Strukturelle Modellierung
(Masterstudiengang Bioinformatik)

Strukturbestimmung mit Röntgenkristallographie und NMR Spektroskopie

Sommersemester 2012

Peter Güntert

X-ray crystallography

Strukturelle Modellierung
(Masterstudiengang Bioinformatik)

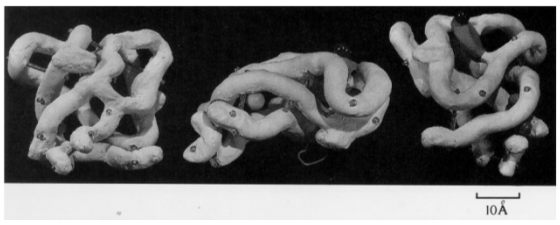
Strukturbestimmung mit Röntgenkristallographie

Sommersemester 2012

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
2012: > 72'000 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

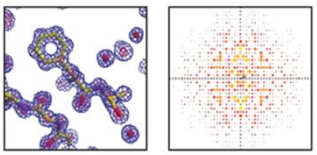
Protein crystals are formed by a loose periodic network of weak, non-covalent interactions and contain large solvent channels. The solvent channels allow relatively free diffusion of small molecules through the crystal and also provide conformational freedom for surface-exposed side chains or loops. 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 or by analysis of correlated displacement by TLS (Translation-Libration-Screw) analysis. 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

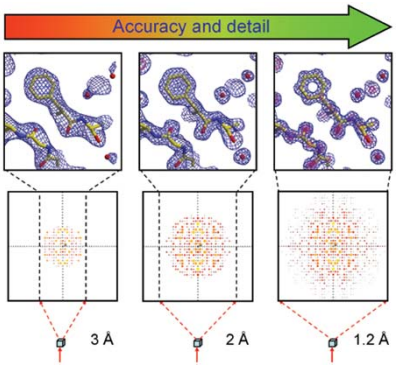
$$\rho(x,y,z) = \frac{1}{V}\sum_{-h}^{h}\sum_{-k}^{k}\sum_{-l}^{l} F_{hkl} \cdot \exp[-2\pi i(hx+ky+lz-\alpha_{hkl})]$$



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



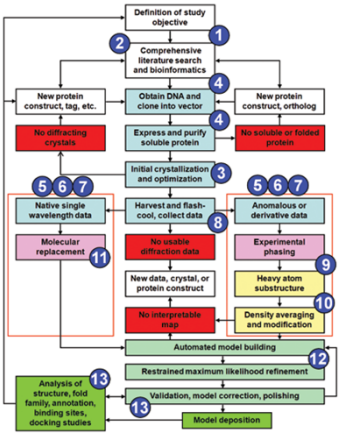
The qualitative relation between the extent of X-ray diffraction, the resulting amount of available diffraction data, and the quality and detail of the electron density reconstruction and protein structure model are evident from this figure: The crystals are labeled with the nominal resolution d_{min} 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. Kristallisation
2. Messung der Beugungsmuster
3. 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 82 000 entries as of June 2012.
- 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

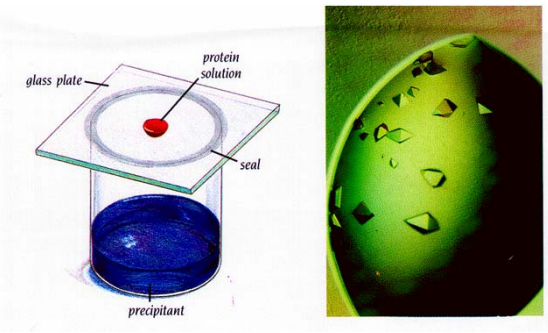
- The X-ray diffraction patterns are not a direct image of the molecular structure.
- The electron density of the scattering molecular structure must be reconstructed by Fourier transform techniques.
- Both structure factor amplitude and relative phase angle of reach reflection are required for the Fourier 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.

Key concepts of protein cystallography III

- Protein crystals are formed by a loose periodic network of weak, non-covalent interactions and 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 regions 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 analyzing 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.

Crystallization

Proteinkristallisation



Protein crystallization basics

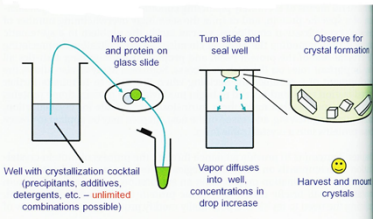
- Protein crystals are periodic self-assemblies of large and often flexible macromolecules, held together by weak intermolecular interactions. Protein crystals are generally fragile and sensitive to environmental changes.
- In order to form crystals, the protein solution must become supersaturated. In 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

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

Figure 3-1 Basic hanging-drop vapor diffusion. Hanging-drop vapor diffusion has been in use for over 30 years for the manual setup of protein crystallization. The reservoir (generally one well of a multi-well assay plate) is partially filled with several hundred µl of crystallization cocktail. A small drop (a few µl or less) of this cocktail is set in the center of a siliconized cover slide, and mixed there with an equal volume of protein stock solution (green). The cover slide is then turned over and placed on the greased rim of the reservoir well. The mixing with protein has reduced the precipitant cocktail concentration to half of the original value, and the sealed system thus equilibrates by water vapor diffusion from the drop into the reservoir solution, thus effectively increasing the concentration of all constituents (protein and precipitation cocktail reagents) in the crystallization drop. During this process the drop becomes supersaturated, nucleation can occur, and protein crystals may grow from the supersaturated solution.



Solubility phase diagram

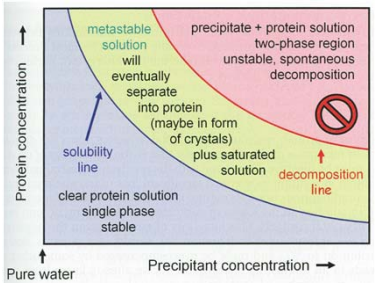


Figure 3-7 A basic solubility phase diagram for a given temperature. The diagram visualizes the general observation that the higher the precipitant concentration in the solution, the lower the maximal achievable protein concentration in the solution and *vice versa*. Between the solubility line and the 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

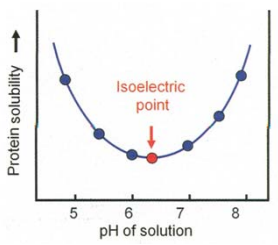


Figure 3-8 Protein solubility versus pH of protein solution. The protein shown in this example has its solubility minimum at its isoelectric point of ~6.3, where the sum of positive and negative charges (the net charge of the protein) is zero. Even at the isoelectric point, there are still numerous (but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

Crystal growth

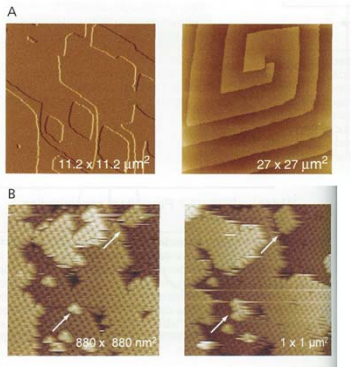


Figure 3-11 Atomic force microscope images of crystal growth. (Panel A) The atomic force microscope images of the 001 surface of glucose isomerase show the two most common growth patterns observed in crystal growth: step growth starting from 2-dimensional nucleation islands (A, left image) and a double-spiral 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

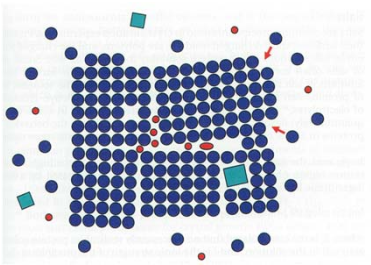


Figure 3-12 Growth of a real mosaic crystal. The schematic drawing shows a crystal growing in a solution of protein molecules (blue spheres). Small impurities (red) and some larger detritus (green squares) are also present in the solution. New molecules attach preferentially to steps and edges (red arrows) 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;[83] or a dual-station layout using separate cocktail preparation with a generic liquid-handling system followed by a dedicated plate-setup robot.[84] 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.

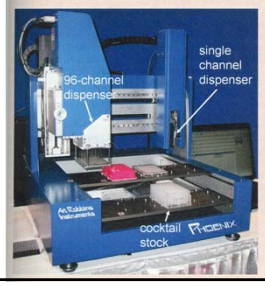
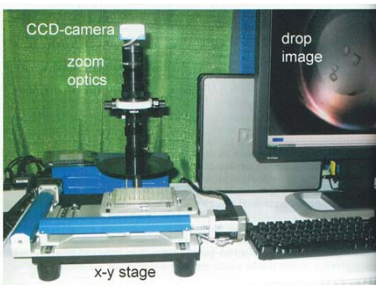


Figure 3-33 A robot for automated crystallization plate setup. The Phoenix robot (Art Robbins Instruments) can set up 96 crystallization trials in about one minute. On the left side, a 96-channel syringe dispenser re-arrays (10 µl each) 96 prefabricated or purchased crystallization cocktails simultaneously 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.[84]

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

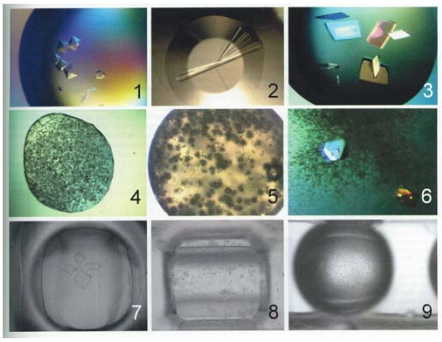


Figure 3-37 Images of crystallization drops with different experimental outcomes. (1) Perfectly formed octahedral single crystals in hanging drop; (2) a cluster of large needles in microbatch reservoir; (3) thin plates in a hanging drop; (4) microcrystal shower in hanging drop; (5) small crystals together 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

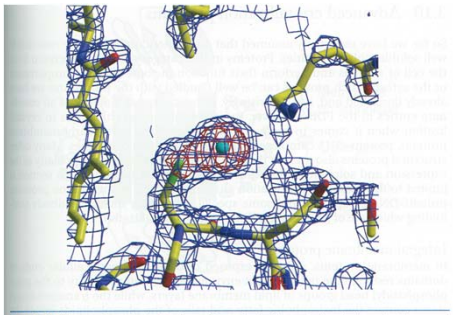


Figure 3-42 **Heavy atom derivatization of a protein.** Shown is the electron density around a gold atom covalently linked to a cysteine residue in the *Clostridium tetani* neurotoxin.[123] A combination of anomalous and isomorphous signals from gold atoms were used to solve the structure of the ganglioside 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	$K_3UO_2F_5$
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

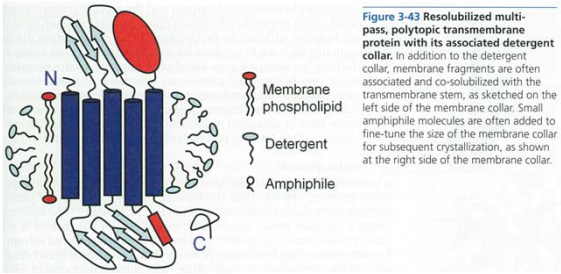
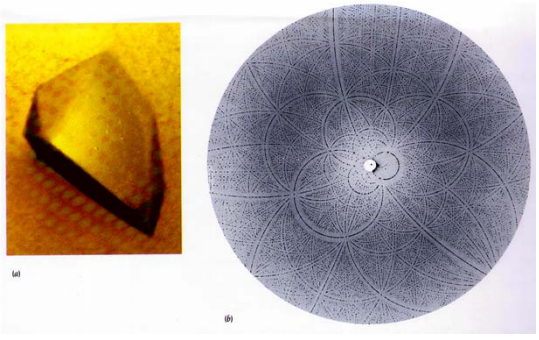


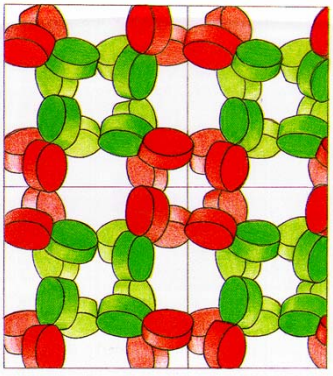
Figure 3-43 **Resolubilized multi-pass, polytopic transmembrane protein with its associated detergent collar.** In addition to the detergent collar, membrane fragments are often associated and co-solubilized with the transmembrane stem, as sketched on the left side of the membrane collar. Small amphiphile 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

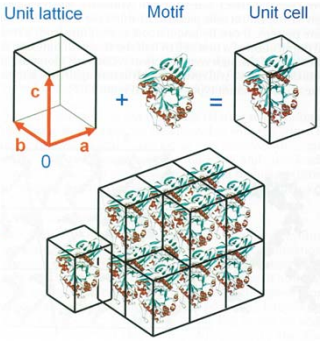


Figure 5-24 Assembly of a primitive triclinic 3-dimensional crystal from unit cells. In analogy to the 2-dimensional case, the unit lattice is filled with a motif, and the crystal is built from translationally stacked unit cells. The basis vectors form a right-handed system [0, **a**, **b**, **c**].

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

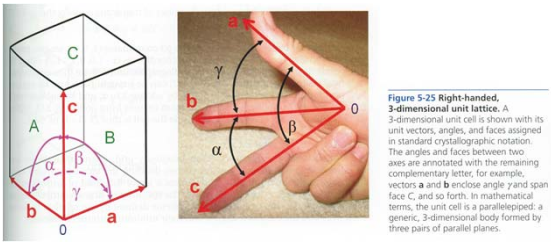


Figure 5-25 Right-handed, 3-dimensional unit lattice. A 3-dimensional unit cell is shown with its unit vectors, angles, and faces assigned in standard crystallographic notation. The angles and faces between two axes are annotated with the remaining complementary letter, for example, vectors **a** and **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

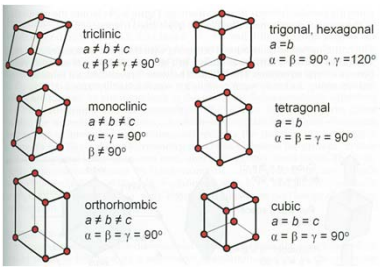


Figure 5-27 The six primitive 3-dimensional lattices. The lattices are derived from a general oblique lattice (in which all six cell parameters are different) and are compatible with increasing internal symmetry (Table 5-2). The trigonal/hexagonal lattice splits into two different crystal systems depending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

Centered 3D Bravais lattices

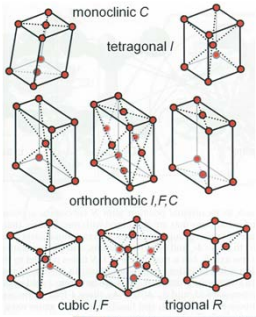
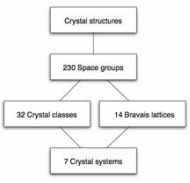


Figure 5-30 Centered 3-dimensional Bravais lattices. In addition to the six primitive 3-dimensional lattices, centered Bravais lattices are derived from the primitive lattices by translational centering. The necessity for additional internal translational symmetry within the unit cells limits the number 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.

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

Kristallsysteme

Bravaisgitter



Bravais lattice	Parameters	Simple (P)	Volume centered (I)	Base centered (C)	Face centered (F)
Triclinic	$a_1 \neq a_2 \neq a_3$ $\alpha_{12} \neq \alpha_{23} \neq \alpha_{31}$				
Monoclinic	$a_1 \neq a_2 \neq a_3$ $\alpha_{23} = \alpha_{31} = 90°$ $\alpha_{12} \neq 90°$				
Orthorhombic	$a_1 \neq a_2 \neq a_3$ $\alpha_{12} = \alpha_{23} = \alpha_{31} = 90°$				
Tetragonal	$a_1 = a_2 \neq a_3$ $\alpha_{12} = \alpha_{23} = \alpha_{31} = 90°$				
Trigonal	$a_1 = a_2 = a_3$ $\alpha_{12} = \alpha_{23} = \alpha_{31} < 120°$				
Cubic	$a_1 = a_2 = a_3$ $\alpha_{12} = \alpha_{23} = \alpha_{31} = 90°$				
Hexagonal	$a_1 = a_2 \neq a_3$ $\alpha_{12} = 120°$ $\alpha_{23} = \alpha_{31} = 90°$				

Protein crystals belong to one of 65 space groups

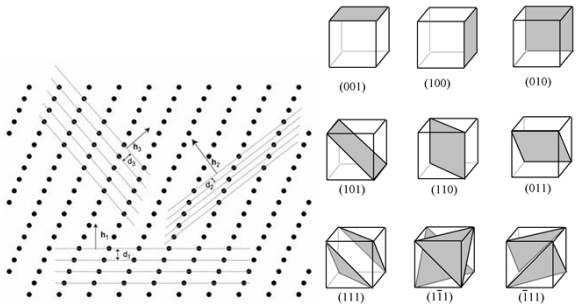
Only 65 discrete and distinct ways exist to assemble 3-dimensional periodiccrystals from asymmetric chiral molecules, through combinations of translational and rotational symmetry. These 65 types of arrangements form 65 chiral space groups, and their symmetry properties and the rules for constructing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

The 65 chiral space groups

Lattice properties	Minimum internal symmetry	Crystal system	Point group	m	Bravais type	β	Lattice type	Chiral space groups	z, M
$a \neq b \neq c$ $\alpha \neq \beta \neq \gamma \neq 90°$	None	Triclinic	1	1	P	1	aP	$P1$	1
$a \neq b \neq c$ $\alpha = \gamma = 90°$ $\beta \neq 90°$	2-fold rotation axis parallel to unique axis **b**	Monoclinic	2	2	P	1	mP	$P2$, $P2_1$	2
					C	2	mC	$C2$	4
$a \neq b \neq c$ $\alpha = \beta = \gamma = 90°$	3 perpendicular, non-intersecting 2-fold axes	Orthorhombic	222	4	P	1	oP	$P222$, $P222_1$, $P2_12_12$, $P2_12_12_1$,	4
					I	2	oI	$I222$, $I2_12_12_1$	8
					C	2	oC	$C222_1$, $C222$	8
					F	4	oF	$F222$	16
$a = b \neq c$ $\alpha = \beta = \gamma = 90°$	4-fold rotation axis parallel to **c**	Tetragonal	4	4	P	1	tP	$P4$, $P4_1$, $P4_2$, $P4_3$	4
					I	2	tI	$I4$, $I4_1$	8
			422	8	P	1	tP	$P422$, $P42_12$, $P4_122$, $P4_12_12$, $P4_222$, $P4_22_12$, $P4_12_2$	8
					I	2	tI	$I422$, $I4_122$	16
$a = b \neq c$ $\alpha = \beta = 90°$ $\gamma = 120°$	3-fold rotation axis parallel to **c**	Trigonal	3	3	P	1	hP	$P3$, $P3_1$, $P3_2$	3
					R	3	hR	$R3$	6
			32	6	P	1	hP	$P312$, $P321$, $P3_112$, $P3_121$, $P3_212$, $P3_221$	9
					R	3	hR	$R32$	18
	6-fold rotation axis parallel to **c**	Hexagonal	6	6	P	1	hP	$P6$, $P6_1$, $P6_5$, $P6_2$, $P6_4$, $P6_3$	6
			622	12	P	1	hP	$P622$, $P6_122$, $P6_522$, $P6_222$, $P6_422$, $P6_322$	12
$a = b = c$ $\alpha = \beta = \gamma = 90°$	Four 3-fold axes along space diagonals	Cubic	23	12	P	1	cP	$P23$, $P2_13$	12
					I	2	cI	$I23$, $I2_13$	24
					F	4	cF	$F23$	48
			432	24	P	1	cP	$P432$, $P4_232$, $P4_332$, $P4_132$	24
					I	2	cI	$I432$, $I4_132$	48
					F	4	cF	$F432$, $F4_132$	96

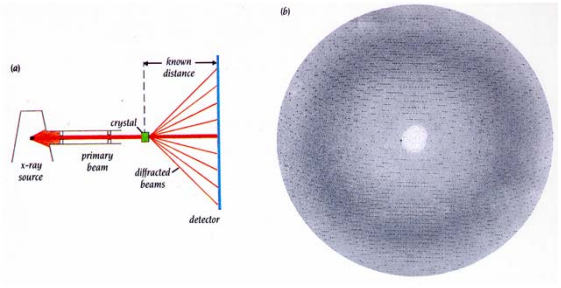
Table 5-2 The 65 chiral space groups. Lattice properties, lattice symmetry, the resulting crystal systems, the 11 enantiomorphic point groups and their multiplicity (m), the Bravais lattice translations and their multiplicity (β), lattice type, and the 65 chiral space groups are listed together with the general position multiplicity (M or z; $M = m \times \beta$). For each enantiomorphic space group. The general position multiplicity M is equivalent to the number of asymmetric units that make up the entire unit cell. In the triclinic lattice symbol aP, a stands for anorthic. The symbols follow the Hermann–Mauguin notation. Augmented Table 6-6 includes additional information relevant for data collection.

Miller Indizes



X-ray diffraction

Röntgenkristallographie Messung



Laboratory X-ray diffractometer

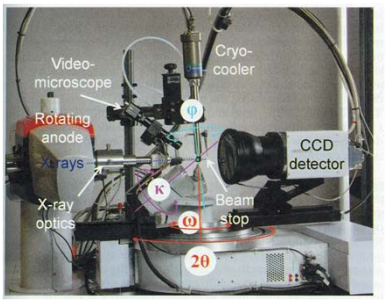
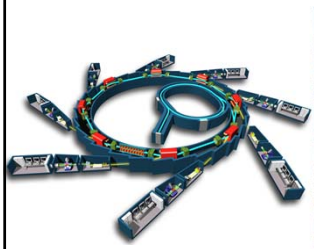
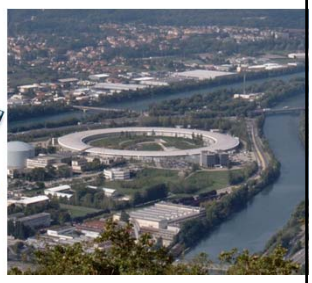


Figure 8-1 A contemporary laboratory X-ray diffractometer for macromolecular crystallography. A rotating anode X-ray source is closely coupled with integrated focusing optics delivering high photon flux at low operating power. In the center of the diffractometer is a full 4-circle κ-goniostat for orienting 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.

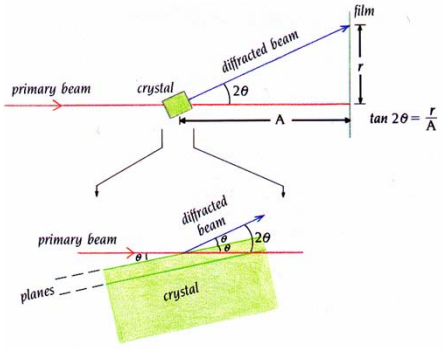
Synchrotron



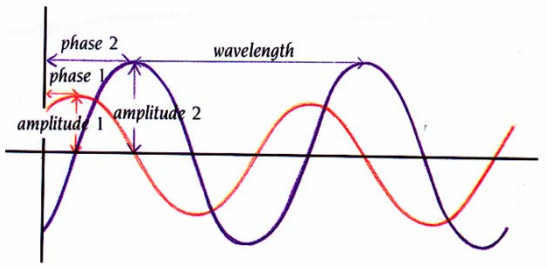


ESRF Grenoble (France)

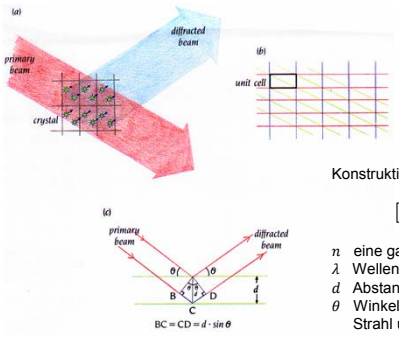
Röntgenstreuung



Superposition of two waves



Röntgenstreuung: Bragg-Bedingung



Konstruktive Interferenz, falls

$$n\lambda = 2d\sin\theta$$

- 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

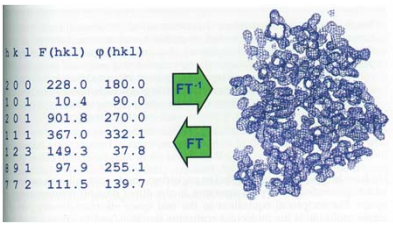
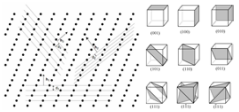


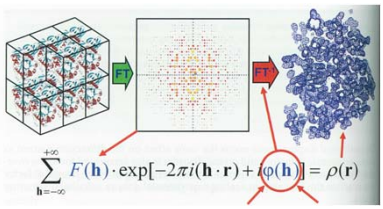
Figure 9-1 Back-transformation of complex structure factors into electron density. The back-transformation of complex structure factors (provided as a list of structure factor amplitudes plus their phases) from the reciprocal into the real space domain by discrete Fourier summation produces the electron 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.



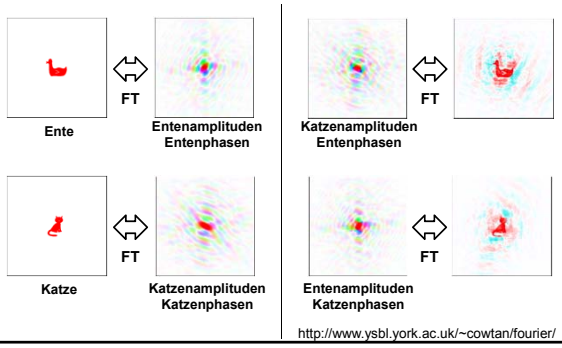
Phases

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 $|\mathbf{F}(\mathbf{h})|$ 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

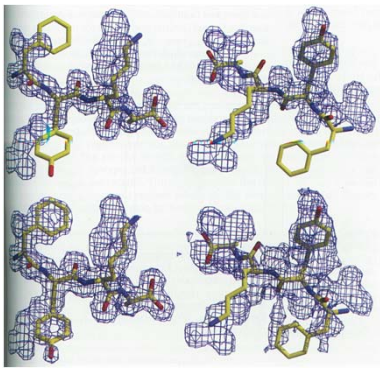


Figure 9-18 Phase bias in electron density maps. The upper panels show a mutant peptide Phe-Tyr-Lys-Ala (left) and the same peptide rotated (leading to reverse chain direction) simply superimposed on the electron density of the original Val-Arg-Tyr-Ala peptide. The lower panels show the electron density reconstructed using the diffraction data from the new models above, but using the old starting phases from the original peptide. The resulting guide-scattering: 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

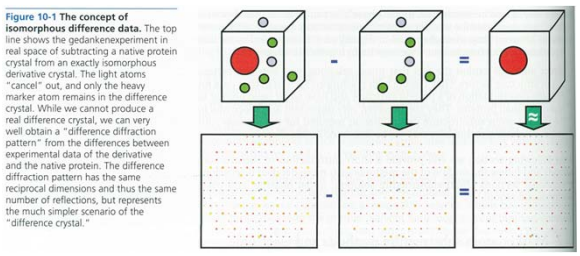
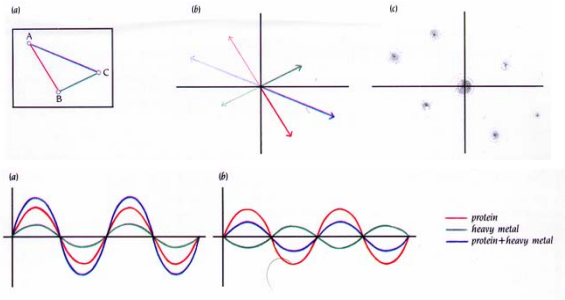
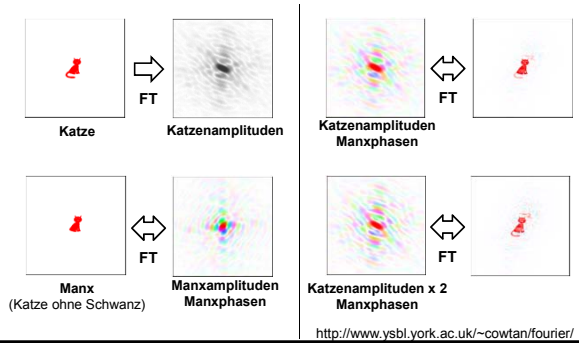


Figure 10-1 The concept of isomorphous difference data. The top line shows the gedankenexperiment in real space of subtracting a native protein crystal from an exactly isomorphous derivative crystal. The light atoms "cancel" out, and only the heavy marker atom remains in the difference crystal. While 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)



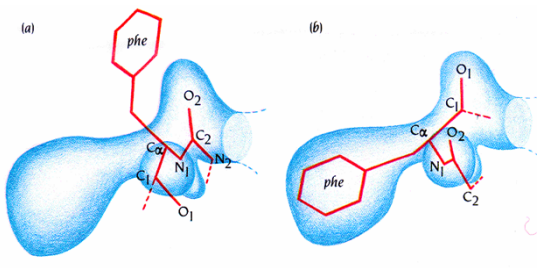
Molecular replacement



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.[3] 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[5] 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

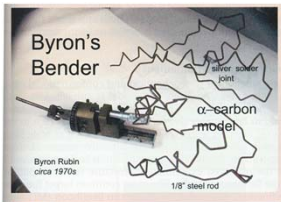


Figure 12-2 Byron's bender. The instrument, invented by Byron Rubin, allows dialing-in of Cα-backbone torsion angles and the bending of steel wires accordingly so that they form the Cα-backbone model of the protein structure. The annotated image of an early bender was kindly provided by Leonard Banaszak,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

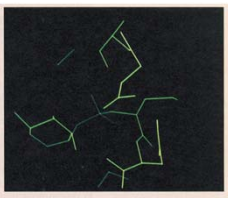


Figure 12-3 A screen shot of FRODO. The picture is a screen shot of a MMS-X CRT graphics system showing part of a thermolysin inhibitor displayed by the original 1978 version of FRODO,[9] which later became O,[12] a venerable graphics program by Alwyn Jones, still used in some laboratories today. Dale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

Electron density at different resolution

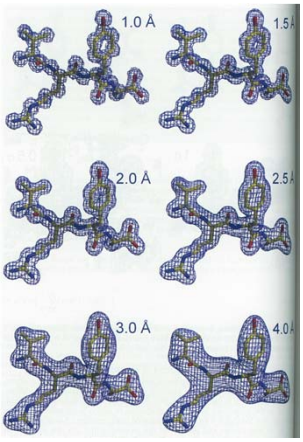


Figure 9-8 Electron density reconstruction at decreasing resolution. The electron density shape progressively changes from distinct atomic spheres discernable at 1.0 Å to a sausage- or tube-like electron density without distinguishable side chain definition at 4.0 Å. The electron densities are reconstructed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

Effect of omitted data

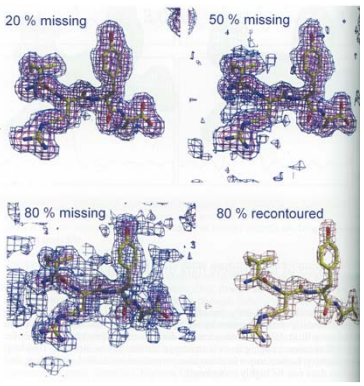


Figure 9-10 Effect of randomly omitted data. The panels show the effect of an increasing amount of randomly missing data. In the top left panel with 20% of the data randomly deleted, there is barely a difference noticeable compared with the maps generated from compete data shown in Figure 9-8. Even when 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

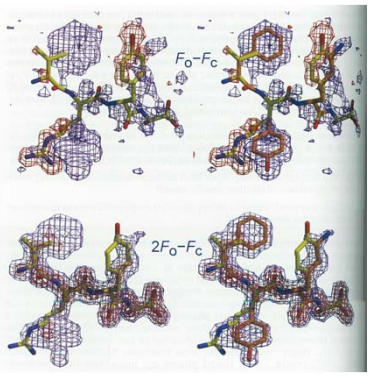


Figure 9-19 $F_o - F_c$ difference maps and $2F_o - F_c$ maps. The $F_o - F_c$ difference maps in the top panel show negative difference density (where there should not be any density) in red and positive difference density (where there should be density) in purple. Both regions correctly reveal the 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 $(F_o)\cdot\exp(i\varphi_c)$ 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

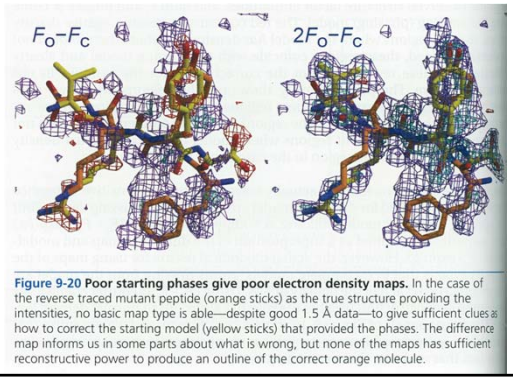


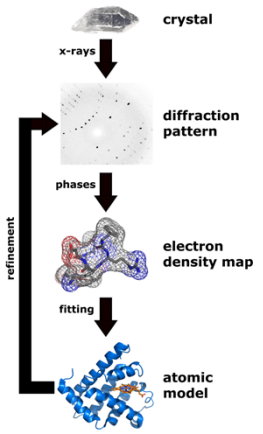
Figure 9-20 Poor starting phases give poor electron density maps. In the case of the reverse traced mutant peptide (orange sticks) as the true structure providing the intensities, no basic map type is able—despite good 1.5 Å data—to give sufficient clues as how to correct the starting model (yellow sticks) 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.

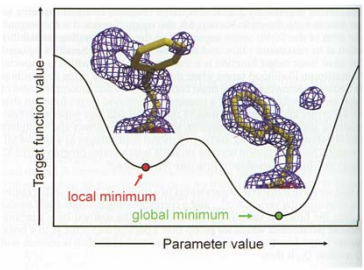
Refinement

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_ε 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}):

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

  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

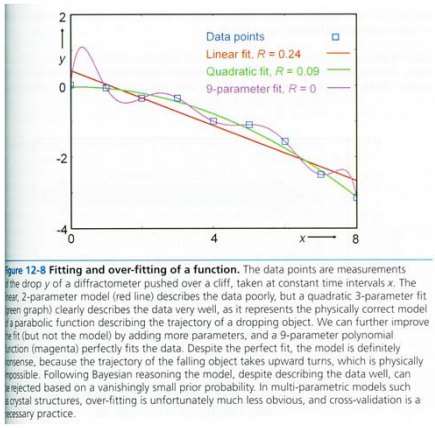


Figure 12-8 Fitting and over-fitting of a function. The data points are measurements of the drop y of a diffractometer pushed over a cliff, taken at constant time intervals x. The linear, 2-parameter model (red line) describes the data poorly, but a quadratic 3-parameter fit (green graph) clearly describes 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

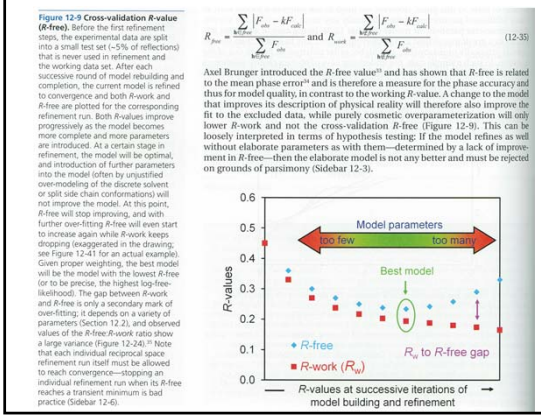


Figure 12-9 Cross-validation *R*-value (*R*-free). Before the first refinement steps, the experimental data are split into a small test set (~5% of reflections) that is never used in refinement and the working data set. After each successive round of model rebuilding and completion, the current model 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).[35] 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).

$$R_{free} = \frac{\sum_{h \in T_{free}} |F_{obs} - kF_{calc}|}{\sum_{h \in T_{free}} F_{obs}} \quad \text{and} \quad R_{work} = \frac{\sum_{h \in T_{free}} |F_{obs} - kF_{calc}|}{\sum_{h \in T_{free}} F_{obs}} \qquad (12\text{-}35)$$

Axel Brunger introduced the *R*-free value[35] and has shown that *R*-free is related to the mean phase error[34] 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

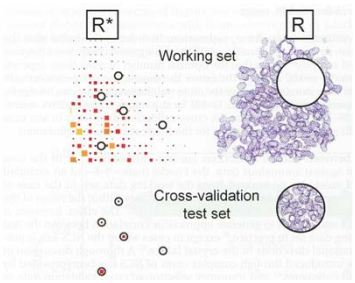


Figure 12-10 Cross-validation in reciprocal space and in real space. A subset of unique reflections is set aside (the test or cross-validation set) before the model is refined and is excluded from any further refinement. The model is then refined against the working data set and the progress of the refinement 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.[36]

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

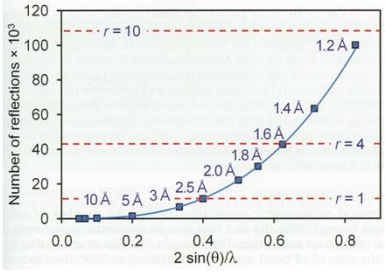


Figure 12-11 Data-to-parameter ratio for protein structures. The graph shows the number of reflections as a function of resolution (in units of 1/*d*). The red dashed lines are drawn at numbers of reflections that correspond to a given data-to-parameter ratio *r*. The number of reflections approaches 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.

NMR

Strukturelle Modellierung
(Masterstudiengang Bioinformatik)

Strukturbestimmung mit NMR Spektroskopie

Sommersemester 2012

Peter Güntert

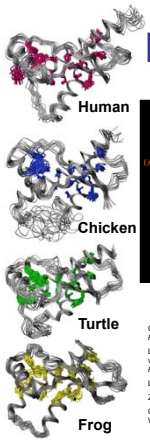
NMR Spektroskopie: Geschichte

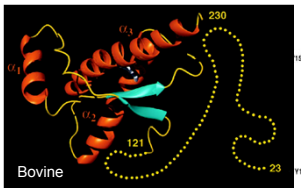
1924, Wolfgang Pauli: Vorhersage des Kernspins
1933, Isidor Rabi: Molekularstrahlmagnetresonanzdetektion
1945: Edward Purcell, Felix Bloch: Kernspinresonanz (NMR)
1953: A. Overhauser, I. Solomon: Nuclear Overhauser Effekt
1966, Richard Ernst: Fouriertransformations-NMR
1971, Jean Jeener: 2D NMR Spektren
1981, Kurt Wüthrich et al.: Resonanzzuordnung in Proteinen
1984, Kurt Wüthrich et al.: 3D Proteinstruktur in Lösung
1991, Ad Bax et al.: Tripelresonanzspektren (^{13}C, ^{15}N, ^{3}H)
1997: TROSY, NMR Spektroskopie von großen Proteinen
2012: ~9400 NMR Strukturen in der Protein Data Bank

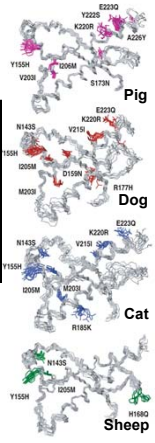
Literatur über NMR Proteinstrukturbestimmung

- K. Wüthrich, *NMR of Proteins and Nucleic Acids*, Wiley, 1986.
- J. Cavanagh, W. J. Fairbrother, A. G. Palmer III, N. J. Skelton & M. Rance, M. *Protein NMR Spectroscopy. Principles and Practice*, Academic Press, ^{2}2006.
- M. Williamson, *How Proteins Work*, Garland, 2012.

Prion proteins







Calzolai, L., Lysek, D. A., Pérez, D. R., Güntert, P., Wüthrich, K. *PNAS* 102, 651-655 (2005).

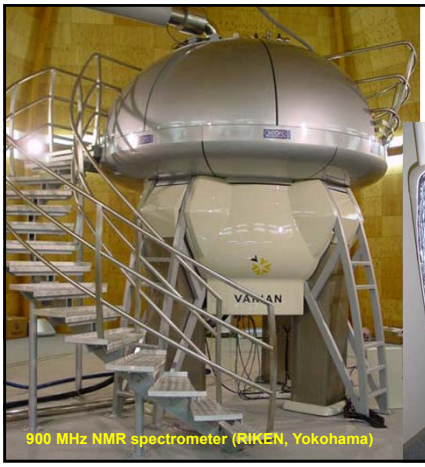
Lysek, D. A., Schorn, C., Nivon, L. G., Esteve-Moya, V., Christen, B., Calzolai, L., von Schroetter, C., Fiorito, F., Herrmann, T., Güntert, P., Wüthrich, K. *PNAS* 102, 640-645 (2005).

Lührs, T., Riek, R., Güntert, P., Wüthrich, K. *JMB* 326, 1549-1557 (2003).

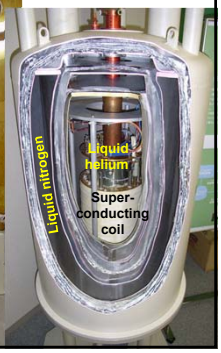
Zahn, R., Güntert, P., von Schroetter, C., Wüthrich, K. *JMB* 326, 225-234 (2003).

Calzolai, L., Lysek, D. A., Güntert, P., von Schroetter, C., Riek, R., Zahn, R., Wüthrich, K. *PNAS* 97, 8340-8345 (2000).

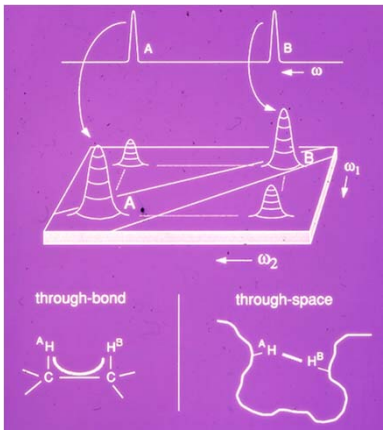
NMR Spectrometer



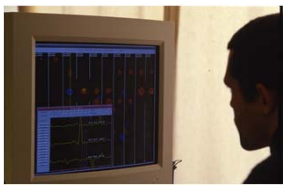
900 MHz NMR spectrometer (RIKEN, Yokohama)



2D NMR Spektren



NMR Spektrenauswertung



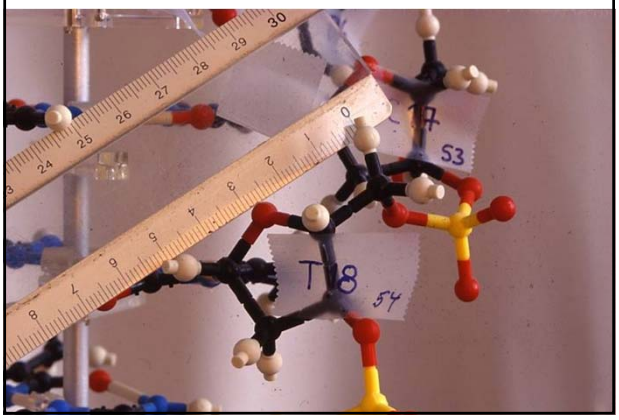
Manuell

Interaktiv

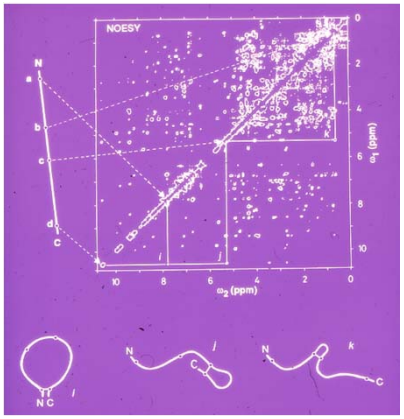
Automatisch

Conformational restraints

NMR measures distances between atoms



NOESY Spektrum



Konformationsdaten aus NMR Messungen

1. Nuclear Overhauser Effects (NOEs)
2. ^3J skalare Kopplungen
3. H-Brücken
4. Chemische Verschiebungen
5. Residuelle dipolare Kopplungen (RDC)

...

Experimental data Systems

- NOEs
  Hydrogen bonds
  Paramagnetic relaxation enhancement
  ambiguous NOEs; docking (HADDOCK)
  "exact" NOEs (eNOEs)
- Chemical shifts (TALOS)
  Scalar coupling constants
  Ramachandran plot; rotamers
- ^3J scalar coupling constants
- Partially aligned proteins
- Paramagnetic proteins
- Partially aligned proteins
- Known size, shape
- Symmetric multimers; fibrils
- Symmetric multimers; fibrils
- Energy refinement

Conformational restraints in CYANA

- Distance restraints
  - exact distances
  - upper bounds, lower bounds
  - ambiguous distance restraints
  - ensemble-averaged restraints
- Torsion angle restraints
  - single torsion angles
  - multiple torsion angles
- ^3J scalar coupling constants
- Residual dipolar couplings (RDC)
- Pseudocontact shifts (PCS)
- Chemical shift anisotropy (CSA)
- Radius of gyration restraints
- Multimer identity restraints
- Multimer symmetry restraints
- AMBER force field

NOE (Nuclear Overhauser Effect)

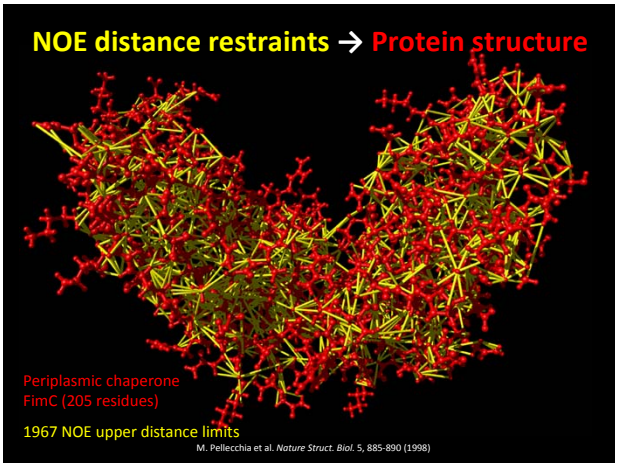
NMR Daten: Integral V von NOESY Kreuzsignalen

Konformationsdaten: obere Schranken für ^{1}H-^{1}H Distanzen, d

Fuer isoliertes Spinpaar im starren Molekül:
$V = C/d^6$ mit C = konstant

Eigenschaften:
- nur kurze Distanzen < 5 Å messbar
- dichtes Netzwerk bzgl. der Sequenz kurz- und langreichweitiger Distanzschranken
- viele ^{1}H Atome im Molekül → "Spindiffusion"
- interne Bewegungen → nicht-lineare Mittelung
- Bestimmung von C?
- Überlapp → mehrdeutige Zuordnung, verfälschte Integrale

NOE distance restraints → Protein structure



^3J skalare Kopplungen

NMR Daten: Aufspaltung eines Signals

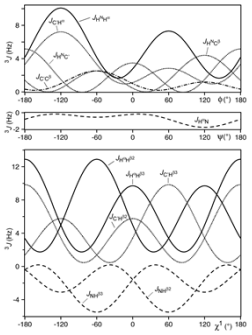
Konformationsdaten: Einschränkungen von Torsionswinkeln, θ

Karplus-Kurve: $^3J(\theta) = A\cos^2\theta + B\cos\theta + C$
mit emprischen Konstanten A, B, C

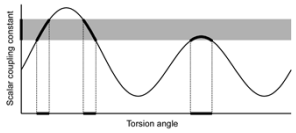
Zum Beispiel: $^3J_{HNH\alpha}(\phi)$, $^3J_{H\alpha H\beta}(\chi^1)$

Eigenschaften:
- Information nur über lokale Konformation
- mehrdeutige Beziehung $^3J \leftrightarrow \theta$

^3J skalare Kopplungen



- $^3J(\theta) = A\cos^2\theta + B\cos\theta + C$
- local information only
- ambiguous relation to torsion angle



H-Brücken

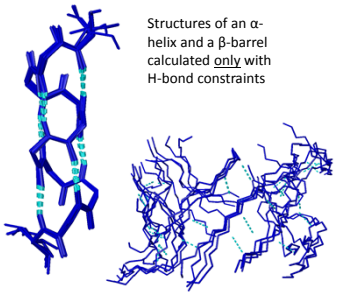
NMR Daten: langsamer $^1H \rightarrow {}^2H$ Austausch + NOEs

Konformationsdaten: Donor-Akzeptor Distanz

Typische H-Brücken: -N-H · · · O=C- in regulären Sekundärstrukturen (Helices, β-Blätter)

Eigenschaften:
- Bzgl. Sequenz mittel- und langreichweitig
- Donor (H) identifizierbar
- Akzeptor (O) nur indirekt bestimmbar (benachbarte NOEs + Annahmen über Sekundärstruktur)

Impact of hydrogen bond restraints



Structures of an α-helix and a β-barrel calculated only with H-bond constraints

- Strong impact on structure
- Direct detection of H-bonds by NMR is possible, but not sensitive
- Without identification of acceptor atom ≈ assumption on secondary structure

Chemische Verschiebungen

NMR Daten: chem. Verschiebungen, δ

Konformationsdaten: (ϕ,ψ) Torsionswinkelbereiche

Komplexe Beziehung: $\delta \leftrightarrow (\phi,\psi)$

Eigenschaften:
- einfache Messung
- (ϕ,ψ)-Werte aus Datenbank von Proteinen mit bekannter Struktur und chem. Verschiebungen (TALOS)
- Information nur über lokale Konformation

Computational challenges

Three principal challenges of NMR protein structure analysis

1. **Efficiency**
   Spectrum analysis requires (too) much time and expertise.
2. **Size limitation**
   Structures of proteins > 30 kDa are very difficult to solve.
3. **Objectivity**
   Agreement between structure and raw NMR data?

Computational tasks in NMR structure determination

Peak picking	→	Signal frequencies
Shift assignments	→	Spin frequencies
NOESY assignment	→	Structural restraints
Structure calculation	→	3D structure
Refinement, validation	→	Final structure

Use of automation for different stages of PDB NMR structures

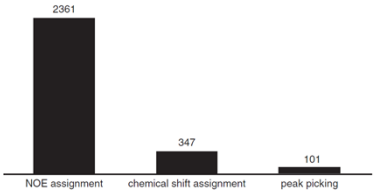


Fig. 4. The use of automation – in terms of PDB depositions – for the different stages of the traditional protocol for NMR protein structure determination. The histograms represent the number of structures returned when searching the PDB for one of the programs published for the respective stages. Exact search strings can be found in the Appendix (Tables A1, A2 and A3).

Guerry, P. & Herrmann, T. *Q. Rev. Biophys.* **44**, 257-309 (2011).

Citations of software in PDB files submitted September 2005–2008

Program	Function	# PDB entries citing	Year of introduction
NMRPipe	Processing, display and peak picking	1,340	1995
CYANA	Structure calculation	1,160	2003
XWinNMR/Topspin	Bruker programs for acquisition and processing	1,043	1997
NMRView	Viewing spectra; peak picking; analysis	910	1994
KUJIRA	Semi-automated processing and structure calc	736	2007
Sparky	Assignment, integration	365	1999
VNMR	Varian programs for acquisition and processing	317	1989
CNS	Structure calculation	242	1998
XPLOR-NIH	Structure calculation	153	2003
XEASY	Semi-automated analysis and assignment	130	1995
ARIA	NOE assignment and structure calculation	122	1995
DYANA	Structure calculation	114	1997
Autostructure	Structure calculation	103	2003
Autoassign	Assignment	82	2001
XPLOR	Structure calculation	75	1992
CcpNmr analysis	Viewing, analysis, assignment	18	2004
Aurelia/Auremol	Semi-automated processing and structure calc	17	2004
ABACUS/CLOUDS	Structure calculation without assignments	4	2002
FLYA	Fully automated structure calculation	3	2006

A much more comprehensive discussion of programs can be found in Gronwald and Kalbitzer (2004). This is a selective list and programs listed here are not necessarily the most cited. References to software that are not given in the text: NMRPipe (Delaglio et al. 1995); Sparky T. D. Goddard and D. G. Kneller, SPARKY 3, University of California, San Francisco, http://www.cgl.ucsf.edu/home/sparky/; CNS (Brünger et al. 1998); XPLOR-NIH (Schwieters et al. 2003); XEASY (Bartels et al. 1995); XPLOR A.T. Brünger, X-PLOR Version 3.1, Yale University Press, NewHaven/London, 1992

Williamson, M. P. & Craven, C. J. *J. Biomol. NMR* **43**, 131–143 (2009).

Peak picking

Computational tasks in NMR structure determination

Peak picking	→	**Signal frequencies**
Shift assignments	→	Spin frequencies
NOESY assignment	→	Structural restraints
Structure calculation	→	3D structure
Refinement, validation	→	Final structure

Automatically picked peaks for the protein ENTH

Spectrum	Expected peaks	Measured peaks [%]	Missing peaks [%]	Artifact peaks [%]	Deviation
^{15}N-HSQC	164	138	14	58	0.138
^{13}C-HSQC	685	113	12	51	0.434
HNCO	134	150	12	63	0.308
HN(CA)CO	269	74	35	16	0.449
HNCA	274	116	18	39	0.331
HN(CO)CA	134	150	10	61	0.395
CBCANH	529	112	29	47	0.458
CBCA(CO)NH	270	149	13	63	0.405
HBHA(CO)NH	365	134	35	75	0.510
(H)CC(CO)NH	451	88	34	25	0.530
H(CCCO)NH	664	56	57	21	0.673
HCCH-COSY	2469	97	66	70	0.609
(H)CCH-TOCSY	2449	136	45	93	0.568
HCCH-TOCSY	3574	44	66	20	0.632
^{15}N-edited NOESY	1776	120	47	74	0.486
^{13}C-edited NOESY	5958	144	48	103	0.495
Total	**20165**	**99**	**49**	**69**	**0.524**

Missing peaks: Percentage of expected peaks that cannot be mapped to a measured peak using the manually determined reference chemical shifts. **Artifact peaks:** Percentage of measured peaks to which no expected peak can be mapped. All percentages are relative to the number of expected peaks. **Deviation:** Root-mean-square deviation between the chemical shift position coordinates of the measured peaks to which an expected peak can be mapped and the corresponding reference chemical shift value, normalized by the chemical shift tolerances of 0.03 ppm for ^{1}H and 0.4 ppm for ^{13}C and ^{15}N.

Resonance assignment

Computational tasks in NMR structure determination

Peak picking	→	Signal frequencies
Shift assignments	→	**Spin frequencies**
NOESY assignment	→	Structural restraints
Structure calculation	→	3D structure
Refinement, validation	→	Final structure

NMR resonance assignment is like solving a puzzle…

…with missing pieces
(incomplete signals)

…with additional pieces
(artifacts)

…in the November mist
(low signal-to-noise, line-broadening)

Table A1. *List of references for programs performing automated chemical shift assignment published in the last 15 years*

1		ABACUS (Lemak *et al.* 2008)	31	●	PACES (Coggins & Zhou, 2003)
2		ASCAN (Fiorito *et al.* 2008)	32		PASA (Xu *et al.* 2006)
3		ASSTOOL (Reed *et al.* 2003)	33	●	PASTA (Leutner *et al.* 1998)
4	●	AUTOASSIGN (Zimmerman *et al.* 1997; Moseley *et al.*	34		PINE (Bahrami *et al.* 2009)
5		Bhavesh *et al.* (Bhavesh *et al.* 2001)	35		PISTACHIO (Eghbalnia *et al.* 2005b)
6	?	'Bipartite Matching' (Xu *et al.* 2002)	36		PRODECOMP/SHABBA (Staykova *et al.* 2008)
7	?	CISA (Wan & Lin, 2007a)	37		RIBRA (Wu *et al.* 2006)
8	?	'Contact Based' (Kamisetty *et al.* 2006)	38		SAGA (Crippen *et al.* 2010)
9	?	'Contact Replacement' (Xiong *et al.* 2008)	39		Sanctuary 1 (Xu & Sanctuary, 1993)
10		CONTRAST (Olson & Markley, 1994)	40		Sanctuary 2 (Xu *et al.* 1994)
11	(aa type)	CRAACK (Benod *et al.* 2006)	41		Sanctuary 3 (Li & Sanctuary, 1997b)
12		DYNASSIGN (Schmucki *et al.* 2009)	42		Sanctuary 4 (Li & Sanctuary, 1997a)
13		GANA (Lin *et al.* 2005)	43	●	TATAPRO (Atreya *et al.* 2000)
14	●	GARANT with structure (Bartels *et al.* 1996)	44	?	Wan *et al.* (Wan *et al.* 2003)
15	●	GARANT without structure (Bartels *et al.* 1997)			
16	?	GASA (Wan & Lin, 2007b)			
17		Hus *et al.* (Hus *et al.* 2002)			
18		Inferential (Vitek *et al.* 2006)			
19		IPASS (Alipanahi *et al.* 2009a)			
20	● ?	JIGSAW (Bailey-Kellogg *et al.* 2000)			
21	●	MAPPER (Güntert *et al.* 2000)			
22	●	MARS with structure (Jung & Zweckstetter, 2004a)			
23		MARS without structure (Jung & Zweckstetter, 2004b)			
24		MATCH (Volk *et al.* 2008)			
25		MC_ASSIGN1 (Tycko & Hu)			
26		MONTE (Hitchens *et al.* 2003)			
27		NOEnet (Stratmann *et al.*)			
28		NVR 1 (Langmead & Donald, 2004)			
29		NVR 2 (Langmead *et al.* 2004)			
30		NVR 3 (Apaydin *et al.* 2008)			

● Backbone from peak lists
○ Backbone with other information, e.g. spin systems, fragments, structure, RDCs
● All/sidechain from peak lists
○ All/sidechain with other info
● more than 50 citations

Guerry, P. & Herrmann, T. *Q. Rev. Biophys.* **44**, 257-309 (2011)

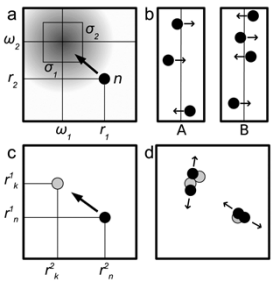
Characteristics of a correct assignment

a) **Shift normality:**
Chemical shifts are consistent with general chemical shift statistics.

b) **Alignment:**
Peaks assigned to the same atom are aligned.

c) **Completeness:**
As many peaks as possible are assigned.

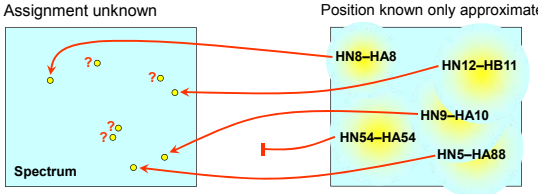
d) **Low degeneracy:**
The number of degenerate peaks is small.



Automated Chemical Shift Assignment

Observed peaks
Position known
Assignment unknown

Expected peaks
Assignment known
Position known only approximately



Assignment = Find **mapping** between expected and observed peaks.

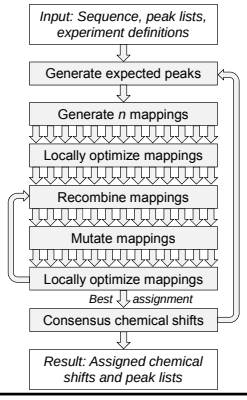
Score for assignment
Presence of expected peaks
Alignment of peaks assigned to the same atom
Normality of assigned resonance frequencies

Optimization of assignment
Evolutionary algorithm combined with local optimization

Elena Schmidt

Christian Bartels *et al.*
J. Comp. Chem. **18**, 139–149 (1997)
J. Biomol. NMR **7**, 207–213 (1996)

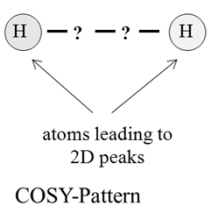
FLYA resonance assignment algorithm

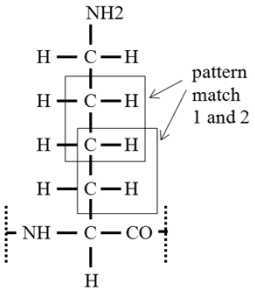


Principles:

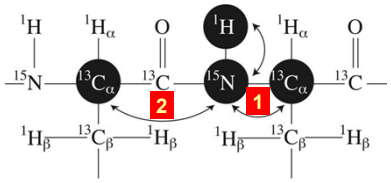
- **Input from NMR experiments:**
  Peak lists (required)
  Chemical shifts, structure (optional)
- **Optimization algorithm:**
  Evolutionary optimization of a population of assignments combined with local optimization.
- **General approach:**
  Use any set of spectra for which expected peaks with can be generated from the primary structure.
- **Exploit redundancy:**
  - Use all information simultaneously
  - Presence of any given peak not required

Peak patterns





Generation of expected peaks
Example: HNCA experiment



Magnetization path entries in CYANA library:

```
SPECTRUM HNCA
1 0.98 H_AMI N_AMI C_ALI
2 0.80 H_AMI N_AMI C_BYL C_ALI
```

Observation probability

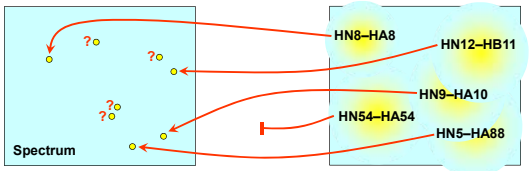
Spectra types

Triple resonance (backbone assignment)	Through-bond (2D & side-chains)	Through-space (NOESY)	Solid-state NMR
• H_CA_NH	• COSY	• NOESY	• NCACB
• HNCA	• TOCSY	• D2ONOESY	• NCACALI
• iHNCA	• D2OCOSY	• N15NOESY	• NCOCACB
• HN_CO_CA	• D2OTOCSY	• C13NOESY	• CANCOCA
• HN_CA_CO	• C13H1 HSQC	• C13NOED2O	• CANCO
• HNCO	• N15H1 HSQC	• CCNOESY	• NCACO
• HCACO	• CB_HARO	• CNNOESY	• CCC
• HCA_CO_N	• N15TOCSY	• NNNOESY	• NCACX
• CBCANH	• HCCH TOCSY		• NCOCA
• CBCACONH	• HCCH COSY		• NCOCA
• HBHACONH	• CCH	2D 3D 4D	• NCOCX
• HNHB	• C_CO_NH		• DARR
• HNHA	• HC_CO_NH		• DREAM
	• HC_CO_NH_4		• PAIN
	• APSY		• NHHC

Automated Chemical Shift Assignment

Observed peaks
Position known
Assignment unknown

Expected peaks
Assignment known
Position known only approximately



Assignment = Find **mapping** between expected and observed peaks.

Score for assignment
Presence of expected peaks
Alignment of peaks assigned to the same atom
Normality of assigned resonance frequencies

Optimization of assignment
Evolutionary algorithm combined with local optimization

Elena Schmidt

Christian Bartels *et al.*
J. Comp. Chem. 18, 139–149 (1997)
J. Biomol. NMR 7, 207–213 (1996)

Global assignment score

Quantifies the quality of the complete assignment of all atoms

$$G = \frac{\sum_{a \epsilon A}\left[w_1(a)Q_1(a) + \sum_{n \epsilon N'_a} w_2(a,n)Q_2(a,n)/b(n)\right]}{\sum_{a \epsilon A_0}\left[w_1(a) + \sum_{n \epsilon N_a} w_2(a,n)\right]}$$

A_0	set of all atoms for which expected peaks exist
$A \subseteq A_0$	subset of assigned atoms
N_a	set of expected peaks for atom a
$N'_a \subseteq N_a$	subset of expected peaks mapped to a measured peak
$Q_1(a)$	measure of normality of the chemical shift of atom a with respect to the general shift statistics; $Q_1(a) \in (-\infty, 1] \ \forall a$
$Q_2(a,n)$	measure of alignment between the chemical shift of atom a obtained from mapping peak n and the average shift of atom a in all its assigned peaks; $Q_2(a,n) \in (-\infty, 1] \ \forall a,n$
$b(n)$	degeneracy of the assignment = number of expected peaks assigned to the same measured peak as expected peak n
$w_1(a), w_2(a,n)$	weights (typically, $w_1(a) \equiv 4$ and $w_2(a,n) \equiv 1 \ \forall a,n$)

Quality measures

Quality measures Q for shift normality and alignment are defined by

$$Q_i = 1 - \frac{q(x_i)}{q(x_i^{(0)})} \quad (i = 1,2)$$

$x_1 = \frac{\bar{f}(a) - f(a)}{\sigma(a)}$ deviation of shift $\bar{f}(a)$ of atom a from statistical average $f(a)$, normalized by its standard deviation $\sigma(a)$

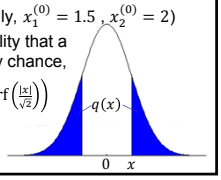
$x_2 = \frac{f(a,n) - \bar{f}(a)}{\epsilon(a)/4}$ deviation of the shift $f(a,n)$ of atom a of peak n and the average shift of atom a in all its assigned peaks $\bar{f}(a)$, normalized by the shift matching tolerance $\epsilon(a)$

$x_i^{(0)}$ deviation "as bad as no assignment" (typically, $x_1^{(0)} = 1.5$, $x_2^{(0)} = 2$)

$q(x) \geq 0$ is the negative logarithm of the probability that a normalized deviation exceeds the given value by chance,

$$q(x) = -\log\left(1 - \int_{-|x|}^{|x|} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt\right) = -\log\left(1 - \mathrm{erf}\left(\frac{|x|}{\sqrt{2}}\right)\right)$$

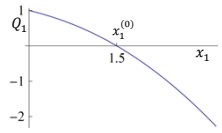
with $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$



Properties of global assignment score

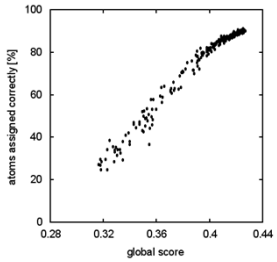
$$G = \frac{\sum_{a \epsilon A}\left[w_1(a)Q_1(a) + \sum_{n \epsilon N'_a} w_2(a,n)Q_2(a,n)/b(n)\right]}{\sum_{a \epsilon A_0}\left[w_1(a) + \sum_{n \epsilon N_a} w_2(a,n)\right]}$$

- Quality measures Q are designed such that
  - $Q = 1$ for a perfect match
  - $Q < 1$ in all other cases
  - $Q = 0$ for a deviation considered "as bad as no assignment"
  - $Q = -\infty$ for an infinitely large deviation

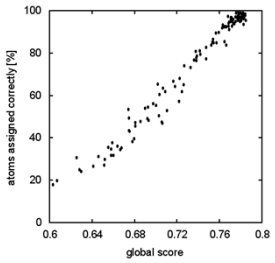


- Hence, the global score G is normalized such that
  - $G = 1$ for a perfect assignment of all atoms
  - $G < 1$ in all other cases
  - $G = 0$ if, for instance, there are either no assignments at all or if all assignments have deviations "as bad as no assignment"
  - $G < 0$ is in principle possible for (very) bad assignments.

Correlation between global score and percentage of correctly assigned atoms



Standard calculation with the full set of 15 peak lists for SH2



Calculation with 7 experiments for the backbone assignment

Data points refer to the current best scored solutions, which were saved during the calculation.

Local assignment score

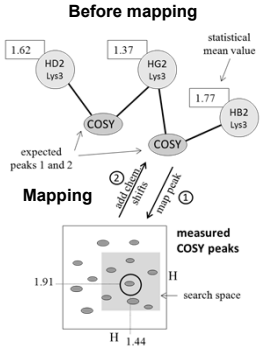
Quantifies the quality of the assignment of a single atom a

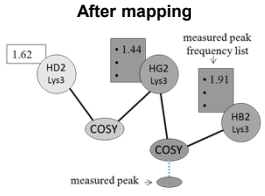
$$L(a) = \frac{\sum_{n \epsilon N'_a} \mathrm{prob}(n)/b(n)}{\sum_{n \epsilon N_a} \mathrm{prob}(n)}$$

N_a set of expected peaks for atom a
$N'_a \subseteq N_a$ subset of expected peaks mapped to a measured peak
$\mathrm{prob}(n)$ probability to observe expected peak n

$L(a) \in [0,1]$
$L(a) = 0$ if *no* expected peaks for atom a are mapped
$L(a) = 1$ if *all* expected peaks for atom a are mapped

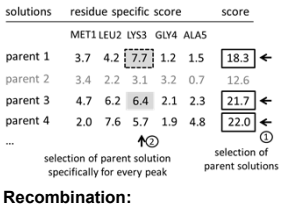
Mapping expected to measured peaks

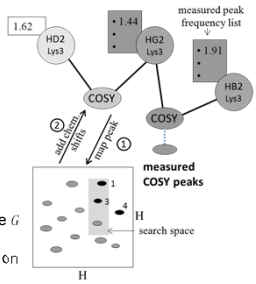




- Not all expected peaks have to be mapped to a measured peak.
- There can be measured peaks to which no expected peak is mapped (mapping is not surjective).
- More than one expected peak may be mapped to the same measured peak (mapping is not injective).

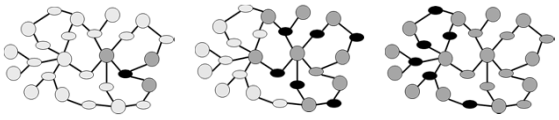
Global optimization





1. **Recombination:**
   - Select parent solutions based on global score G
   - Select an expected peak that fits the search space from a selected parent solution based on the residue-specific part of G
   - Repeat for all peaks through shells of neighboring peaks, if possible. Otherwise, select a new peak
2. **Mutation:**
   - Generate assignments that are not part of a parental solution with probability $e^{-\kappa/T}$, where the "temperature" T is decreased during the calculation.

Consistent mapping of expected peaks during global optimization



First expected peak (black) is mapped to a measured peak. The adjacent atoms obtain a chemical shift entry.

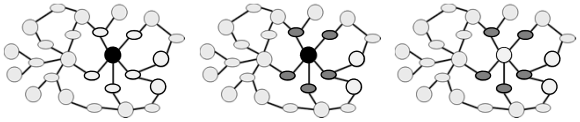
The first shell of neighboring peaks are mapped.

The second shell of neighboring peaks are mapped.

Expected peaks:
- not mapped
- being mapped
- mapped

Atoms:
- unassigned
- being assigned or assigned

Local optimization



An atom with low local assignment score $L(a)$ ( ) is selected, and its assignment removed.

Adjacent peaks ( ) are remapped to measured peaks.

Atom a becomes reassigned.

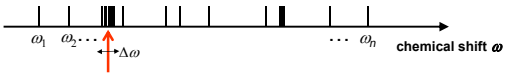
This process is repeated 15,000 times.

Atoms:
- assigned
- unassigned

Expected peaks:
- mapped
- remapped

Consensus chemical shifts

- Ensemble of *n* independently calculated chemical shift values $\omega_1,\ldots,\ \omega_n$ for each nucleus:



- **Consensus chemical shift:** Value ω that maximizes the function

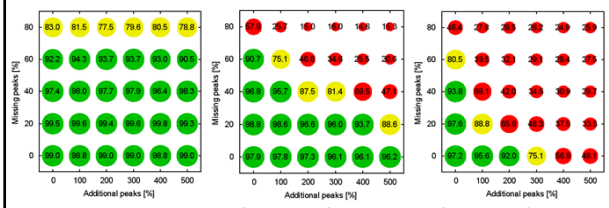
$$\mu(\omega) = \frac{1}{n}\sum_{j=1}^{n} \exp\left(-\frac{1}{2}\left(\frac{\omega-\omega_j}{\Delta\omega}\right)^2\right)$$

$\Delta\omega$ = chemical shift tolerance, e.g. 0.03 ppm for ^{1}H, 0.4 ppm for ^{13}C/^{15}N

- Most individual shifts $\omega_1,\ldots,\ \omega_n$ near consensus value
  → **"safe" assignment**
  Otherwise → unreliable (tentative) assignment

Dependence on quality of input data

Calculations using simulated data for SH2 (15 spectra) with 0–80% missing peaks and 0–500% additional artifact peaks



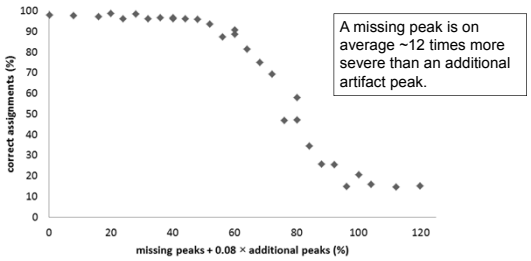
Accurate peak positions

Chemical shift tolerance
0.04 ppm for ^{1}H
0.4 ppm for ^{13}C/^{15}N

Chemical shift tolerance
0.08 ppm for ^{1}H
0.8 ppm for ^{13}C/^{15}N

- >90% correct
- 70–90% correct
- <70% correct

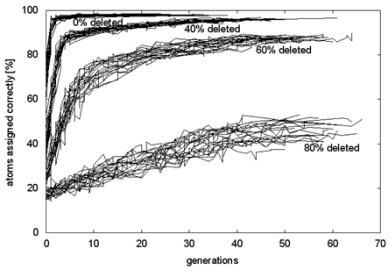
Severity of missing or artifact peaks



A missing peak is on average ~12 times more severe than an additional artifact peak.

Calculations using simulated data for SH2 (15 spectra) with chemical shift tolerance 0.04 ppm for ^{1}H, 0.4 ppm for ^{13}C/^{15}N, 0–80% missing peaks, and 0–500% additional artifact peaks.

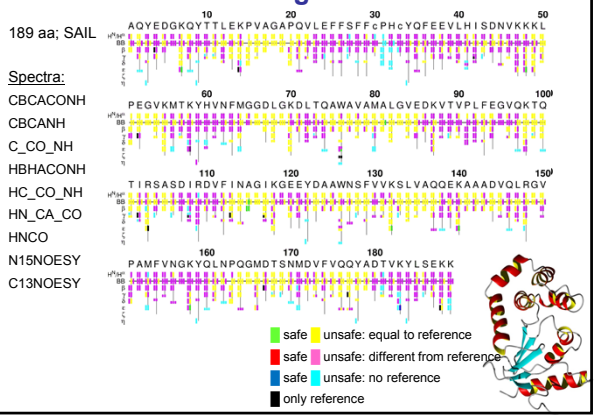
Course of optimization



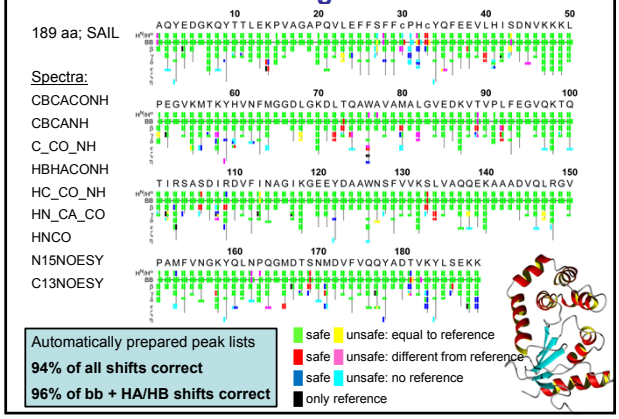
Lower quality input data
↓
Fewer correct assignments
Slower convergence
Greater divergence among individual runs

20 calculations each, using simulated data for SH2 (15 spectra) with chemical shift tolerance 0.04 ppm for ^{1}H, 0.4 ppm for ^{13}C/^{15}N, 0–80% missing peaks, and no additional artifact peaks.

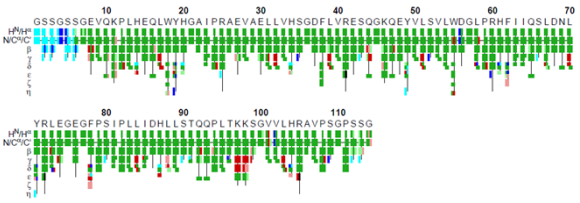
DsbA automated assignment with CYANA



DsbA automated assignment with CYANA



SH2: FLYA assignment



91% correct assignments
99% correct for backbone

Automatically prepared peak lists

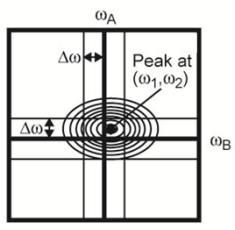
safe unsafe: equal to reference
safe unsafe: different from reference
safe unsafe: no reference
only reference

NOE assignment

Computational tasks in NMR structure determination

Peak picking	→	Signal frequencies
Shift assignments	→	Spin frequencies
NOESY assignment	→	**Structural restraints**
Structure calculation	→	3D structure
Refinement, validation	→	Final structure

Ambiguity of chemical shift based NOE assignment



$|\omega_1 - \omega_A| < \Delta\omega \quad |\omega_2 - \omega_B| < \Delta\omega$

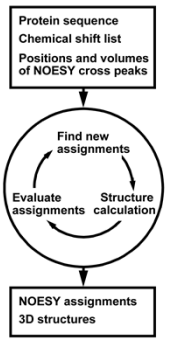
In general, several different ^{1}H chemical shifts ω_A, ω_B match the position of a NOESY peak within the experimental uncertainty $\Delta\omega$.

→ Assignment ambiguity

Manual assignment is very cumbersome!

Automated NOESY assignment and structure calculation

- Automated methods are
  - much faster
  - more objective
- Problems may arise because of
  - imperfect input data
  - limitations of the algorithms used
- Iterative process: All but the first cycle use the structure from the preceding cycle.
- The first cycle is important for the reliability of the method.



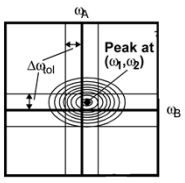
Automated NOE Assignment and Structure Calculation

- Distance restraints from not uniquely assigned NOEs:
  → Ambiguous distance restraints
- Reduction of assignment ambiguity prior to the structure calculation:
  → Network-anchored assignment
- Robustness against erroneous assignments:
  → Constraint combination

T. Herrmann, P. Güntert, K. Wüthrich. *J. Mol. Biol.* **319**, 209-227 (2002)
P. Güntert. *Prog. NMR Spectrosc.* **43**, 105-125 (2003)

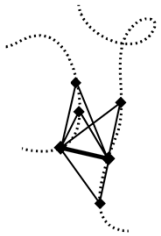
Conditions for valid NOESY assignments

Chemical shift agreement

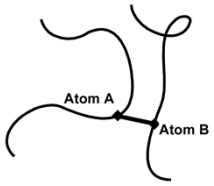


$|\omega_1 - \omega_A| < \Delta\omega_{1ol}$
$|\omega_2 - \omega_B| < \Delta\omega_{1ol}$

Network-Anchoring



Consistency with preliminary structure



$d_{AB} < d_{max}$

NOE assignment probability

(CYANA 2.1, 3.0)

Probability(*assignment to atoms A-B is correct*) =
Probability(*chemical shifts match*) x
Probability(*distance A-B < upper limit*) x
Probability(*other assignments predict NOE A-B*)

$$P_{tot} = P_{shift} \cdot P_{structure} \cdot P_{network}$$

Accept assignments with $P_{tot} > P_{min}$ (= 20%)

Ambiguous distance restraints

$$d_{\text{eff}} = \left(\sum_k d_k^{-6} \right)^{-1/6} \leq b$$

b: upper distance bound
d_k: distance for assignment possibility k
$\sum_k$: sum over all assignment possibilities

- Restraint with multiple assignments
- If one assignment possibility leads to a sufficiently short distance, then the ambiguous distance restraint will be fulfilled.

→ The presence of wrong assignment possibilities has no (or little) influence on the structure,
as long as the correct assignment possibility is present.

Nilges et al., *J. Mol. Biol.* **269**, 408–422 (1997)

Properties of ambiguous distance restraints

$$d_{eff} = \left(\sum_k d_k^{-6} \right)^{-1/6}$$

- d_{eff} is never longer than any of the individual distances d_k:

  $d_{eff} \leq d_k$ for all k

- d_{eff} is close to the smallest individual distance:

  $d_{eff} \approx d_1$ if $d_1 << d_2, d_3, \ldots$

- Examples: $d_1 = 3$ Å, $d_2 = 10$ Å → $d_{eff} = 2.9996$ Å
  $d_1 = 3$ Å, $d_2 = \ldots = d_{10} = 10$ Å → $d_{eff} = 2.9967$ Å

Information content of NOEs

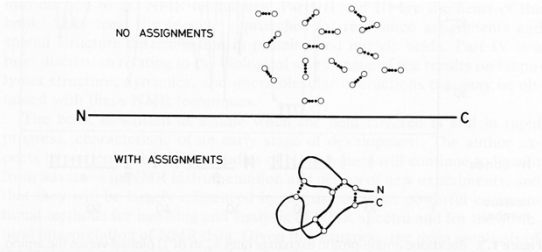
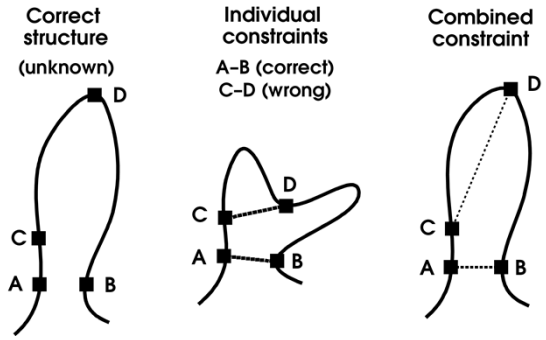


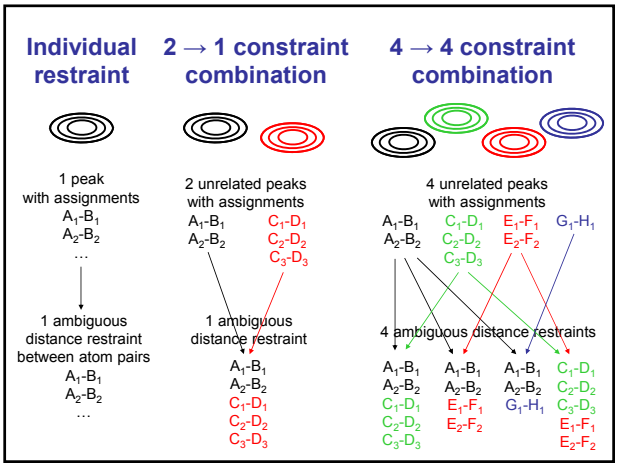
Figure 1.1. Information content of ^1H–^1H NOE's in a polypeptide chain with and without sequence-specific resonance assignments. Open circles represent hydrogen atoms of the polypeptide. The polypeptide chain is represented by the horizontal line in the center.

Constraint Combination



Constraint combination

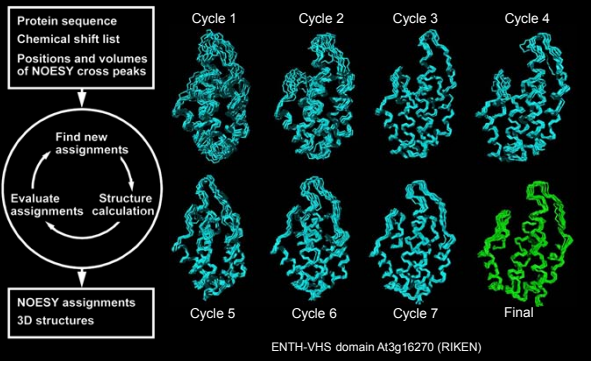
- **Problem:** Peaks with wrong medium- or long-range assignments may severely distort the structure, especially in the first cycles of automated NOE assignment and structure calculation, and may lead to convergence to a wrong structure.
- **Idea:** From two long-range peaks each, combine the assignments into a single distance restraint.
  → Occurrence of erroneous restraints is reduced.



Effect of constraint combination

- Example: 1000 long-range peaks, 10% of which would lead to erroneous restraints.
- Individual restraints:
  1000 constraints, 1000 x 0.1 = 100 wrong (10 %)
- 2 → 1 constraint combination:
  500 restraints, ~500 x 0.1^2 = 5 wrong (~1%)
- 4 → 1 constraint combination:
  1000 restraints, ~1000 x 0.1^2 = 10 wrong (~1%)

Automated NOESY assignment and structure calculation with CYANA



Structure calculation

Computational tasks in NMR structure determination

Peak picking	→	Signal frequencies
Shift assignments	→	Spin frequencies
NOESY assignment	→	Structural restraints
Structure calculation	→	**3D structure**
Refinement, validation	→	Final structure

Structure calculations

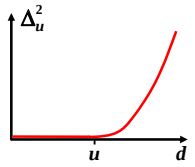
- Structure calculation programs try to fold a protein into a three-dimensional structure that agrees with the measured data.
- Differences between measured data and the structure are manifested as violations of conformational restraints.
- Violations cause forces that act on the molecule, driving it towards minimal (pseudo)energy and optimal agreement with the measured data.
- The target function (pseudoenergy) is the sum of squares of the violations.
- The energy landscape of this target function is complex and has many local minima.

CYANA target function

$$T = \underbrace{\sum \Delta_u^2}_{\text{upper distance limits (NOEs)}} + \underbrace{\sum \Delta_l^2}_{\text{lower distance limits (steric)}} + \underbrace{\sum \Delta_a^2}_{\text{torsion angle restraints}} + \cdots$$

$\Delta_u, \Delta_l, \Delta_a$: restraint violations,

e. g., $\Delta_u = \begin{cases} d-u & \text{if } d > u \\ 0 & \text{otherwise} \end{cases}$

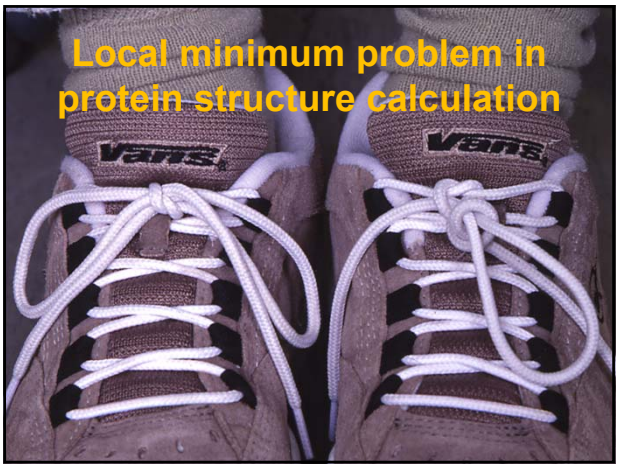


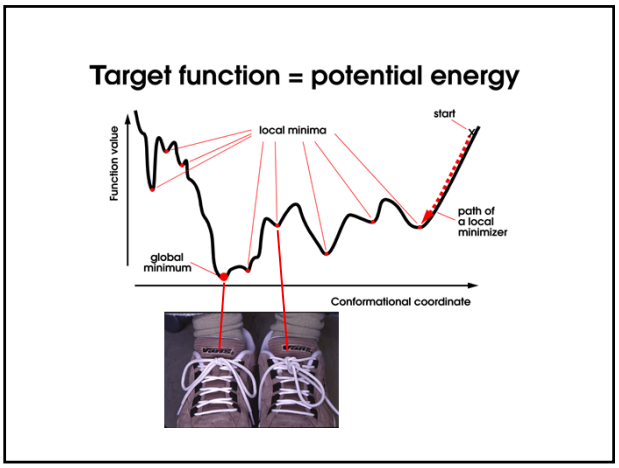
Strukturberechnungsalgorithmen

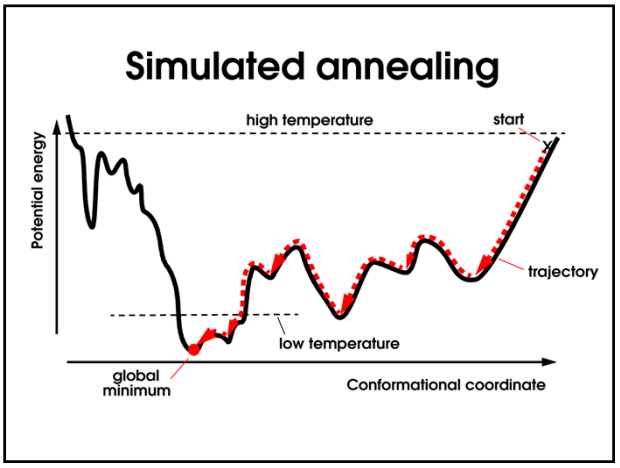
- Frühere Methoden:
  - Interaktiver Modellbau
  - Distanzgeometrie
  - Minimierung einer variablen Zielfunktion
- Simulated annealing:
  - Monte Carlo
  - Moleküldynamiksimulation im kartesischen Raum
  - Moleküldynamiksimulation im Torsionswinkelraum

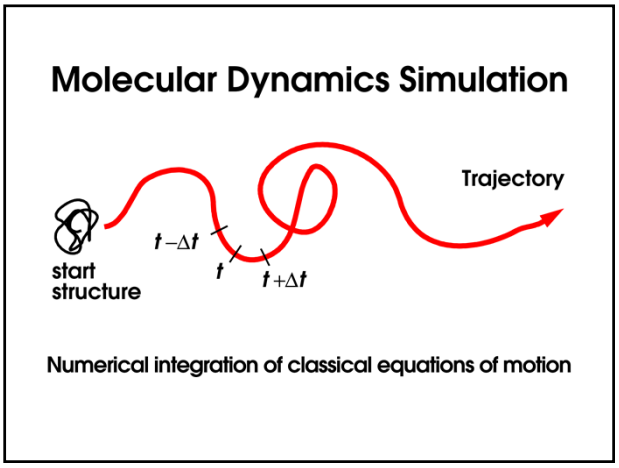
Ist NMR Strukturberechnung möglich?

- Grundsätzlich:
  - NOEs messen nur kurze Distanzen < 5 Å
  - ungenaue obere Schranken
  - Kann damit die globale Struktur eines 30 Å großen Proteins bestimmt werden?
    JA, wenn genügend Daten da sind.
- Praktisch:
  - Zielfunktion hat viele lokale Minima
  - Kann eine (fast) optimale Struktur gefunden werden?
    JA.





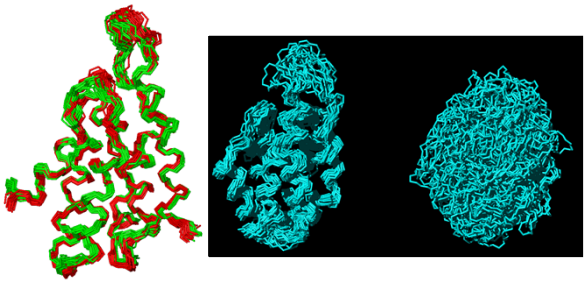




Strukturbündel

- 100 Startstrukturen mit zufälligen Torsionwinkeln
- 100 unabhängige simulated annealing Läufe mit:
  - gleichen experimentellen Daten
  - unterschiedlichen Starttrukturen
- Auswahl der 20 “besten” Strukturen mit den tiefsten Zielfunktionswerten
- Sampling des Konformationsraums?

Strukturbündel



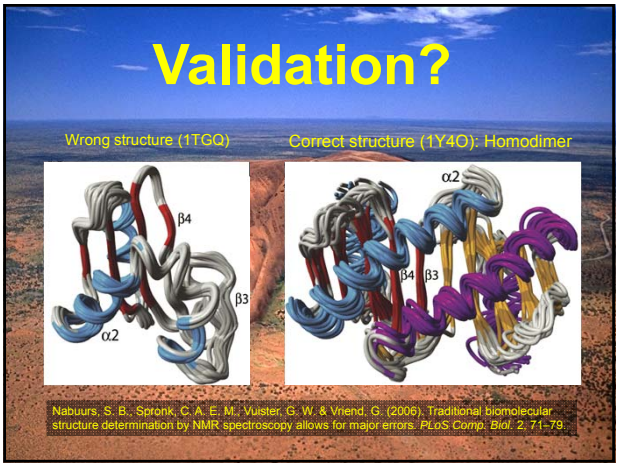
RMSD 0.8 Å RMSD 1.3 Å RMSD 6.3 Å

ENTH-VHS domain At3g16270

Validation

Computational tasks in NMR structure determination

Peak picking	→	Signal frequencies
Shift assignments	→	Spin frequencies
NOESY assignment	→	Structural restraints
Structure calculation	→	3D structure
Refinement/validation	→	**Final structure**

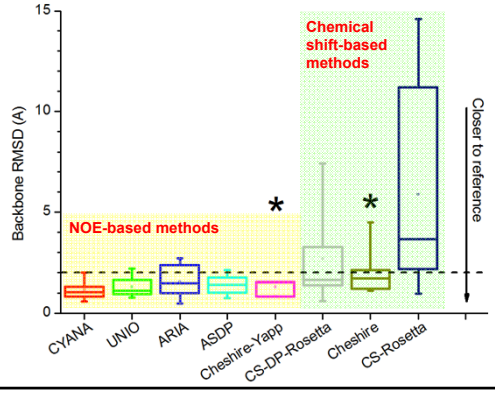


CASD-NMR: Critical Assessment of Structure Determination by NMR

- **Evaluation of current algorithms for automated NOESY assignment and structure calculation**
- **Blind test (analogous to CASP):**
  - *NMR data are provided 8 weeks before the release of the structure by the PDB.*
  - *Structures obtained by different algorithms are collected before the original PDB structure is released.*
- **Open to anybody for providing data and for calculating structures by automated methods**
  - *In 1st round: 10 protein NMR data sets, 7 algorithms.*

Rosato, A. *et al.*, *Nature Methods* 6, 625–626 (2009)

CASD-NMR results: Structure accuracy



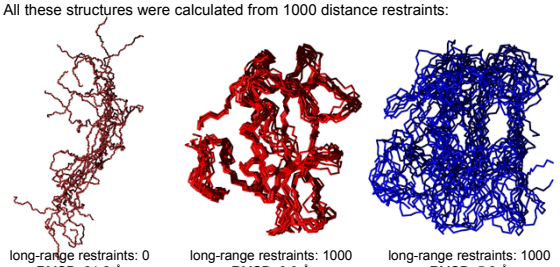
CASD-NMR results: Correlation between accuracy and validation scores

	DP-score	Verify3D	Prosall	Procheck (phi-psi)	Procheck (all)	MolProbity Clashscore
RMSD	-0.66	-0.14	-0.16	0.11	0.26	0.07

Information content

How to characterize conformational restraints for protein structure determination in a concise but meaningful way?

All these structures were calculated from 1000 distance restraints:



long-range restraints: 0
RMSD: 21.2 Å
Information content: 125

long-range restraints: 1000
RMSD: 0.9 Å
Information content: 719

long-range restraints: 1000
RMSD: 5.8 Å
Information content: 125

Information content of distance restraints

Table I
Statistics of the 20 Final Solution Structures of the hPrp18 SFM Domain

Completeness of resonance assignments	
Backbone (%)	96.8
Side chain (%)	99.2
Distance restraints	
Total NOE	1021
Intraresidue	320
Sequential ($\|i-j\|=1$)	278
Medium-range ($1<\|i-j\|<5$)	243
Long-range ($\|i-j\|\geq 5$)	180
Total dihedral angle restraints (TALOS)	
ϕ/ψ	28/27
CYANA target function (Å^2)	0.019
Structure statistics	
NOE restraint violations	
Number > 0.10 Å	0
Maximum (Å)	0.12
Dihedral angle restraint violations	
Number > 2.5°	0
Maximum (°)	0.10
Energies (kcal/mol)	
Mean restraint violation energy	2.91
Mean AMBER energy	−3102.07
Ramachandran plot statistics (%)[a]	
Residues in most favored regions	94.7
Residues in additional allowed regions	5.3
Residues in generously allowed regions	0
Residues in disallowed regions	0
RMSD from the average structure (Å)[a]	
Backbone atoms	0.39
Heavy atoms	1.02

He et al. *Proteins* **80**, 968-974 (2012)

- Counting the number of restraints is not very informative because individual restraints have widely different impact on defining the three-dimensional structure.