Strukturelle Modellierung
(Masterstudiengang Bioinformatik)

Strukturbestimmung mit Röntgenkristallographie

Sommersemester 2013

Peter Güntert

Introduction

Myoglobin Struktur

"*Vielleicht die bemerkenswerteste Eigenschaft des Moleküls ist seine Komplexität und die Abwesenheit von Symmetrie. Der Anordnung scheinen die Regelmässigkeiten, die man instinktiv erwartet, fast völlig zu fehlen, und sie ist komplizierter als von irgendeiner Theorie der Proteinstruktur vorhergesagt.*" — John Kendrew, 1958

Kristallographie: Geschichte

1839, William H. Miller: Miller Indices für Gitterebenen
1891: 230 Raumgruppen für Kristalle
1895, Wilhelm Conrad Röntgen: Röntgenstrahlung
1912, Max von Laue: Röntgenstreuung
1912, William L. Bragg: Braggsches Gesetz
1914, Bragg: Kristallstrukturen von NaCl und Diamant
1937: Dorothy Hodgkin: Kristallstruktur von Cholesterin
1945: Dorothy Hodgkin: Kristallstruktur von Vitamin B12
1952: Rosalind Franklin: DNA Röntgenbeugungsdiagramme
1955: Rosalind Franklin: Tabakmosaikvirus (TMV) Struktur
1958: John Kendrew: Erste Proteinstruktur (Myoglobin)
2000: Kristallstruktur des Ribosoms
2013: > 79'000 Kristallstrukturen in der Protein Data Bank

Literatur über Kristallstrukturbestimmung

- B. Rupp, *Biomolecular Crystallography*, Garland, 2010.
- W. Massa, *Kristallstrukturbestimmung*, Teubner, ^{5}2007.
- C. Branden & J. Tooze, *Introduction to Protein Structure*, Garland, ^{2}1999.

Crystallographic structure models versus proteins in solution

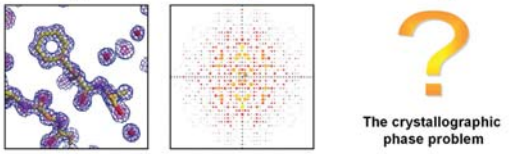
- Protein crystals are formed by a loose periodic network of weak, non-covalent interactions.
- Protein crystals contain large solvent channels. The solvent channels allow relatively free diffusion of small molecules through the crystal and also provide conformational freedom for surface-exposed side chains or loops.
- The core structure of protein molecules in solution as determined by NMR is identical to the crystal structure. Even enzymes generally maintain activity in protein crystals.
- Crystal packing can affect local regions of the structure where surface-exposed side chains or flexible surface loops form intermolecular crystal contacts.
- Large conformational movements destroy crystals and cannot be directly observed though a single crystal structure. Limited information about the dynamic behavior of molecules can be obtained from analysis of the *B*-factors as a measure of local displacement.
- The quality of a protein structure is a local property. Surface-exposed residues or mobile loops may not be traceable in electron density, no matter how well defined the rest of the structure is.

Challenges of protein cystallography

- Proteins are generally difficult to crystallize and without crystals there is no crystallography. Preparing the material and modifying the protein by protein engineering so that it can actually crystallize is nontrivial.
- Prevention of radiation damage by ionizing X-ray radiation requires cryocooling of crystals and many crystals are difficult to flash-cool.
- The X-ray diffraction patterns do not provide a direct image of the molecular structure. The electron density of the scattering molecular structure must be reconstructed by Fourier transform techniques.
- Both structure factor amplitude and relative phase angle of each reflection are required for the Fourier reconstruction. While the structure factor amplitudes are readily accessible being proportional to the square root of the measured reflection intensities, the relative phase angles must be supplied by additional phasing experiments. The absence of directly accessible phases constitutes the phase problem in crystallography.
- The nonlinear refinement of the structure model is nontrivial and prior stereochemical knowledge must generally be incorporated into the restrained refinement.

The crystallographic phase problem

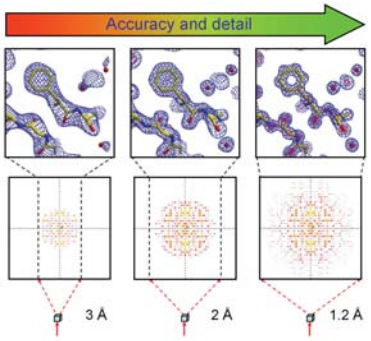
$$\rho(x,y,z) = \frac{1}{V}\sum_{-h}^{h}\sum_{-k}^{k}\sum_{-l}^{l} F_{hkl} \cdot \exp[-2\pi i(hx+ky+lz+\alpha_{hkl})]$$



The crystallographic phase problem

In order to reconstruct the electron density of the molecule, two quantities need to be provided for each reflection (data point): the structure factor amplitude, F_{hkl}, which is directly obtained through the experiment and is proportional to the square root of the measured intensity of the diffraction spot or reflection; and the phase angle of each reflection, α_{hkl}, which is not directly observable and must be supplied by additional phasing experiments.

Data quality determines structural detail and accuracy



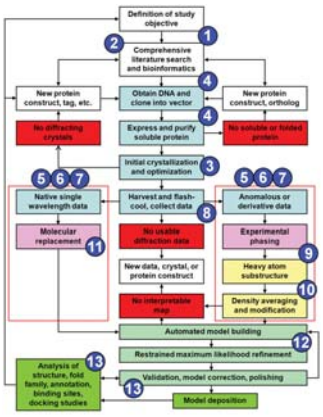
The qualitative relation between the extent of X-ray diffraction, the resulting amount of available diffraction data, and the quality and detail of the electron density reconstruction and protein structure model are evident from this figure: The crystals are labeled with the nominal resolution d_{min} given in Å (Ångström) and determined by the highest diffraction angle (corresponding to the closest sampling distance in the crystal, thus d_{min}) at which X-ray reflections are observed. Above each crystal is a sketch of the corresponding diffraction pattern, which contains significantly more data at higher resolution, corresponding to a smaller distance between discernable objects of approximately *dmin*. As a consequence, both the reconstruction of the electron density (blue grid) and the resulting structure model (stick model) are much more detailed and accurate.

Kristallstrukturbestimmung

1. Proteinherstellung
2. Kristallisation
3. Messung der Beugungsmuster
4. Datenauswertung
   a) Bestimmung der Einheitszelle und Raumgruppe
   b) Phasenbestimmung
   c) Modellbau
   d) Verfeinerung der Phasen und der Struktur

Key stages in X-ray structure determination

The flow diagram provides an overview about the major steps in a structure determination project, labeled with the chapter numbers treating the subject or related general fundamentals. Blue shaded boxes indicate experimental laboratory work, while all steps past data collection are conducted *in silico*.



Crystallographic computer programs

- Protein crystallography depends heavily on computational methods.
- Crystallographic computing has made substantial progress, largely as a result of abundant and cheap high performance computing.
- It is now possible to determine and analyze complex crystal structures entirely on inexpensive laptop or desktop computers with a few GB of memory. Automation and user interfaces have reached a high level of sophistication (although compatibility and integration issues remain).
- As a result, the actual process of structure solution, although the theoretically most sophisticated part in a structure determination, is commonly not considered a bottleneck in routine structure determination projects.
- Given reliable data of decent resolution (~2.5 Å or better) and no overly large or complex molecules, many structures can in fact be solved de novo and refined (although probably not completely polished) within several hours.
- Automated model building programs—many of them available as web services—have removed much of the tedium of initial model building.

Key concepts of protein cystallography I

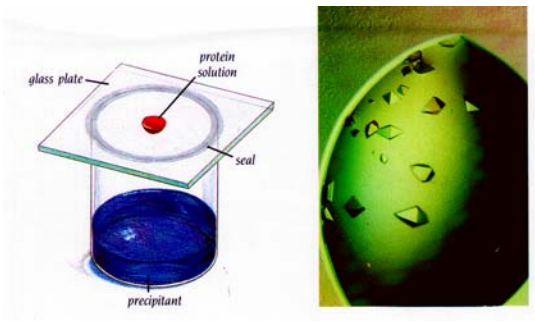
- The power of macromolecular crystallography lies in the fact that highly accurate models of large molecular structures and molecular complexes can be determined at often near atomic level of detail.
- Crystallographic structure models have provided insight into molecular form and function, and provide the basis for structural biology and structure guided drug discovery.
- Non-proprietary protein structure models are made available to the public by deposition in the Protein Data Bank, which holds more than 83 000 entries as of April 2013.
- Proteins are generally difficult to crystallize; without crystals there is no crystallography.
- Preparing the material and modifying the protein by protein engineering so that it can actually crystallize is nontrivial.
- Radiation damage by ionizing X-ray radiation requires cryocooling of crystals, and many crystals are difficult to flash-cool.

Key concepts of protein cystallography II

- The X-ray diffraction patterns are not a direct image of the molecular structure.
- The electron density of the scattering molecular structure must be reconstructed by Fourier transform techniques.
- Both structure factor amplitude and relative phase angle of reach reflection are required for the Fourier reconstruction.
- While the structure factor amplitudes are readily accessible, being proportional to the square root of the measured reflection intensities, the relative phase angles must be supplied by additional phasing experiments.
- The absence of directly accessible phases constitutes the phase problem in crystallography.
- The nonlinear refinement of the structure model is nontrivial and prior stereochemical knowledge must generally be incorporated into the restrained refinement.

Crystallization

Proteinkristallisation



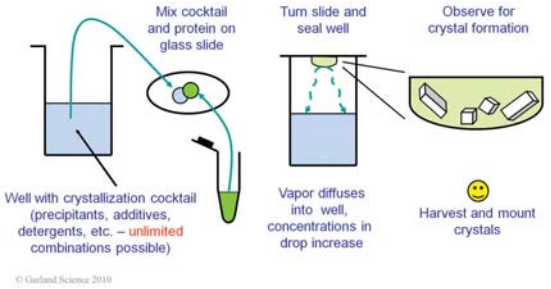
Protein crystallization basics

- Protein crystals are periodic self-assemblies of large and often flexible macromolecules, held together by weak intermolecular interactions. Protein crystals are generally fragile and sensitive to environmental changes.
- In order to form crystals, the protein solution must become supersaturated. In the supersaturated, thermodynamically metastable state, nucleation can occur and crystals may form while the solution equilibrates.
- The most common technique for protein crystal growth is by vapor diffusion, where water vapor equilibrates from a drop containing protein and a precipitant into a larger reservoir with higher precipitant concentration.
- Given the large size and inherent flexibility of most protein molecules combined with the complex nature of their intermolecular interactions, crystal formation is an inherently unlikely process, and many trials may be necessary to obtain well-diffracting crystals.

The protein is the most crucial factor in determining crystallization success

- Given that a crystal can only form if specific interactions between molecules can occur in an orderly fashion, the inherent properties of the protein itself are the primary factors determining whether crystallization can occur.
- A single-residue mutation can make all the difference between successful crystallization and complete failure.
- Important factors related to the protein that influence crystallization are its purity, the homogeneity of its conformational state, the freshness of the protein, and the additional components that are invariably present, but often unknown or unspecified, in the protein stock solution.

Hanging drop vapor diffusion



Solubility phase diagram

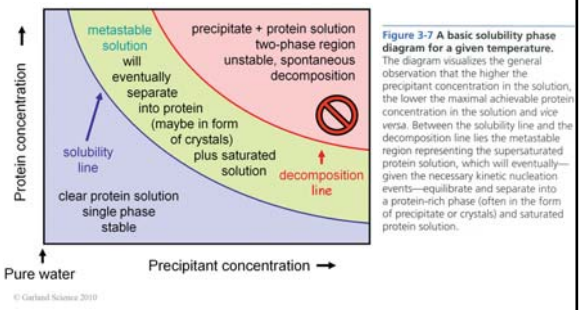


Figure 3-7 A basic solubility phase diagram for a given temperature. The diagram visualizes the general observation that the higher the precipitant concentration in the solution, the lower the maximal achievable protein concentration in the solution and *vice versa*. Between the solubility line and the decomposition line lies the metastable region representing the supersaturated protein solution, which will eventually—given the necessary kinetic nucleation events—equilibrate and separate into a protein-rich phase (often in the form of precipitate or crystals) and saturated protein solution.

Protein solubility versus pH

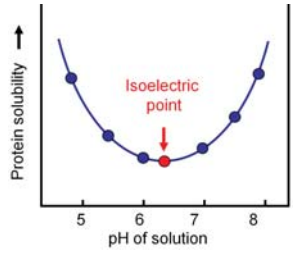


Figure 3-8 Protein solubility versus pH of protein solution. The protein shown in this example has its solubility minimum at its isoelectric point of ~6.3, where the sum of positive and negative charges (the net charge of the protein) is zero. Even at the isoelectric point, there are still numerous (but net compensating) local charges present on the surface of the protein.

Crystal growth

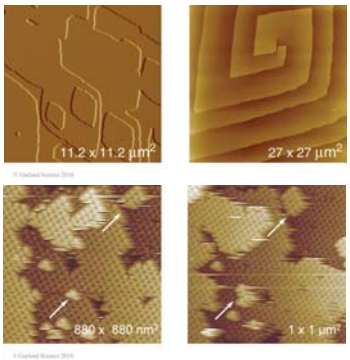


Figure 3-11 Atomic force microscope images of crystal growth. (Panel A) The atomic force microscope images of the 001 surface of glucose isomerase show the two most common growth patterns observed in crystal growth: step growth starting from 2-dimensional nucleation islands (A, left image) and a double-spiral growth pattern (A, right image). Panel B shows formation of supercritical 2-dimensional nuclei on the 001 surface of cytomegalovirus (CMV), a member of the herpes virus family. As indicated by the arrows, in this case only two virions (B, left image) suffice to generate a critical nucleus from which new step growth commences (B, right image). Images courtesy of Alexander McPherson and Aaron Greenwood, University of California, Irvine.

Mosaic crystals

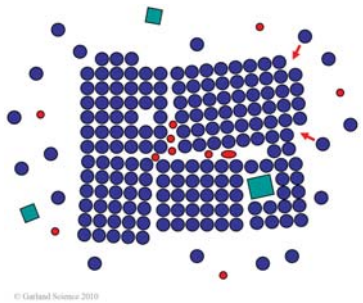


Figure 3-12 Growth of a real mosaic crystal. The schematic drawing shows a crystal growing in a solution of protein molecules (blue spheres). Small impurities (red) and some larger detritus (green squares) are also present in the solution. New molecules attach preferentially to steps and edges (red arrows) and we can recognize a growth defect in the form of a hole; impurities are enclosed at the domain boundaries; and a larger piece of detritus is incorporated at a domain boundary. Individual domains can be substantially misaligned, in this case about 6°; such a highly mosaic crystal would not be useful for diffraction experiments.

Crystallization techniques

- The inability to predict *ab initio* any conditions favoring protein crystallization means that, in general, several hundred crystallization trials must be set up in a suitable format and design.
- Crystallization screening experiments are commonly set up manually or robotically in multi-well format crystallization plates.
- The most common procedure for achieving supersaturation is the vapor-diffusion technique, performed in sitting-drop or hanging-drop format. In vapor-diffusion setups, protein is mixed with a precipitant cocktail, and the system is closed over a reservoir into which water vapor diffuses from the protein solution. During vapor diffusion, both precipitant and protein concentration increase in the crystallization drop and supersaturation is achieved.
- As a rule of thumb, low supersaturation favors controlled crystal growth, while high supersaturation is required for spontaneous nucleation of crystallization nuclei. Seeding is a method to induce heterogeneous nucleation at low supersaturation, which is more conducive to controlled crystal growth.

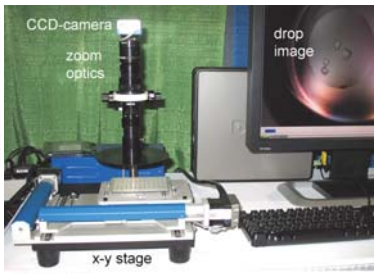
Robot for automated crystallization

Sidebar 3-13 Automated crystallization setup for the small laboratory. Based on the assumption of modest throughput requirements, and no necessity for full walk-away automation, two low-budget approaches to automation are conceivable: selection of a single system that can prepare crystallization cocktails (perhaps in a limited fashion) and also set up the crystallization plates,[45] or a dual-station layout using separate cocktail preparation with a generic liquid-handling system followed by a dedicated plate-setup robot.[46] The major reason for separating plate setup from cocktail production is differing requirements for dispensing precision, volume, and speed. Fast, small volume (µl to nl), and very accurate (also in geometric terms) dispensing is mandatory for plate crystallization setup, whereas large volume (ml) handling with modest speed and precision requirements suffices for cocktail production. Another advantage of the separation between the cocktail stage and the plate setup is that simple one-to-one dispensing into reservoir wells and drop aliquots followed by protein addition with a single needle dispenser suffices (Figure 3-33) once the cocktails are produced in a 96-well format deep-well block. Deep-well blocks prefilled with crystallization cocktails are also commercially available. In addition, compared with a single-stage setup, failure of one system component does not affect the other. For example, cocktail production can continue while the plate setup robot is inoperative. Figure 3-33 shows a popular robot for 96-well crystallization plate setup.

Figure 3-33 A robot for automated crystallization plate setup. The Phoenix robot (Art Robbins Instruments) can set up 96 crystallization trials in about one minute. On the left side, a 96-channel syringe dispenser re-arrays (100 µl each) 96 prefabricated or purchased crystallization cocktails simultaneously from a standard deep-well block into the reservoirs of an SBS-format, 96-well sitting-drop crystallization plate, and places between 1 µl and 100 nl into the drop shelves or wells. From the right side, a contact-less microvalve dispenser nozzle immediately adds the pre-aspirated protein (stock vials in the red block) rapidly and without contact onto each of the precipitant drops. To minimize evaporation, the plate is then immediately sealed with a sheet of pressure-sensitive adhesive. Taking all losses into account, about 12 to 15 µl of protein stock is required for 96 100 + 100 nl drops. The robot design has been based on a prototype developed in an academic laboratory setting.[46]

Cystallization plate imaging

Figure 3-36 A low-cost automated crystallization plate imaging station. The crystallization plate is positioned by an x-y translation stage, and a digital zoom camera takes high-resolution images of the crystallization drops. The images taken in about 2 minutes can then be manually inspected on a computer screen, or processed by automated image recognition software. The depicted instrument is the CrysCam microscope manufactured by Art Robbins Instruments.



Crystallization outcomes

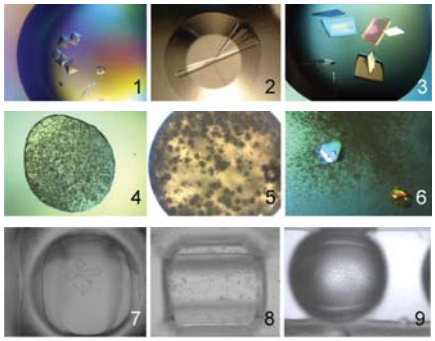


Figure 3-37 Images of crystallization drops with different experimental outcomes. (1) Perfectly formed octahedral single crystals in hanging drop; (2) a cluster of large needles in microbatch reservoir; (3) thin plates in a hanging drop; (4) microcrystal shower in hanging drop; (5) small crystals together with spherical microcrystal clusters; (6) single crystals growing from a grainy precipitate; (7) irregular dendritic growth of crystals in sitting drop; (8) grainy precipitate in sitting drop; (9) amorphous precipitate and protein "oil" in sitting drop. The black and white pictures have been taken by automated crystal imaging stations. The false colors apparent in images 1 to 6 result from polarization effects in the optically anisotropic (birefringent) protein crystals and the varying depolarization in the injection-molded plastic. The proteins (and their crystals shown in this figure) are actually colorless.

Heavy atom derivatives

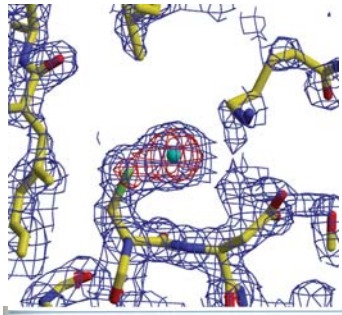


Figure 3-42 Heavy atom derivatization of a protein. Shown is the electron density around a gold atom covalently linked to a cysteine residue in the *Clostridium tetani* neurotoxin.[123] A combination of anomalous and isomorphous signals from gold atoms were used to solve the structure of the ganglioside binding domain of the neurotoxin from bacillus *C. tetani*, the causative agent of tetanus infections. PDB entry 1a8d.

Heavy atom reagents

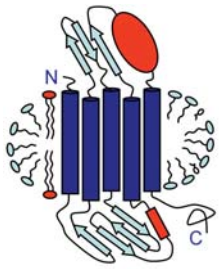
Name	Formula
Platinum potassium chloride, potassium tetrachloroplatinate(II)	K_2PtCl_4
Aurous potassium cyanide, potassium dicyanoaurate(I)	$KAu(CN)_2$
Mercuric potassium iodide, potassium tetraiodo mercurate(II)	K_2HgI_4
Uranyl acetate, uranium(VI) oxyacetate	$UO_2(C_2H_3O_2)_2$
Mercuric(II) chloride	$HgCl_2$
Potassium uranyl fluoride, potassium uranium(VI) oxyfluoride	$K_2UO_2F_5$
Para-chloromercurobenzosulfonate, PCMBS	$Hg(C_6H_4)SO_4$
Trimethyllead acetate	$(CH_3)_3Pb(CH_3COO)$
Methylmercuric acetate	$CH_3Hg(CH_3COO)$
Ethylmercuric thiosalicylate, thiomersal	$C_2H_5HgSC_6H_4COO^-$
Hexatantalum tetradecabromide	$(Ta_6Br_{12})Br$

Table 3-1 Selected heavy atom reagents. The listed reagents are frequently used for derivatization. The top seven entries are historically the most well used, the alkylated compounds below and the powerful Ta-clusters are more recent and very successful derivatization reagents. Many more are listed in the heavy atom data bank[116] and in the review by M.A. Rould[117]. All these substances are quite toxic when ingested because they bind to proteins and taking corresponding precautions is prudent. The uranium salts are generally prepared from natural uranium (0.7% ^{235}U) or depleted uranium (^{238}U), which both are only a weak α-particle source.

Less than 1% of all deposited protein structures are membrane protein structures

- About a third of all expressed human proteins are presumed to be membrane proteins, and over 60% of all current drug targets are membrane receptors. Their primary functions include transport of material and signals across cell membranes as well as motor functions.
- Despite membrane proteins being a significant class of proteins, it was nearly 30 years, and 195 deposited protein structures, after Kendrew's first myoglobin structure in 1958 that the first integral membrane protein structure, the photosynthetic reaction center isolated from the bacterium *Rhoda pseudomonas viridis,* was published in 1985. That research led to a Nobel Prize for crystallographic work being awarded to Johann Deisenhofer, Hartmut Michel, and Robert Huber in 1988.
- In early 2007, there were 242 coordinate entries of 122 different membrane proteins out of 35100 total entries in the PDB, still a factor of 1/145 disfavoring the membrane proteins. Clearly, membrane protein crystallization remains a major challenge for crystallography.

Resolubilized membrane protein

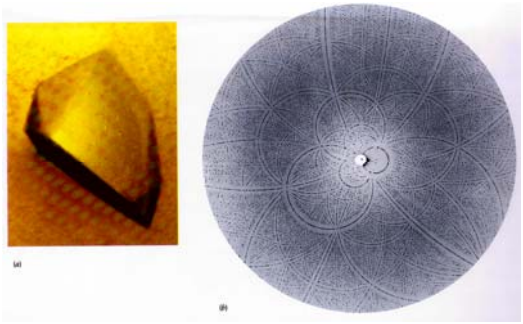


Membrane phospholipid
Detergent
Amphiphile

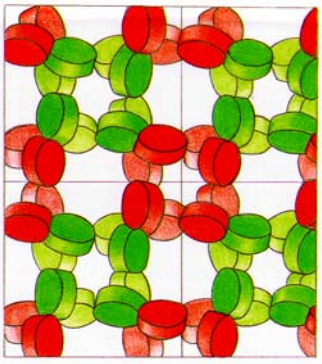
Figure 3-43 Resolubilized multi-pass, polytopic transmembrane protein with its associated detergent collar. In addition to the detergent collar, membrane fragments are often associated and co-solubilized with the transmembrane stem, as sketched on the left side of the membrane collar. Small amphiphile molecules are often added to fine-tune the size of the membrane collar for subsequent crystallization, as shown at the right side of the membrane collar.

Crystals

Kristall und Beugungsmuster



Proteinkristall



Unit lattice + Motif = Unit cell

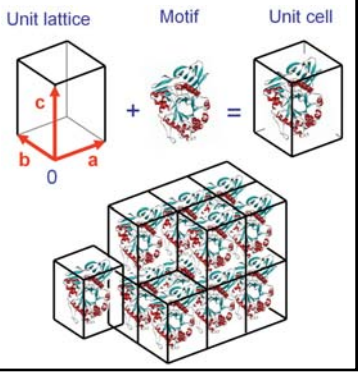


Figure 5-24 Assembly of a primitive triclinic 3-dimensional crystal from unit cells. In analogy to the 2-dimensional case, the unit lattice is filled with a motif, and the crystal is built from translationally stacked unit cells. The basis vectors form a right-handed system [0, **a**, **b**, **c**].

Unit cell parameters

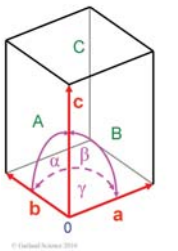
The three basis vectors of a unit lattice [0, **a**, **b**, **c**] extend from a common origin in a right-handed system; that is, if going counterclockwise from basis vector **a** to basis vector **b**, the third basis vector **c** points upwards (Figure 5-25). The vector product $\mathbf{a} \times \mathbf{b}$ generates a third vector **c** perpendicular to **a** and **b**, and the vector product $\mathbf{a} \times \mathbf{b}$ is *positive defined* in a right-handed system. The magnitude of this vector, $|\mathbf{a} \times \mathbf{b}|$, is equal to the area spanned by the vectors **a** and **b**. The unit cell volume V_{uc} is given by the triple vector product, $V_{uc} = \mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})$.

The angle between **a** and **b** is γ, the angle between **b** and **c** is α, and the angle between **a** and **c** is β. Similarly, the plane spanned by **a** and **b** is denoted as *C*, the plane between **b** and **c** is *A*, and the plane between **a** and **c** is labeled *B*.

The length of a unit cell vector is given by its norm: $|\mathbf{a}| = a$, $|\mathbf{b}| = b$, and $|\mathbf{c}| = c$.

The cell dimensions and angles are the six cell parameters (or cell constants) a, b, c, α, β, and γ.

Right-handed unit lattice



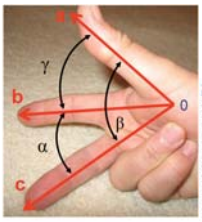


Figure 5-25 Right-handed, 3-dimensional unit lattice. A 3-dimensional unit cell is shown with its unit vectors, angles, and faces assigned in standard crystallographic notation. The angles and faces between two axes are annotated with the remaining complementary letter, for example, vectors **a** and **b** enclose angle γ and span face C, and so forth. In mathematical terms, the unit cell is a parallelepiped: a generic, 3-dimensional body formed by three pairs of parallel planes.

The 6 primitive 3D lattices

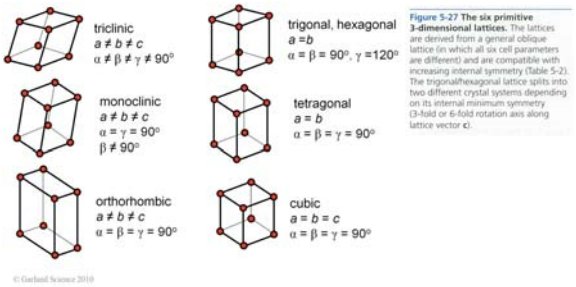


Figure 5-27 The six primitive 3-dimensional lattices. The lattices are derived from a general oblique lattice (in which all six cell parameters are different) and are compatible with increasing internal symmetry (Table 5-2). The trigonal/hexagonal lattice splits into two different crystal systems depending on its internal minimum symmetry (3-fold or 6-fold rotation axis along lattice vector **c**).

Centered 3D Bravais lattices

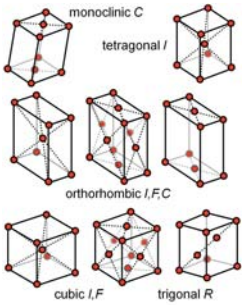
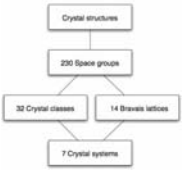


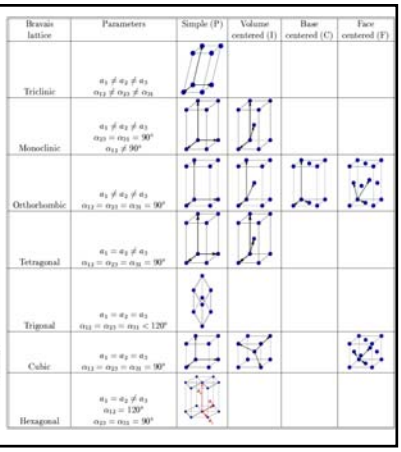
Figure 5-30 Centered 3-dimensional Bravais lattices. In addition to the six primitive 3-dimensional lattices, centered Bravais lattices are derived from the primitive lattices by translational centering. The necessity for additional internal translational symmetry within the unit cells limits the number of combinations to eight. Together, there are thus 14 Bravais lattices. Lattice points located in the rear of the cells are shown with lighter borders for increased clarity.

ox 5-8 The 14 three-dimensional Bravais lattices belong to seven crystal ystems. In 3-dimensional space, the combination of internal lattice transla- ions together with the six basic translational lattices leads to 14 Bravais lattices. hese lattices fall into seven crystal systems, which are defined by their minimal nternal symmetry. Lattice types and crystal systems are listed in Table 5-2.

Kristallsysteme

Bravaisgitter

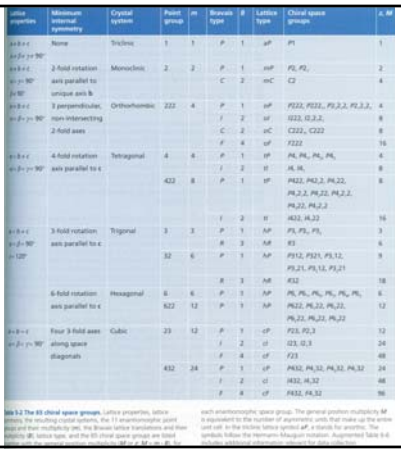




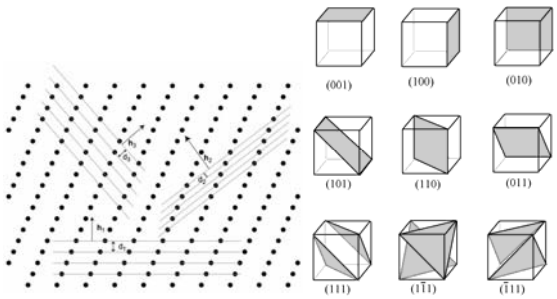
Protein crystals belong to one of 65 space groups

Only 65 discrete and distinct ways exist to assemble 3-dimensional periodic crystals from asymmetric chiral molecules, through combinations of translational and rotational symmetry. These 65 types of arrangements form 65 chiral space groups, and their symmetry properties and the rules for constructing each crystal structure are described in the *International Tables for Crystallography, Volume A*.

The 65 chiral space groups

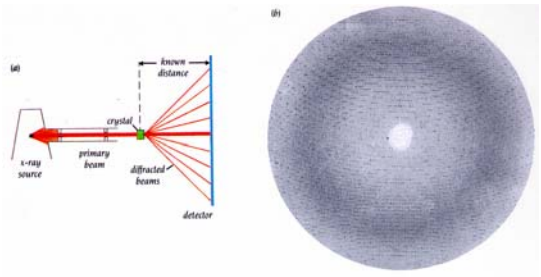


Miller Indizes



X-ray diffraction

Röntgenkristallographie Messung



Laboratory X-ray diffractometer

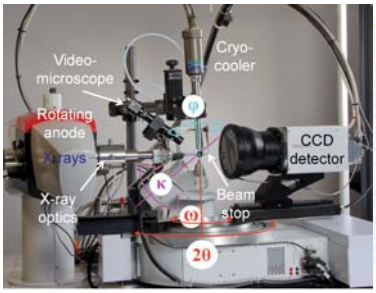
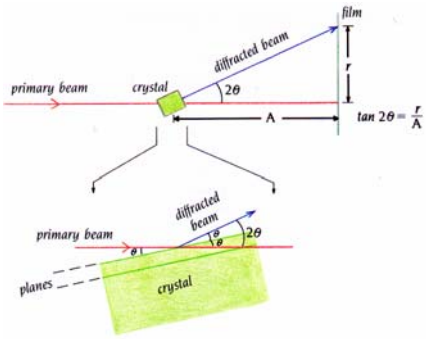


Figure 8-1 A contemporary laboratory X-ray diffractometer for macromolecular crystallography. A rotating anode X-ray source is closely coupled with integrated focusing optics delivering high photon flux at low operating power. In the center of the diffractometer is a full 4-circle κ-goniostat for orienting and rotating the crystal in multiple positions in the X-ray beam, thus enabling redundant data collection and in-house S-SAD phasing experiments. The CCD area detector is located to the right, and the diffractometer is also equipped with a cryocooler and a video microscope. The 2θ- and the ω-axis are collinear, with 2θ the detector offset angle. Image courtesy Matt Benning, Bruker AXS.

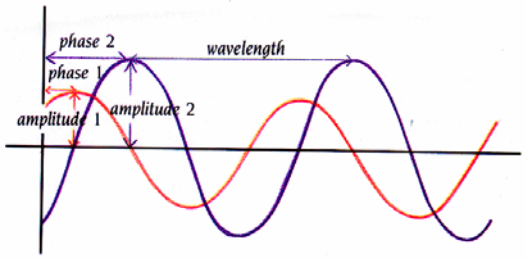
Synchrotron

ESRF Grenoble (France)

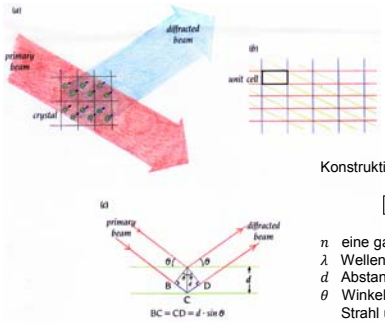
Röntgenstreuung



Superposition of two waves



Röntgenstreuung: Bragg-Bedingung



Konstruktive Interferenz, falls

$$\boxed{n\lambda = 2d\,\sin\theta}$$

- n eine ganze Zahl
- λ Wellenlänge
- d Abstand der Gitterebenen
- θ Winkel zwischen einfallendem Strahl und den Gitterebenen

Fourier transform

Fourier transform relates structure factors and electron density

$$F(\boldsymbol{k}) = \int_R \rho(\boldsymbol{r})\, e^{2\pi i \boldsymbol{r}\cdot\boldsymbol{k}} d\boldsymbol{r}$$

$$\rho(\boldsymbol{r}) = \int_{R^*} \boldsymbol{F}(\boldsymbol{k})\, e^{-2\pi i \boldsymbol{r}\cdot\boldsymbol{k}} d\boldsymbol{k}$$

- $\rho(\boldsymbol{r})$ electron density at position $\boldsymbol{r}$ in real space R
  $\rho(\boldsymbol{r}) \in \mathbb{R}$ is real
- $F(\boldsymbol{k})$ structure factor at position $\boldsymbol{k}$ in reciprocal space R^*
  $F(\boldsymbol{k}) \in \mathbb{C}$ is complex with (measurable) amplitude $|F(\boldsymbol{k})|$ and (not measurable) phase $\alpha(\boldsymbol{k})$, i.e. $F(\boldsymbol{k}) = |F(\boldsymbol{k})|e^{i\alpha(\boldsymbol{k})}$

Structure factors ←→electron density

h	k	l	F(hkl)	φ(hkl)
2	0	0	228.0	180.0
1	0	1	10.4	90.0
2	0	1	901.8	270.0
1	1	1	367.0	332.1
1	2	3	149.3	37.8
8	9	1	97.9	255.1
7	7	2	111.5	139.7

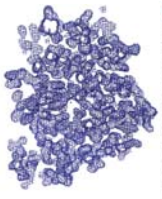


Figure 9-1 Back-transformation of complex structure factors into electron density. The back-transformation of complex structure factors (provided as a list of structure factor amplitudes plus their phases) from the reciprocal into the real space domain by discrete Fourier summation produces the electron density of the scattering molecule. Using the same formalism, any electron density can be transformed into complex structure factors (map inversion). The Fourier transformation from/to the reciprocal space domain (complex structure factors) to/from the real space domain (electron density) is completely reversible without loss of information.

Phases

The crystallographic phase problem

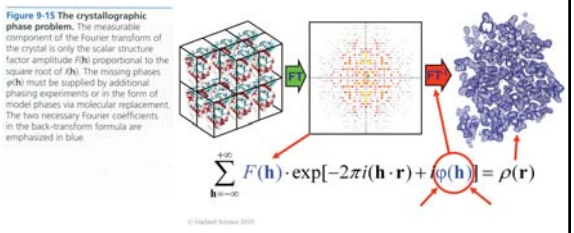
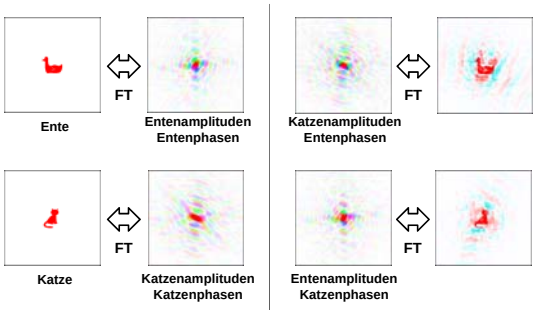


Figure 9-15 The crystallographic phase problem. The measurable component of the Fourier transform of the crystal is only the scalar structure factor amplitude $F(\mathbf{h})$ proportional to the square root of $I(\mathbf{h})$. The missing phases $\varphi(\mathbf{h})$ must be supplied by additional phasing experiments or in the form of model phases via molecular replacement. The two necessary Fourier coefficients in the back-transform formula are emphasized in blue.

Fourier Transformation: Phasen und Amplituden



http://www.ysbl.york.ac.uk/~cowtan/fourier/

Phase bias in electron density maps

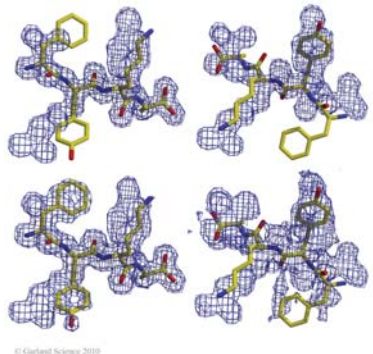


Figure 9-18 Phase bias in electron density maps. The upper panels show a mutant peptide Phe-Tyr-Lys-Ala (left) and the same peptide rotated (leading to reverse chain direction) simply superimposed on the electron density of the original Val-Arg-Tyr-Ala peptide. The lower panels show the electron density reconstructed using the diffraction data from the new models above, but using the old starting phases from the original peptide. The result is quite sobering: the shape of the electron density is still dominated by the starting model, and only weak outlines of the correct molecule density are visible. In the lower left panel, not even the direction of the peptide could be assigned for the reversed peptide with any certainty.

Determination of phases

- **Ab initio phasing (direct methods):** Exploit theoretical phase relationships. Requires high resolution (< 1.4 Å) data.
- **Heavy atom derivatives (multiple isomorphous replacement; MIR):** Crystallize the protein in the presence of several heavy metals without significantly changing the structure of the protein nor the crystal lattice.
- **Anomalous X-ray scattering at multiple wavelengths (multi-wavelength anomalous dispersion; MAD):** Incorporation of Seleno-methionine.
- **Molecular replacement:** Use structure of a similar molecule as the initial model.

Isomorphus difference data

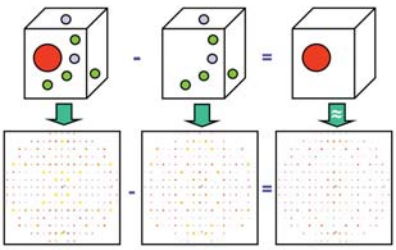
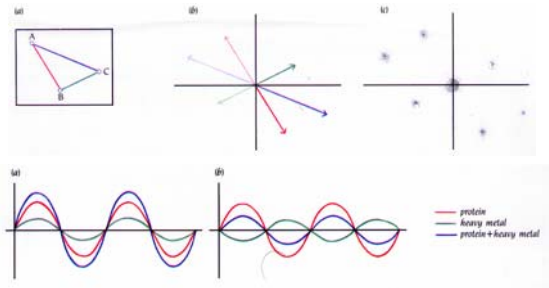
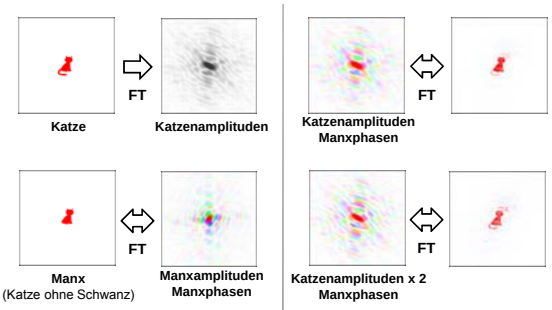


Figure 10-1 The concept of isomorphous difference data. The top line shows the gedankenexperiment in real space of subtracting a native protein crystal from an exactly isomorphous derivative crystal. The light atoms "cancel" out, and only the heavy marker atom remains in the difference crystal. While we cannot produce a real difference crystal, we can very well obtain a "difference diffraction pattern" from the differences between experimental data of the derivative and the native protein. The difference diffraction pattern has the same reciprocal dimensions and thus the same number of reflections, but represents the much simpler scenario of the "difference crystal."

Multiple isomorphous replacement (MIR)



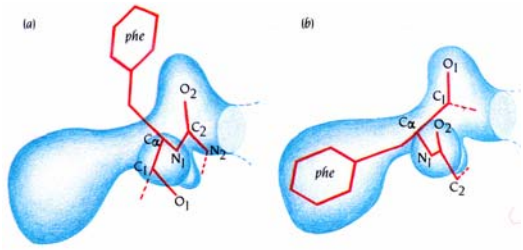
Molecular replacement



http://www.ysbl.york.ac.uk/~cowtan/fourier/

Electron density

Interpretation der Elektronendichte



Manual model building

Figure 12-1 **The Richards box used to build the model of thermolysin in 1972.** Prior to the development of computer graphics, the "Richards Box," also known as "Fred's Folly," was used to build physical models of protein structures assembled from prefabricated parts. The panel above shows on the left side the wire model of the crystal structure of the cro-repressor assembled from "Kendrew parts" at a scale of 2 cm Å^{-1} together with a Watson–Crick DNA model; there were no crystal structures of DNA available before 1979.[7] To the right of the model is a storage crate for the electron density sections plotted on clear plastic sheets, each suspended on a 3 by 3 feet square aluminum frame. A block of 11 map sections pulled out from the storage crate that represent the "active" part of the electron density map can be seen right of the model. A large, semi-transparent mirror is mounted vertically between the model and the electron density map. A viewer standing at the extreme left in front of the model and looking toward the mirror would see the view photographed in the right panel. The electron density sections, visible through the mirror, are superimposed on the image of the model reflected from the face of the mirror. The physical model (out-of focus in the foreground) is assembled from the prefabricated Kendrew metal parts secured together by screw fasteners recognizable in the virtual image together with the electron density. In the original version introduced by Richards[8] the mirror was mounted at an angle of 45° to the map sections.Brian Matthews and Dale Tronrud (University of Oregon) kindly provided the photographs of the box.

Manual model building

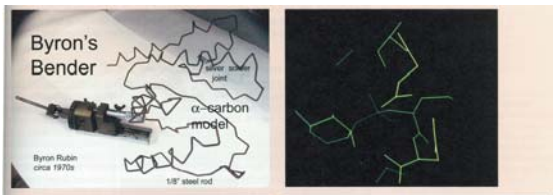


Figure 12-2 **Byron's bender.** The instrument, invented by Byron Rubin, allows dialing-in of Cα-backbone torsion angles and the bending of steel wires accordingly so that they form the Cα-backbone model of the protein structure. The annotated image of an early bender was kindly provided by Leonard Banaszak, University of Minnesota, Twin Cities.

Figure 12-3 **A screen shot of FRODO.** The picture is a screen shot of a MMS-X CRT graphics system showing part of a thermolysin inhibitor displayed by the original 1978 version of FRODO,[9] which later became O,[10] a venerable graphics program by Alwyn Jones, still used in some laboratories today. Dale Tronrud kindly provided the picture taken in Brian Matthews' laboratory.

Electron density at different resolution

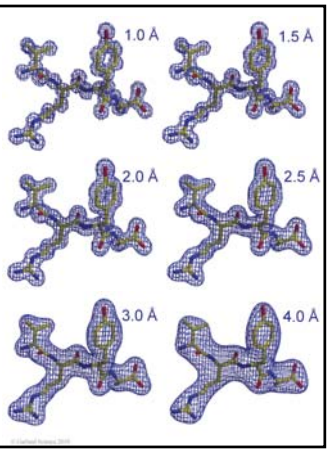


Figure 9-8 **Electron density reconstruction at decreasing resolution.** The electron density shape progressively changes from distinct atomic spheres discernable at 1.0 Å to a sausage- or tube-like electron density without distinguishable side chain definition at 4.0 Å. The electron densities are reconstructed from error-free, *B*-factor attenuated $\mathbf{F}_{calc}$ data and thus represent noise-free, best case scenarios.

Effect of omitted data

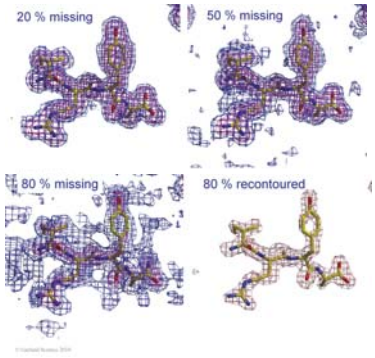


Figure 9-10 Effect of randomly omitted data. The panels show the effect of an increasing amount of randomly missing data. In the top left panel with 20% of the data randomly deleted, there is barely a difference noticeable compared with the maps generated from compete data shown in Figure 9-8. Even when the reconstruction misses 50% and 80% of data, the molecule is still traceable despite the increase in noise. About 800 out of 4000 reflections are all that is left in the reconstruction of the bottom electron density. The density in the bottom right panel is recontoured 80% missing at a higher σ-level, and the molecule is still traceable. Comparing the bottom left and bottom right panels emphasizes the importance of selecting a suitable density level for model density visualization and model building.

Difference maps

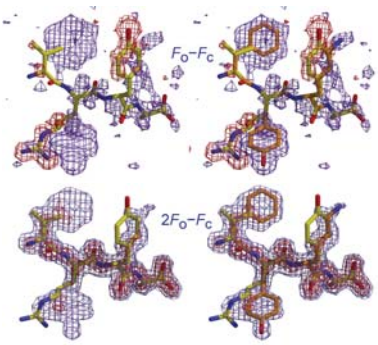


Figure 9-19 $F_o - F_c$ difference maps and $2F_o - F_c$ maps. The $F_o - F_c$ difference maps in the top panel show negative difference density (where there should not be any density) in red and positive difference density (where there should be density) in purple. Both regions correctly reveal the difference between the starting model (yellow sticks) and the correct model (orange sticks, shown in the right panels). The sensitive difference maps are thus particularly valuable for detailed model correction. The $2F_o - F_c$ maps in the bottom panel can be interpreted as a combination of the difference map $(F_o - F_c)\cdot\exp(i\varphi_c)$ and a $(F_c)\cdot\exp(i\varphi_c)$ map. The $2F_o - F_c$ map is contoured to amplify positive density and is well suited to early model building stages. Another common color scheme is red for negative difference density and green for positive.

Poor start phases → poor electron density maps

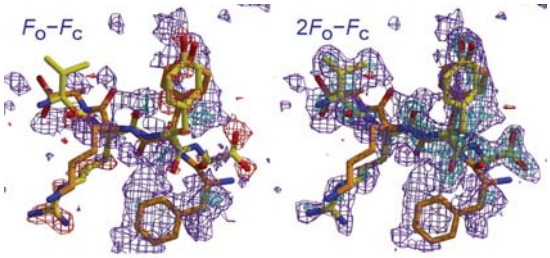


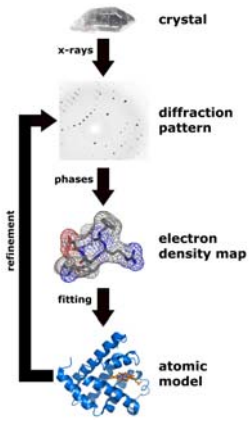
Figure 9-20 Poor starting phases give poor electron density maps. In the case of the reverse traced mutant peptide (orange sticks) as the true structure providing the intensities, no basic map type is able—despite good 1.5 Å data—to give sufficient clues as how to correct the starting model (yellow sticks) that provided the phases. The difference map informs us in some parts about what is wrong, but none of the maps has sufficient reconstructive power to produce an outline of the correct orange molecule.

Key concepts of model building

- The key to successful protein structure modeling is the cycling between local real space model building and model correction and global reciprocal space refinement.
- The molecular model is built in real space into electron density using computer graphics.
- Local geometry errors remaining after real space model building are corrected during restrained reciprocal space refinement by optimizing the fit between observed and calculated structure factor amplitudes.
- Successive rounds of rebuilding, error correction, and refinement are needed to obtain a good final protein model.
- While experimental electron density maps constructed from poor phases will be hard to interpret, an initial experimental map will not be biased toward any structure model.
- In contrast, when molecular replacement models are the sole source of phases, the electron density maps will be severely biased, and the map will reflect the model features.

Refinement

Struktur-ermittlung



Local minima during refinement

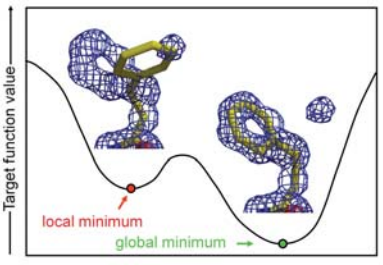


Figure 12-7 Local minima and radius of convergence. The figure visualizes the concept of trapping in a local minimum for a real space scenario. The $C\varepsilon_2$ atom of the misplaced Phe ring is trapped in the electron density of a water molecule, in which it happens incidentally to fit quite well. In such cases, a refinement program may not be able to proceed upwards over the "activation" barrier—or may allow only limited positional parameter shifts—that prevent the large movement of the entire ring out of the partial density until it snaps into the correct electron density. Increased ability to overcome local minima by allowing "upwards movement" during parameter search implies higher radius of convergence and higher probability to approach the global minimum, generally at the cost of more computation and lower accuracy.

X-ray crystallography: *R*-factor

- Measures agreement between measured data (reflections) and 3D structure
- Definition: Relative difference between structure factors, $F(hkl)$, that were observed (F_{obs}) and back-calculated from the 3D structure (F_{calc}):

$$R = \frac{\sum \left| |F_{\text{obs}}| - |F_{\text{calc}}| \right|}{\sum |F_{\text{obs}}|} \quad \text{with} \quad I_{hkl} \propto |F(hkl)|^2$$

I_{hkl} = intensity of reflection (hkl)

- Perfect agreement: $R = 0$
  Good protein X-ray structure: $R < 0.2$
  Random structure: $R \approx 0.6$

Over-fitting

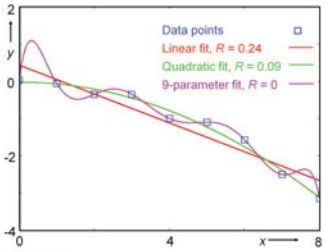


Figure 12-8 Fitting and over-fitting of a function. The data points are measurements of the drop y of a diffractometer pushed over a cliff, taken at constant time intervals x. The linear, 2-parameter model (red line) describes the data poorly, but a quadratic 3-parameter fit (green graph) clearly describes the data very well, as it represents the physically correct model of a parabolic function describing the trajectory of a dropping object. We can further improve the fit (but not the model) by adding more parameters, and a 9-parameter polynomial function (magenta) perfectly fits the data. Despite the perfect fit, the model is definitely nonsense, because the trajectory of the falling object takes upward turns, which is physically impossible. Following Bayesian reasoning the model, despite describing the data well, can be rejected based on a vanishingly small prior probability. In multi-parametric models such as crystal structures, over-fitting is unfortunately much less obvious, and cross-validation is a necessary practice.

X-ray: Free *R*-factor

- Use, say, 90% of the data (reflections) for the structure determination
- Use the remaining 10% to compute the *R* value → "free" *R* value, obtained from independent data
- Detects errors better than conventional *R*-factor
- Each reflection influences whole electron density
- Many reflections → No problem to omit 10% of the reflections from the structure determination

Brünger, A. T. (1992). Free *R* value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* 355, 472-475.

Cross-validation: *R*-free value

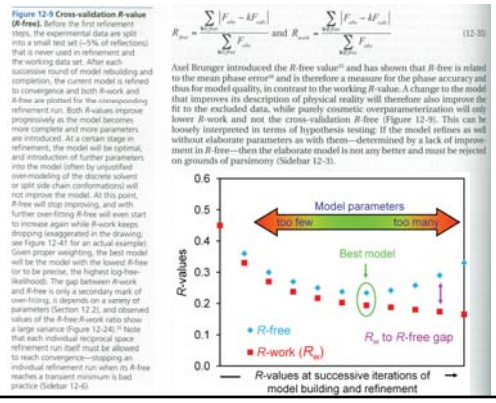


Figure 12-9 Cross-validation *R*-value (*R*-free). Before the first refinement steps, the experimental data are split into a small test set (~5% of reflections) that is never used in refinement and the working data set. After each successive round of model rebuilding and completion, the current model is refined to convergence and both *R*-work and *R*-free are plotted for the corresponding refinement run. Both *R*-values improve progressively as the model becomes more complete and more parameters are introduced. At a certain stage in refinement, the model will be optimal, and introduction of further parameters into the model (often by unjustified over-modeling of the discrete solvent or split side chain conformations) will not improve the model. At this point, *R*-free will stop improving, and with further over-fitting *R*-free will even start to increase again while *R*-work keeps dropping (exaggerated in the drawing; see Figure 12-41 for an actual example). Given proper weighting, the best model will be the model with the lowest *R*-free (or to be precise, the highest log-free-likelihood). The gap between *R*-work and *R*-free is only a secondary mark of over-fitting; it depends on a variety of parameters (Section 12.2), and observed values of the *R*-free:*R*-work ratio show a large variance (Figure 12-24).[10] Note that each individual reciprocal space refinement run itself must be allowed to reach convergence—stopping an individual refinement run when its *R*-free reaches a transient minimum is bad practice (Sidebar 12-6).

$$R_{free} = \frac{\sum_{h\,k\,l \in \text{test}} |F_{obs} - kF_{calc}|}{\sum_{h\,k\,l \in \text{test}} F_{obs}} \text{ and } R_{work} = \frac{\sum_{h\,k\,l \in \text{work}} |F_{obs} - kF_{calc}|}{\sum_{h\,k\,l \in \text{work}} F_{obs}} \quad (12\text{-}3)$$

Axel Brunger introduced the *R*-free value[11] and has shown that *R*-free is related to the mean phase error[12] and is therefore a measure for the phase accuracy and thus for model quality, in contrast to the working *R*-value. A change to the model that improves its description of physical reality will therefore also improve the fit to the excluded data, while purely cosmetic overparameterization will only lower *R*-work and not the cross-validation *R*-free (Figure 12-9). This can be loosely interpreted in terms of hypothesis testing: If the model refines as well without elaborate parameters as with them—determined by a lack of improvement in *R*-free—then the elaborate model is not any better and must be rejected on grounds of parsimony (Sidebar 12-3).

Cross-validation in reciprocal and real space

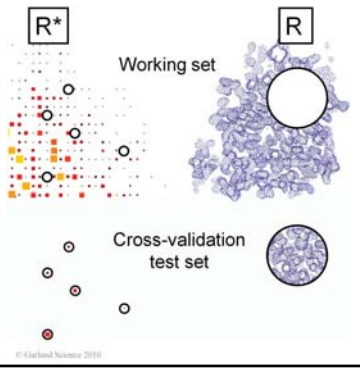


Figure 12-10 Cross-validation in reciprocal space and in real space. A subset of unique reflections is set aside (the test or cross-validation set) before the model is refined and is excluded from any further refinement. The model is then refined against the working data set and the progress of the refinement is tracked against the test data set. In a similar fashion, the electron density or model in a questionable region can be removed, the model is again refined (often combined with a bias removal step), and the omitted region is inspected for new electron density. The figure layout follows an idea by A.T. Brunger.[14]

Data-to-parameter ratio for X-ray protein structure determination

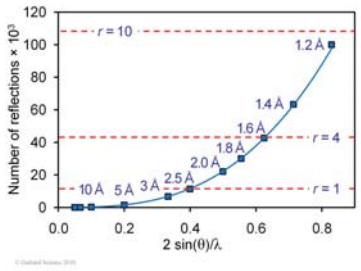


Figure 12-11 Data-to-parameter ratio for protein structures. The graph shows the number of reflections as a function of resolution (in units of $1/d$). The red dashed lines are drawn at numbers of reflections that correspond to a given data-to-parameter ratio r. The number of reflections approaches the number of refined parameters for positional and individual B-factor refinement around 2.5 Å. Below that resolution, unrestrained refinement is underdetermined, and only at atomic resolution does the redundancy of measured data become high enough ($r > 10$) that unrestrained refinement becomes even remotely conceivable. The redundancy levels are generally valid for proteins with a solvent content around 50%. Tighter packing means a smaller unit cell and thus fewer reflections compared with loose packing. For torsion angle only refinement, the n/p ratio is slightly better (Section 12.2); therefore it is often the only available refinement protocol for low resolution (below ~3.5 Å) structures.

Key concepts of refinement I

- During refinement the parameters describing a continuously parameterized model are adjusted so that the fit of discrete experimental observations to their computed values calculated by a target function is optimized.
- Observations can be experimental data specific to the given problem, such as structure factor amplitudes, or general observations that are valid for all models.
- Stereochemical descriptors valid for all models such as bond lengths, bond angles, torsion angles, chirality, and non-bonded interactions are incorporated as restraints to improve the· observation-to-parameter ratio of the refinement.
- The most accurate target functions are maximum likelihood target functions that account for errors and incompleteness in the model.
- Various optimization algorithms can be used to achieve the best fit between parameterized model and all observations, which include measured data and restraints.

Key concepts of refinement II

- The radius of convergence for an optimization algorithm describes its ability to escape local minima and approach the global minimum, generally with increased cost in time and lower accuracy.
- Indiscriminate introduction of an increasing number of parameters into the model can lead to overparameterization, where the refinement residual measured as linear *R*-value still decreases, but the description of reality, i.e., the correct structure, does not improve.
- The evaluation of the residual against a data set excluded from refinement provides the cross-validation *R*-value or *R*-free. If parameters are introduced that do not improve the phase error of the model, *R*-free will not decrease any further or may even increase.
- Refined models carry some memory of omitted parts, which can be removed by slightly perturbing the coordinates and re-refining the model without the questionable part of the model.
- The known geometry target values for bond lengths, bond angles, and torsion angles as well as planarity of certain groups can be regarded as additional observations contributing to a higher data-to-parameter ratio.

Key concepts of refinement III

- In addition, geometry targets constitute prior knowledge that keeps the molecular geometry in check with reality during restrained refinement.
- The geometry targets, chirality values, and non-bonded interactions are implemented as stereochemical restraints and incorporated into the target function generally in the form of squared sum of residuals in addition to the structure factor amplitude residual.
- The structure factor amplitude residual is commonly called the X-ray term (or X-ray energy) and the restraint residuals the chemical (energy) term.
- In terms of maximum posterior estimation, geometry target values and their variance define the prior probability of our model without consideration or knowledge of the experimental (diffraction) data.
- Geometric relations and redundancies between identical molecules in the asymmetric unit can be exploited through NCS restraints.
- Particularly at low resolution, strong NCS restraints are an effective means of stabilizing and improving the refinement.

Key concepts of refinement IV

- In the early stages of model building, experimental phase restraints are also an effective means to stabilize and improve the refinement.
- The data-to-parameter ratio in protein structures is greatly increased through the introduction of stereochemical restraints.
- A protein of 2000 non-hydrogen atoms has about 8000 adjustable parameters and about the same number of restraints.
- At 2 Å about 15 000 to 25 000 unique reflections are observed for a 2000 nonhydrogen atom protein, which yields a total data to parameter ratio of about 2-3 at 2 Å.
- Anisotropic *B*-factor refinement consumes 5 additional parameters per atom, and is generally not advisable at resolutions <1.4 Å.
- The most difficult point in the parameterization of macromolecular structure models is accounting for correlated dynamic or static displacement.
- Isotropic *B*-factors are inadequate to describe any correlated dynamic molecular movement, and anisotropic *B*-factors, except at very high resolution, lead to overparameterization of the model.

Key concepts of refinement V

- Molecular and lattice packing anisotropy can also affect diffraction, and adequate correction by anisotropic scaling, or in severe cases additional anisotropic resolution truncation, is necessary.
- Maximum likelihood target functions that account for incompleteness and errors in the model are superior to basic least squares target functions, particularly in the early, error-prone stages of refinement.
- Maximum likelihood target functions are implemented in REFMAC, Buster/ TNT, and CNS as well as the PHENIX/ cctbx programs, together with all commonly used restraint functions including phase restraints, which is of advantage at low resolution or in the early stages of refinement.
- Optimization algorithms are procedures that search for an optimum of a nonlinear, multi-parametric function.
- Optimization algorithms can be roughly divided into analytic or deterministic procedures and stochastic procedures.
- Deterministic optimizations such as gradient-based maximum likelihood methods are fast and work well when reasonably close to a correct model, at the price of becoming trapped in local minima.

Key concepts of refinement VI

- Stochastic procedures employ a random search that also allows movements away from local minima. They are slow but compensate for it with a large radius of convergence.
- Evolutionary programming as used in molecular replacement or simulated annealing in refinement is a stochastic optimization procedure. This is generally of advantage if we do not know (MR) or are far from (initial model refinement) the correct solution.
- Deterministic optimizations can be classified depending on how they evaluate the second derivative matrix. They generally descend in several steps or cycles from a starting parameter set (model) downhill toward a hopefully but not necessarily global minimum.
- Energy refinement of a molecular dynamics force field and torsion angle refinement are two parameterizations that are used together with the stochastic optimization method of simulated annealing.
- In molecular dynamics the target function is parameterized in the form of potential energy terms and the development of the system is described by equations of motion. In torsion angle parameterization, the structure model is described by its torsion angles, which requires fewer parameters than coordinate parameterization.

Key concepts of refinement VII

- Both molecular dynamics and torsion angle parameterization are often combined with simulated annealing optimization, where the molecular system is perturbed and returns to equilibrium according to an optimized slow cooling protocol.
- Dummy atom placement and refinement is used for discrete solvent building, model completion, and phase improvement in general.
- Dummy atoms are placed in real space in difference electron density peaks, the new model is refined unrestrained in reciprocal space, and in the new map poorly positioned atoms are removed and new ones placed again.
- Dummy atom refinement can be combined with multi-model map averaging where it forms the basis of bias minimization protocols and the automated model building program ARP/wARP.

Model building and refinement practice I

- Building of a model into an empty map begins with the tracing of the backbone.
- Tracing is aided by density skeletonization, followed by placement of C^{α} atoms into positions where side chains extend from the backbone.
- The sequence is docked from known atom positions from the heavy atom substructure or sequences of residues of characteristic shapes.
- The initial model is refined in reciprocal space with geometric restraints and phase restraints, and the next map is constructed from maximum likelihood coefficients.
- The model is then further completed and refined in subsequent rounds with increasing X-ray weights while tracking *R*-free and stereochemistry. Nuisance errors are removed after analysis in a polishing step.
- Automated model building programs greatly simplify model building, and auto-built models often only need to be completed and polished. Autobuilding programs follow similar steps as manual model building and employ pattern recognition algorithms to identify residues.

Model building and refinement practice II

- Rebuilding poor initial molecular replacement models can be aided by a first step of torsion angle-simulated annealing (TA-SA) refinement.
- The large radius of convergence of TA-SA facilitates the necessary large corrections and escape from local minima. Also, before automated model rebuilding and correction, TA-SA can improve the amount and quality of the model that is automatically rebuilt.
- In low resolution structures the backbone can be traced correctly, but the sequence may be shifted. Such register errors can be hard to detect from electron density shape alone and are usually detected by poor side chain interactions or unusual environment.
- A common mistake leading to overparameterization of the model is overbuilding of the solvent. Discrete water molecules should have hydrogen bonded contact(s) to other solvent molecules or to protein.
- Poorly placed waters tend to drift away during refinement because of lack of density and restraints and often end up far away from other molecules and with high *B*-factors.

Model building and refinement practice III

- Binding sites have a tendency to attract various detritus from the crystallization cocktail, and will therefore often contain some weak, unidentifiable density that can be (wishfully) mistaken for desired ligand density.
- Plausible binding chemistry, ligand conformation, and independent evidence are necessary to avoid misinterpretation.
- The three major criteria for abandoning refinement and rebuilding are:
  (i) No more significant and interpretable difference density in $mF_{\mathrm{obs}} - DF_{\mathrm{calc}}$ maps remains.
  (ii) No more unexplained significant deviations from stereochemical target values and from plausible stereochemistry remain.
  (iii) The model makes chemical and biological sense.
- Global measures such as absolute values of *R* and *R*-free (or the level of boredom) do not determine when refinement is finished.

Literatur über Kristallstrukturbestimmung

- **B. Rupp, *Biomolecular Crystallography*, Garland, 2010.**
- W. Massa, *Kristallstrukturbestimmung*, Teubner, ^{5}2007.
- C. Branden & J. Tooze, *Introduction to Protein Structure*, Garland, ^{2}1999.

Skript

www.bpc.uni-frankfurt.de/guentert/wiki/index.php/Teaching