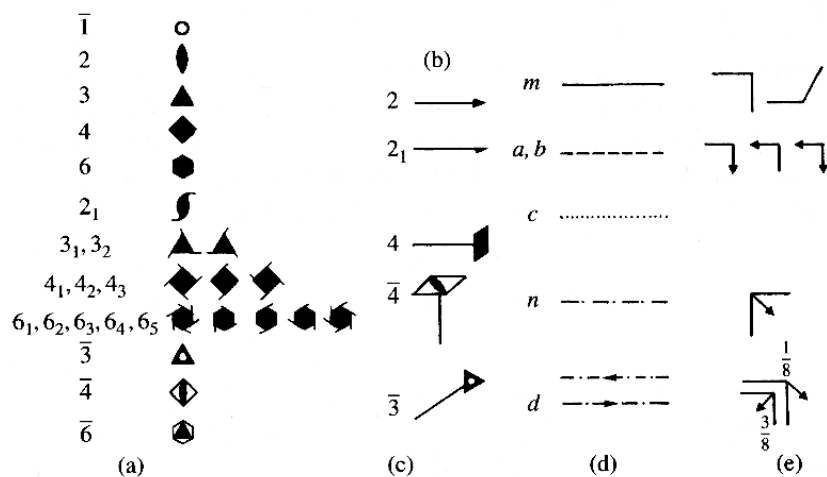
What does this sentence mean?

- The crystals belong to **space group C2** with $a = 157.5 \text{ \AA}$, $b = 122.9 \text{ \AA}$, $c = 115.3 \text{ \AA}$, $\alpha = \gamma = 90^\circ$ and $\beta = 130^\circ$.

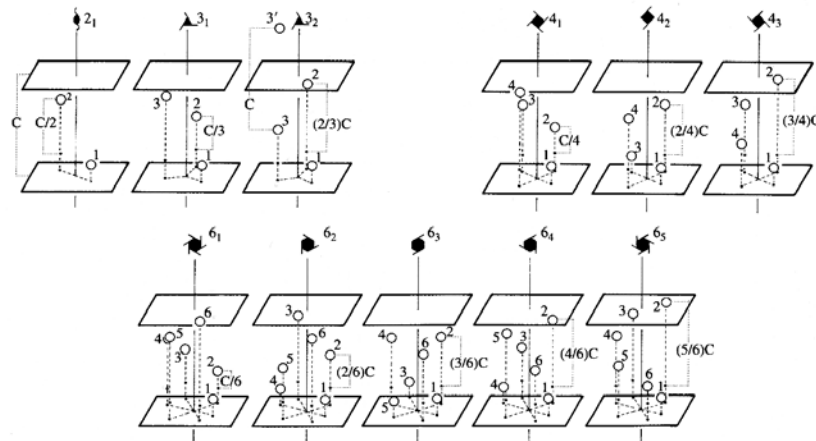
Crystals & Lattices: How to achieve the most favorable space filling



Symmetry (1)



Symmetry (2)



The combination of Bravais-lattice and symmetry lead to 230 space groups

- The combination of 14 Bravais lattices with 32 point groups and additional translational components such as screw axes and glide planes gives in total **230 space groups**.
- Of these, only **65 space groups** without mirror planes and inversion centers are possible for protein crystals.
- All space groups are described in the "International Tables for Crystallography, Vol. A".

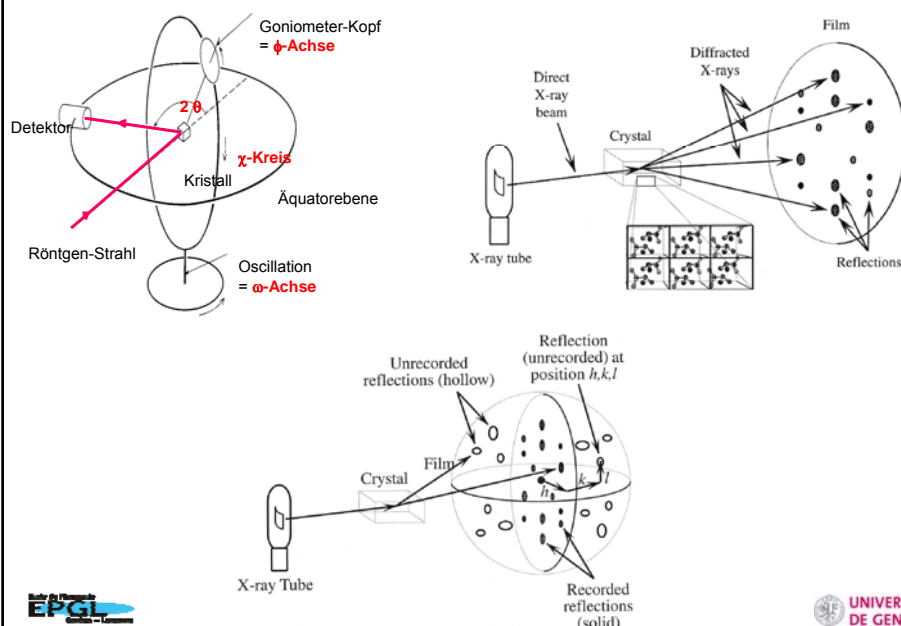
Data collection: SLS at PSI



EPGL

UNIVERSITÉ DE GENÈVE

Data collection and diffraction pattern

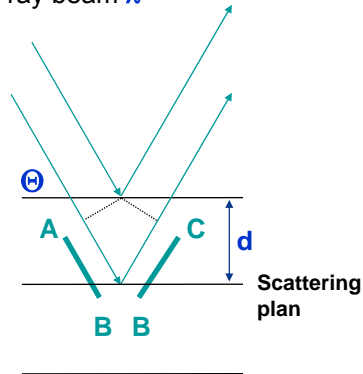


EPGL

UNIVERSITÉ DE GENÈVE

Diffraction of a lattice is visible only when both the von Laue conditions and Bragg's law are fulfilled. This results in diffraction spots called reflections.

X-ray beam λ



$$n\lambda = \overline{AB} + \overline{BC} \quad \text{da } \overline{AB} = \overline{BC}$$

$$n\lambda = 2 \overline{AB}$$

Bragg's Law: $n\lambda = 2d \sin\Theta$

n : is an integer,
 λ : is the wavelength of x-rays, and moving electrons, protons and neutrons,
 d : is the spacing between the planes in the atomic lattice, and
 Θ : is the angle between the incident ray and the scattering planes

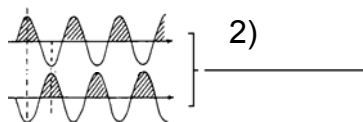
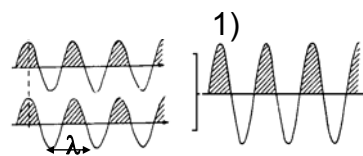
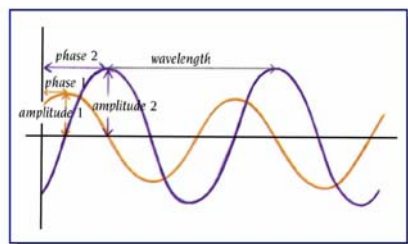
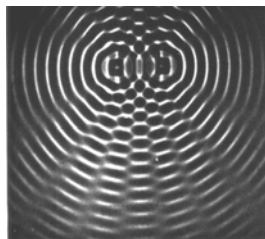


http://en.wikipedia.org/wiki/Bragg%27s_law



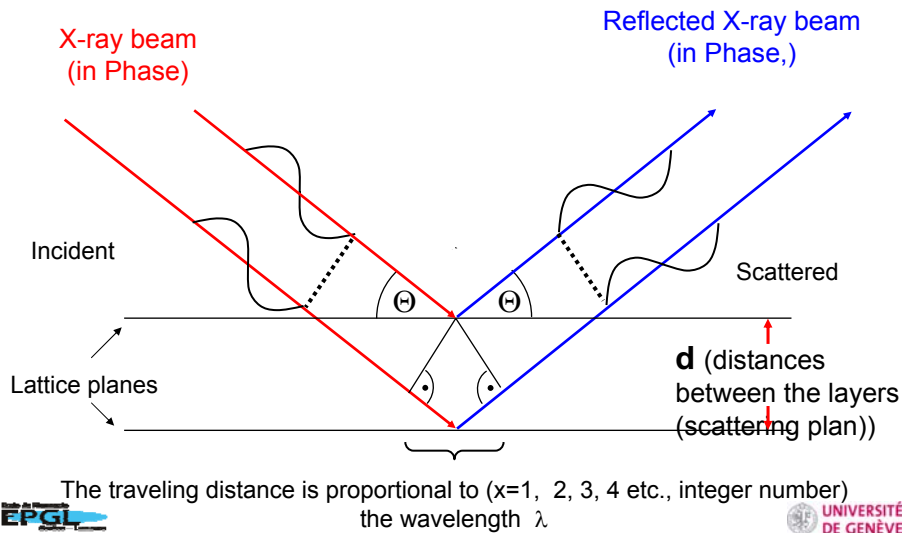
The X-ray diffraction pattern: Analogy with the waves formed in water

- 1) constructive Interference
- 2) destructive Interference

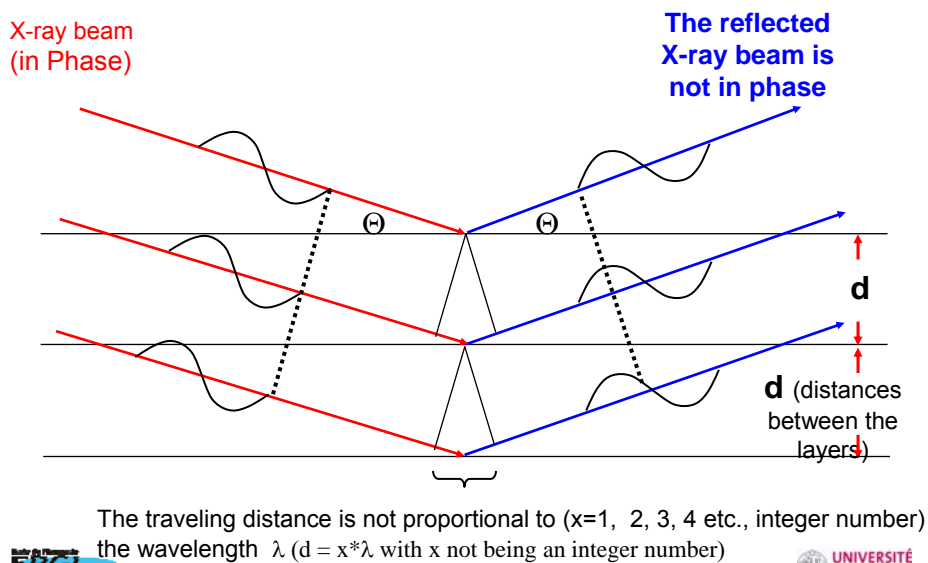


The crystal layers diffract the X-ray

constructive Interference

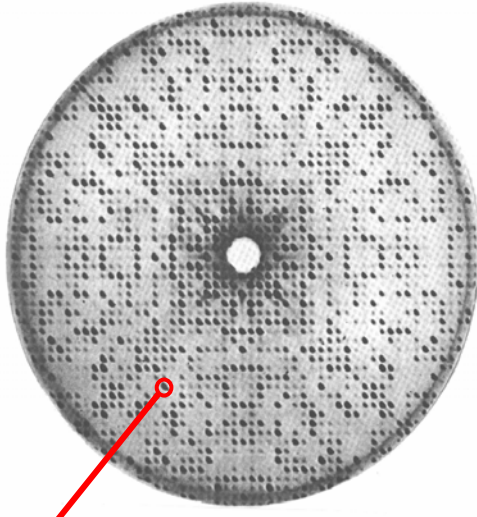


Destructive Interference



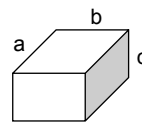
Pattern of diffraction of Lysozyme

(ca. 8500 Reflexes)



The pattern of diffraction allow direct determination of the unit cell and geometry (space group) :

- Symmetrie of the crystal: tetragonal
- Unit cell



$$a = b = 79 \text{ \AA}, c = 38 \text{ \AA}$$

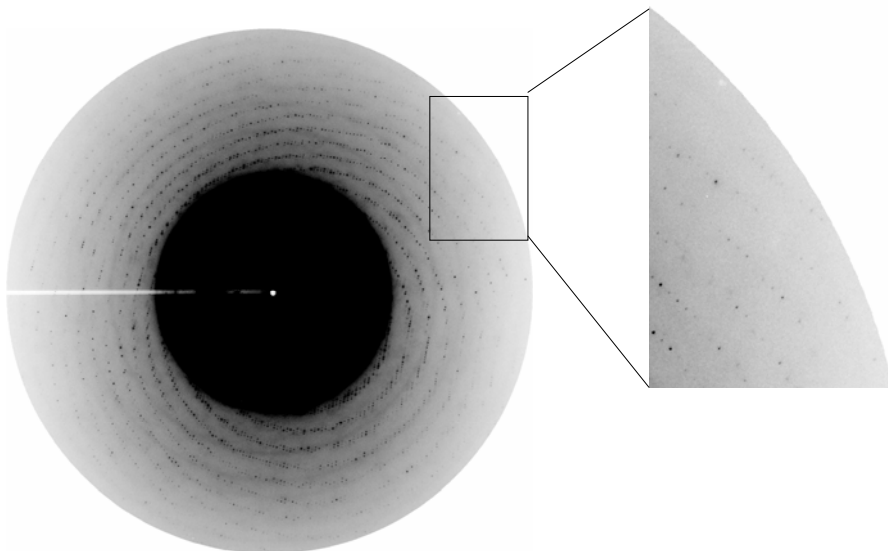
$$\alpha = \beta = \gamma = 90^\circ$$

- Space group $P4_32_12$
- Resolution 2Å

Intensity F is the result of constructive interference.

UNIVERSITÉ
DE GENÈVE

Estimation of the resolution



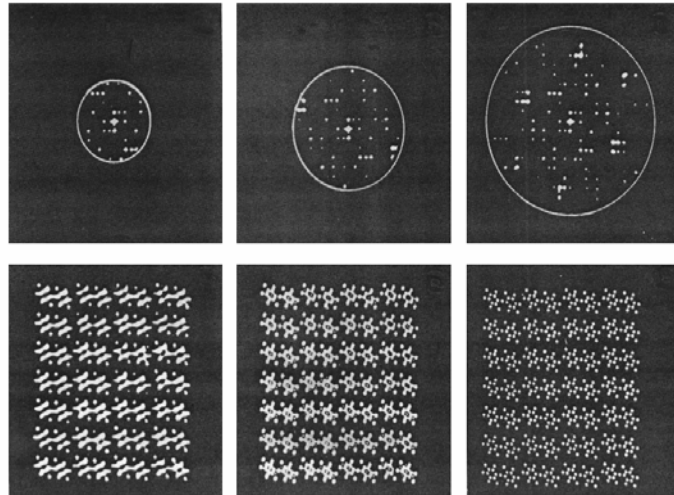
EPGL

UNIVERSITÉ
DE GENÈVE

The Resolution

$$\text{Resolution } d_{\min} = \frac{\lambda}{2 \sin \Theta_{\max}}$$

High resolution 1 - 2 Å
Low resolution 6 Å



EPGL

UNIVERSITÉ
DE GENÈVE

Data Processing Flow

- Indexing
- Integration
- Scaling
- Merging & Data Reduction
- Quality Indicators

EPGL

UNIVERSITÉ
DE GENÈVE

Statistical Quality indicators (in Table 1)

Completeness: $N_{\text{unique}}^{\text{obs}} / N_{\text{unique}}^{\text{theor}}$

Redundancy: $N_{\text{total}}^{\text{obs}} / N_{\text{unique}}^{\text{theor}}$

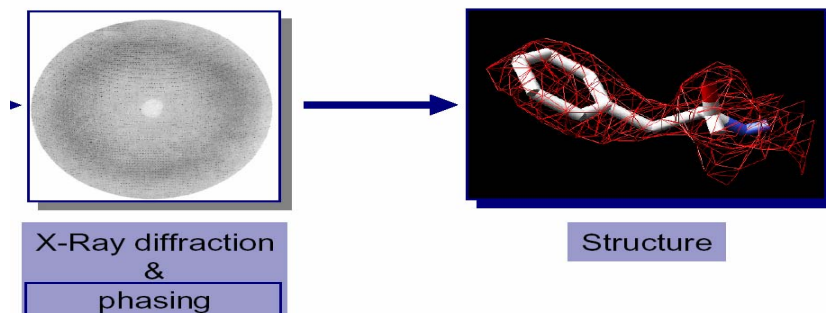
Signal-to-Noise: $\langle I / \sigma(I) \rangle$

$$R_{\text{sym}} = \frac{\sum_{hkl} \sum_i |I_i(hkl) - \langle I(hkl) \rangle|}{\sum_{hkl} \sum_i I_i(hkl)}$$

$$R_{\text{meas}} = \frac{\sum_{hkl} \sqrt{\frac{N}{N-1}} \sum_i |I_i(hkl) - \langle I(hkl) \rangle|}{\sum_{hkl} \sum_i I_i(hkl)}$$



From Pattern of Diffraction to Electron Density



The Phase Problem: Analysis of the diffraction pattern

Electron density $\rho(x,y,z)$ in the unit cell:

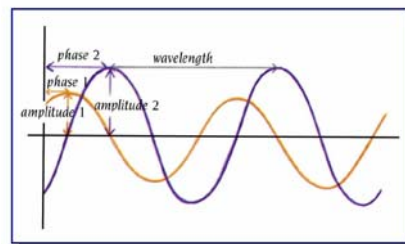
$$\rho(x,y,z) = \frac{1}{V} \sum_{hkl} F(hkl) e^{i\Phi(hkl) - 2\pi i(hx + ky + lz)}$$

- V: Volume of the unit cell
- $F(hkl)$: Intensity with Index hkl
- h,k,l : Index of diffraction (Miller indices): every combination h,k,l (z. 3,7,1) corresponds to planes in the atomic lattice
- i : Type of atom
- $\Phi(hkl)$: Phases (unknown from the data)



The phase problem

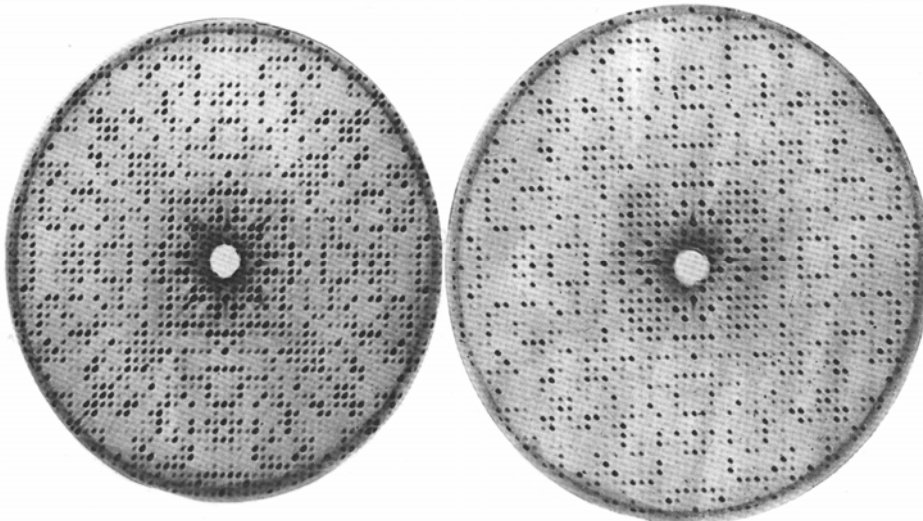
- Diffraction beam properties
 - Amplitude
 - Strength of beam
 - Intensity of recorded spot
 - Wavelength
 - Set by X-ray source
 - Phase
 - Interference of beams
 - Phases cannot be measured
 - Phase problem



Solving the phase problem

- Molecular replacement (MR)
 - Phases of a similar known protein structure (*phasing model*) used as estimate of phases of the unknown protein “Easy”
- Anomalous scattering (MAD; multiple wavelength anomalous diffraction)
- Multiple isomorphous replacement (MIR)
 - Introducing heavy metal ions as new X-ray scatters into crystal
 - Heavy atoms give strong signals and act as reference atoms
 - Should not change structure of protein thus Isomorphous structure
 - Cysteine groups (-SH) can bind metal
 - Replacing e.g. Zn by Hg
 - Heavy metals contain more electrons
 - Stronger scattering
 - Diffracted beams more, less, equally intense depending on proximity to heavy metal

MIR on Lysozyme



Lysozyme native

Lysozyme p-chlormercuribenzenesulphonat

Solving the phase problem

- Multiple isomorphous replacement (MIR)
 - Diffractions calculated from these ions only
 - Intensity differences used to identify positions of heavy atoms in unit cell
 - Mathematical calculations (Fourier transformation) lead to atomic arrangement in space for heavy atoms
 - From positions of heavy atoms in unit cell calculation of
 - Amplitude
 - Phase contribution to diffracted beam



Solving the phase problem

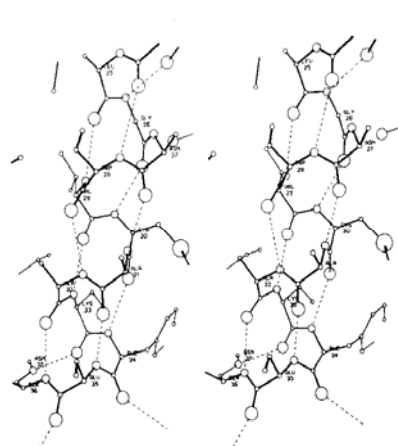
- How can we use this information to solve the phase problem of the protein?
 - Known
 - Amplitude and phase of heavy atoms
 - Amplitude of protein
 - Amplitude of protein + heavy metal
 - Unknown
 - Phase of protein
- Three amplitudes + one phase
- Calculation of interference of scattering between protein and heavy metal
 - Positive
 - Negative
- Estimation of protein phase
- In practice many such complexes are built to identify correct phase angle



Solution of the structure when phases have been determined

1. The first electron density map $\rho(x,y,z)$ is calculated by means of the Fourier transformation
2. A first model (mostly Alanine model) is built.
3. The introduction of the model gives more information about the phases $\alpha(hkl)$ and an iterative process of refinement in which a more and more complete model is built is started.

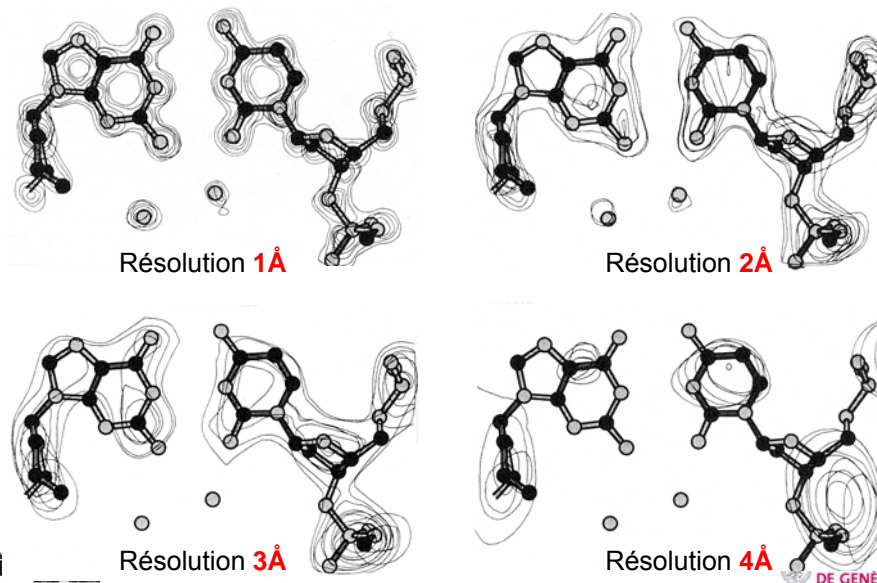
Lysozyme: From electron density to atoms coordinates



What are the parameters for assessing the quality of an X-ray structure?

- Resolution
- How well fit the built model into the experimentally measured density?
- Geometry

High resolution = more accurate details



How well fit the built model into the experimentally measured density?

- **R-factors:** There are different aspects to validation. Some types of validation look at the **fit to the diffraction data**. The agreement of observed and calculated structure factors is often measured with the traditional R-factor, which is the average fractional disagreement:

$$R = \Sigma(|F_o - F_c|) / \Sigma(F_o)$$

- As we have noted several times, it is possible to overfit the data, especially at moderate resolutions. This problem can be circumvented if you use most of the data (working set, 95% of data) to refine the atomic model, and the remaining data (test set, 5% of data) to verify how well you really are doing. **The test set data are used to compute R-free**, which is computed in the same way as the conventional R-factor but using only that subset of data. **If R-free drops, then the model must really have improved** because there is (almost) no pressure to overfit R-free. (The word "almost" will be explained in the advanced series.) This idea, called cross-validation in the statistical community, was introduced into crystallography by Axel Brünger and it has made a great contribution to keeping our models honest.



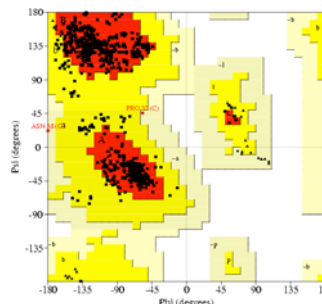
http://www-structmed.cimr.cam.ac.uk/Course/Basic_refinement/Refinement.html



Geometry

- **Quality of geometry:** Other types of validation are completely model based. One of the first entries was the program Procheck, which evaluates structures on various criteria: are the torsion angles (main chain and side chain) typical of those seen in high resolution protein structures? are the bond lengths, bond angles and van der Waals contacts consistent with the databases? One example of Procheck output is a Ramachandran plot (plot of main-chain torsion angles).

- ProCheck
- What_Check
- Rotamer
- MolProbity



http://www-structmed.cimr.cam.ac.uk/Course/Basic_refinement/Refinement.html



Geometry

- **Position of side chains:** A further type of information that can be used is residue environment preference. Some side chains (e.g. leucine, phenylalanine) are hydrophobic and tend to be buried in protein structures, surrounded by other hydrophobic side chains. Others (serine, asparagine) are hydrophilic and tend to be exposed on the surface or surrounded by other polar groups. If the sequence has become out of register, through an error in tracing the main chain, it can be detected by a series of amino acids being found in unfavourable environments. Such errors can be found by various threading programs, including the Profile program from Eisenberg's group.



http://www-structmed.cimr.cam.ac.uk/Course/Basic_refinement/Refinement.html



Quality indicators

- A well refined crystal structure should have:
 - R-factor < 0.20, Free-R < 0.27
 - RMSD bond lengths < 0.02 Å, bond angles < 2°
 - Only a few outliers in the Ramachandran plot
 - No large deviation from ideal stereochemistry without a good reason
 - Water molecules with reasonable hydrogen bonds and B-factor



Acknowledgement

- Dr. Remo Perozzo (Pharm. Biochemistry, Geneva)
- Dr. Dirk Kostrewa (Structural Biology Group, PSI, Villigen)