Protein Crystallography

T. E. Schrader

This document is a slightly revised version of an article originally published in

Manuel Angst, Thomas Brückel, Dieter Richter, Reiner Zorn (Eds.):
Scattering Methods for Condensed Matter Research: Towards Novel Applications at Future Sources
Lecture Notes of the 43rd IFF Spring School 2012
Schriften des Forschungszentrums Jülich / Reihe Schlüsseltechnologien / Key Technologies, Vol. 33
JCNS, PGI, ICS, IAS
Forschungszentrum Jülich GmbH, JCNS, PGI, ICS, IAS, 2012
ISBN: 978-3-89336-759-7

E 10 Protein Crystallography

T. E. Schrader
Jülich Centre for Neutron Science
Forschungszentrum Jülich GmbH

Contents

1 Introduction

Apart from water proteins are the most abundant molecules in living cells. They are constantly synthesized in the cell by the well known transcription and translation mechanism where first the DNA is read out to produce a messenger RNA which encodes the proteins and in the subsequent translation process the protein is synthesized by the ribosomes which itself is a protein/RNA-complex. Proteins fulfil numerous functions in the cell, for example enzymatic catalysis which enhances the speed with which molecules like fatty acids are synthesized or transport and storage of important molecules like oxygen or they are involved in immunology, just to name a few [1]. In order to perform these functions proteins adopt a unique three dimensional structure with a carefully controlled mixture of flexibility and stiffness. For an understanding of their function knowledge of this three dimensional structure is a prerequisite. Ideally one would like to produce a movie where one can follow the functioning protein in action in slow motion with atomic resolution. In practice, there are techniques available which have a sufficient time resolution (in the fs-regime) but do only provide very limited structural information like time resolved infrared spectroscopy. On the other hand there are methods which provide full atomic resolution but with essentially no time resolution. With those methods one often stops the functioning process of a protein under investigation in an intermediate state by trapping methods using inhibitor molecules which stop the catalytic process of the protein leaving it trapped it in a certain intermediate state.

This article will focus on the latter static techniques among which X-ray protein crystallography is the most widely used one. It will also introduce the method of neutron protein crystallography since there are some similarities but also some differences to X-rays as probes. Finally an example case study is discussed where both techniques give complimentary information. But at first a short introduction into the basic structural properties of proteins is given.

2 Some Basics about Proteins

2.1 Amino acids as the building blocks of proteins

Proteins consist of a chain of amino acids. In that sense they are biopolymers and since they in general have charged side chains they can also be called polyelectrolytes. The first information about a protein is therefore the number of amino acid residues it contains. This number can span quite a wide range between 10 and 25 000. In a historical nomenclature often the term “polypeptide” is used for a small protein containing between 10 and 100 amino acids. Amino acid chains with a smaller number of amino acids than 10 are often named oligopeptides. Typically a protein contains 100 amino acids. The average molecular weight per amino acid is around 100 g/mol. This provides a possibility to calculate an estimate for the molecular weight of a protein from the number of its residues. Despite there are many different amino acids (or to be more exact: 2-amino carboxylic acids) present in living organisms not all of them are used to build up proteins. The proteinogenic amino acids are shown in Figure 1. Covalently attached to the central C-atom, the so called C_α atom are an amino functional group, a carboxylic group, the side chain atoms and finally one hydrogen atom. Since the C_α atom has a sp^3 hybridisation sterically all four constituents point into the corners of a tedrahedron. Since (apart from the amino acid glycine) this C_α atom has four different constituents it forms a chiral centre in the Fischer sense. So two different arrangements of these four constituents are possible which lead to the L- and D-enantiomers of the corresponding amino acid, according to Fischer's convention. But in nature only the L-enantiomers are found as building blocks of proteins.

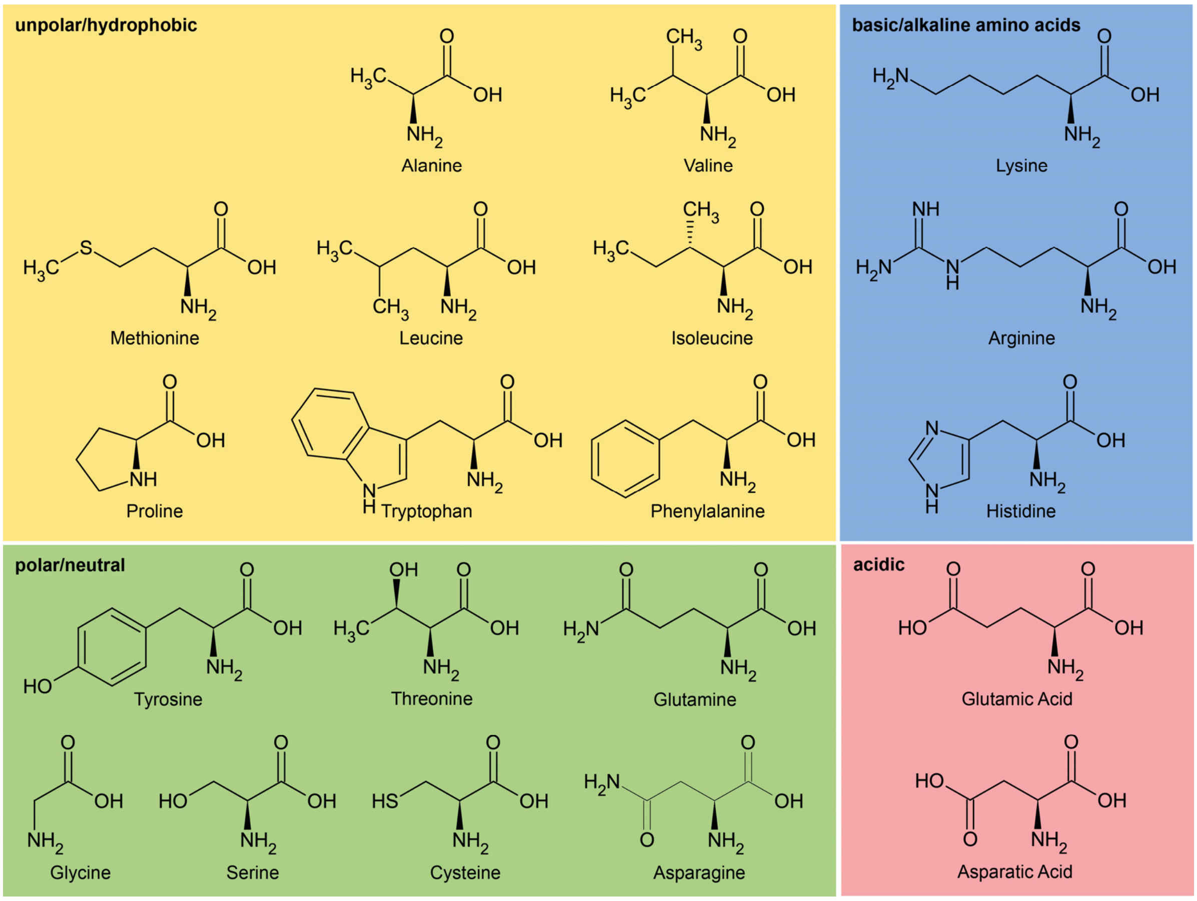

Fig. 1: *A compilation of all 20 amino acids found in natural proteins. The N-terminal amino group is here shown in its neutral charge state pointing to the bottom of the page. Covalently attached to it is the Cα atom which carries the corresponding side chain group (Adapted from http://upload.wikimedia.org/wikipedia/ commons/7/7d/Overview_proteinogenic_amino_acids-DE.sv).*

The ribosomes in a living cell synthesize the proteins according to the code read from the messenger RNA. This process is called translation[1]. Hereby in a step by step fashion one amino acid after the other is linked via a peptide bond between the carboxylic group of the existing amino acid chain and the amino functional group of the newly added amino acid. A water molecule is released per peptide bond formed (see Fig. 2). Hence, the peptide bond formation is a poly-condensation. The inverse process requires adding a water molecule and is named hydrolysis. Formation of a peptide bond requires free energy, so the inverse process is

[1] The term transcription is used to describe the production of the messenger RNA by reading the DNA code.

exothermic but happens on a very long time scale between 10 to 1000 years without enzymatic catalysis.

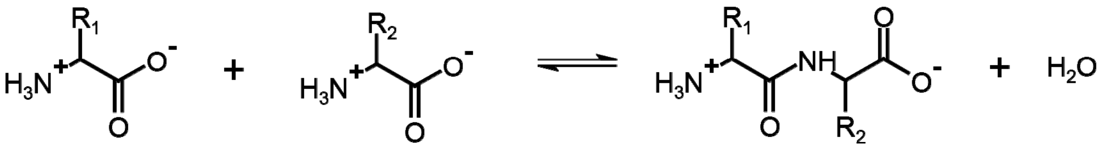

Fig. 2: *A peptide bond forms between two amino acids. As a result a water molecule is released. The inverse process is called hydrolysis.*

The sequence of amino acids is the **primary structure** of the protein. It can be displayed as a line of text with a three letter code representing one amino acid. It is a common convention that the line of text starts at the N-terminal end of the amino acid chain i. e. with the amino acid with side chain R_1 in the example given in Fig. 2 on the right.

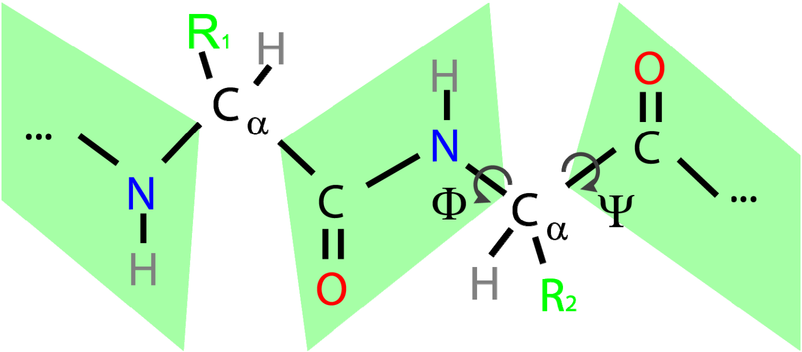

Fig. 3: *Due to a partial double bond character of the C-N peptide bond the rotation around it is hindered by a steep potential. This causes the four atoms OCNH to form a planar structure (marked in green). The only remaining degrees of freedom per amino acid along the backbone are the two dihedral angles Φ and Ψ.*

Another specialty of the peptide bond is its partial double bond character of the chemical bond between the carbon and nitrogen atom. The lone electron pair at the nitrogen atom is delocalized and has some existence probability between the atoms forming the peptide bond. This causes the bond length to shrink below the value of a single bond. As a consequence, the rotation around the CN-bond is hindered and the four atoms OCNH form a planar geometry denoted by a green polygon in Fig. 3[2]. This is why one amino acid only contributes two degrees of freedom to the amino acid backbone which are denoted by the dihedral angles Φ and Ψ, for their definition see Fig. 3. Another property of this planar set of atoms is their potential to from hydrogen bonds. Hereby, the oxygen atom is partly negatively charged and

[2] Only in rare cases one finds a distorted planar geometry. The dihedral angle ω for the rotation around the CN-bond is in this case not equal to +180°. The potential for rotation around the CN-bond has a second local minimum at $\omega=0°$, which corresponds to the peptide bond in its cis configuration. This is frequently found in conjunction with the amino acid proline.

is a hydrogen bond acceptor. The hydrogen atom covalently linked to the backbone nitrogen atom carries a positive partial charge and is therefore a hydrogen bond donor. It is this hydrogen atom which can be easily replaced by a deuterium atom when the protein is dissolved in heavy water (D_2O) for a certain time. This can be seen as a proof for the hydrogen donor capabilities of this hydrogen atom.

2.2 The three dimensional structure of proteins

After leaving the exit tunnel of the ribosome the polypeptide chain folds into a unique three dimensional structure. This process is sometimes assisted by chaperones, which provide a special electrostatic environment, which helps the proteins to fold correctly. Since the backbone of all proteins is the same (i. e. the covalently linked atoms N- C_α -C-N- C_α -C and so forth) the side chains determine this unique three dimensional structure. This structure is stabilized by four different interactions. First of all there is the possibility of establishing **hydrogen bonds** between two parts of the backbone, but also between side chains or between a side chain and a part of the backbone. Another stabilizing mechanism is a formed **salt bridge** between a negatively and a positively charged side chain, e. g. aspartic acid and lysine. The third interaction is the formation of a **hydrophobic cluster** or core. Hereby the surrounding water plays a major role which makes it a mostly entropic effect. It is more favorable for the water molecules to form hydrogen bonds with each other than to stick between some hydrophobic side chains. This is why those side chains tend to be packed together in the folding process resulting in van der Waals interactions among them. The fourth stabilizing moment of a three dimensional fold of a protein is a **disulfide bridge** between two cysteine residues. Often this is used to link two different amino acid chains to form one protein.

When the primary structure only gives the linear sequence of amino acids, the **secondary structure** of a protein denotes all arrangements of the protein backbone stabilized by a regular hydrogen bonding pattern. These hydrogen bonds are solely between different parts of the backbone. There are several structural motifs of that kind which occur frequently in proteins. Some of these motifs have been given a name e .g α-Helix or parallel β-sheet (Fig. 4).

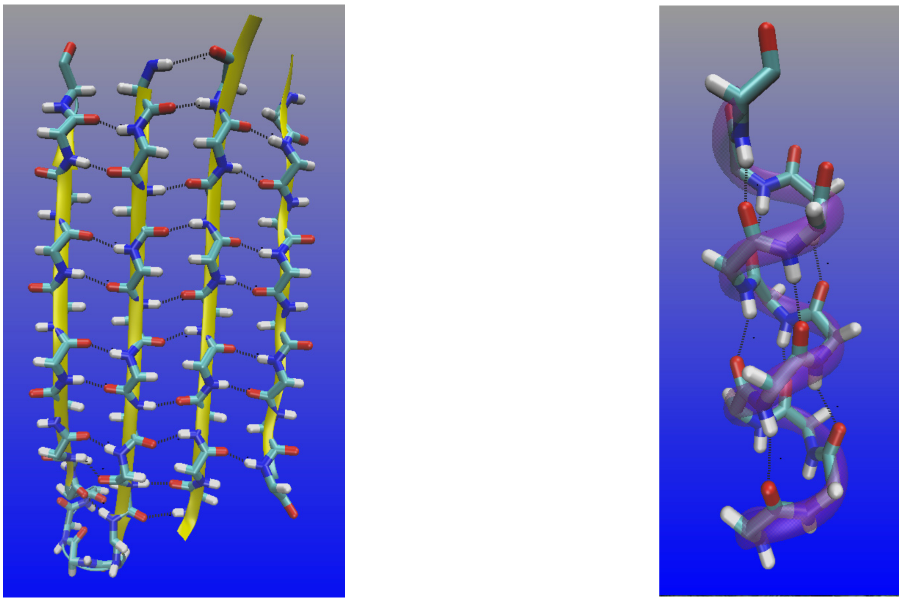

Fig. 4: *On the left the residues 48 to 55, 107 to 130 and 187 to 196 of concanavalin A (pdb code 1XQN) are shown as an example of an anti-parallel β-sheet [2] On the right residues 104 through 114 of sperm whale myoglobin (pdb code1L2K), are shown forming an α-Helix (Images made with VMD version 1.8.7).*

The side chains of the amino acids forming an α-Helix point to the outside perpendicular to the helix axis. Per winding 3.6 amino acids form one winding. The hydrogen bonding pattern can be seen in Fig. 4 or taken from Table 1. Since all hydrogen bonds have a dipole moment which is aligned in parallel in an α-helix a large dipole moment is formed by an α-helix which makes it energetically unfavourable when the number of residues involved exceeds 40 [4].

The β-sheet comes in two flavors a parallel one and an anti-parallel one (shown in Fig. 4). They differ by the hydrogen bonding pattern, but in both cases the side chains point roughly perpendicular to the plane defined by the backbone alternatingly upwards and downwards.

Table 1 Geometic properties of some secondary structure elements (values taken from [4]).

Secondary structure	Frequency	H-bonding	Handedness	Typical Φ	Typical Ψ
α-helix (3.6_{13})	abundant	i to i+4	right	-57°	-47°
3_{10} helix	infrequent	i to i+3	right	-20°	-54°
π-helix (4.4_{16})	rare	i to i+5	right	-57°	-80°
polyproline II	rare	-	left	-78°	+149°
polyclycine II	rare	i to i+3	left	-80°	+150°
parallel β-sheet	abundant	wide pair	-	-119°	+113°
antiparallel β-sheet	abundant	close pair	-	-139°	+135°

Apart from the structures mentioned in Table 1 certain turns form regular hydrogen bonding patterns such they can be considered secondary structural elements. On the bottom left of Fig.

4 with the backbone marked in blue a turn motif is visible. This is also an example of a secondary structure element of proteins. It allows for the amino acid chain to reverse its direction to form the anti-parallel β-sheet.

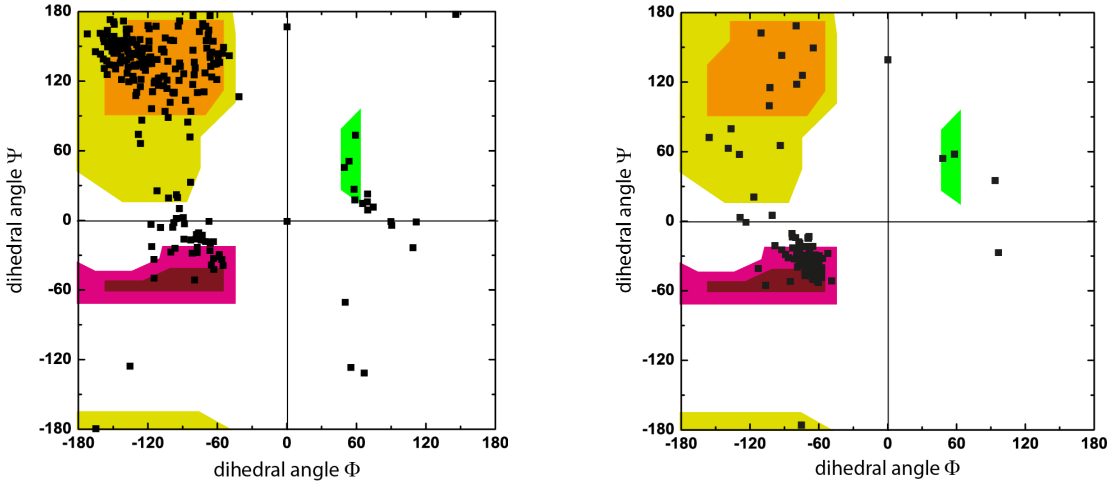

Fig. 5: *Ramachandran plots of the protein concanavalin A on the left and myoglobin on the right. The red area shows dihedral angles typical for α-helices. In the yellow and orange areas dihedral angles typical for β-sheets can be found. The green region corresponds to rare left handed helical arrangements of the protein backbone (Images made with VMD version 1.8.7 [5]).*

Ramachandran plots are especially suited to judge the secondary structure content of a protein. They consist of a scattered plot of all dihedral angles found per residue in the protein (see Fig. 5). On the x-axis all the dihedral angles of Φ and on the y-axis the dihedral angle Ψ is drawn for each amino acid residue resulting in one black filled symbol per residue. Obviously, myoglobin is mostly an α-helical and concanavalin A a β-sheet rich protein.Due to steric hindrances because of the presence of the side chain many combinations of Φ and Ψ are unfavourable (white areas in Fig. 5).

The **tertiary structure** of a protein denotes the three dimensional arrangement of all atoms of the protein in space, including the side chains. This information can be obtained by structural techniques like protein crystallography (will be discussed below), NMR. On that level all interactions mentioned above play a role. Supersecondary structural elements e. g. the α-helix bundle found in myoglobin and the βαβαβ-structure found in dehydrogenases are often seen in a tertiary structure. Another concept to divide the tertiary structure into sub-motifs is to define certain domains which are parts of the protein which can fold to this domain structure without the context of the complete protein. Often these domains provide functional sub-units and their structure is highly conserved throughout the protein family. Figure 6 shows three domains DI to DIII of human urokinase plasminogen activator receptor protein as an example.

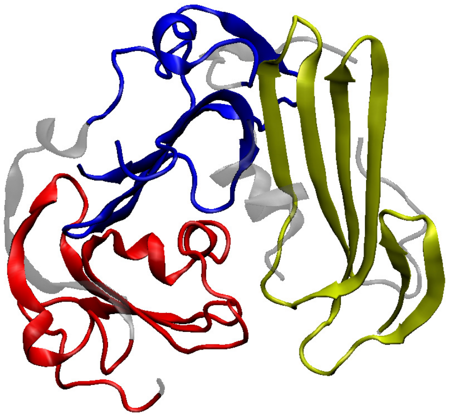

Fig. 6: *Secondary structure plot of human urokinase plasminogen activator receptor [6], a complex protein (pdb code 1YWH). The domains DI (yellow), DII (blue) and DIII (red) are shown. Amino acids not belonging to any domain are depicted in grey.*

Some proteins need more than one amino acid chain to be functional. The arrangement of the different amino acid chains is then referred to as the protein's **quarternary structure**.

2.3 The protein folding problem

The process with which the proteins reach this three dimensional structure is called protein folding and is under intense investigations after C. B. Anfinsen has performed pioneering experiments on denaturation and re-folding [7]. Considering the time a typical protein needs for folding which is of the order of seconds there must be some directive force leading to the correct fold. An exhaustive search of the overall parameter space of all possible dihedral angles Φ and Ψ would take too long for proteins of a typical size of 100 amino acids (Levinthal’s paradox). A possible mechanism for such a directive force is the hydrophobic collapse where all hydrophobic side chains move together to from a hydrophobic core inside the protein. In a different hypothesis secondary structure elements of proteins form first and lead then to the final three dimensional fold.

3 Protein Crystallography

The previous chapter was intended to define some technical terms which describe protein structure in general. The following chapter will show how most of this structural information has been obtained.

The protein data bank (www.rcsb.org) is a well known source of structural information on proteins. Data from many different experimental techniques are entered in a standardized format, a .pdb file which essentially contains not only three dimensional coordinates of all observed atoms in a protein but also information on their mean square displacements. The number of stored entries exceeds 70 000. Among them X-ray crystallography has contributed more than 85 %. The next in line method with more than 8000 entries in the data bank is solution NMR spectroscopy. Electron microscopy as a method was used in more than 250 entries. The remaining methods count less than 100 entries each including neutron protein crystallography. Since the latter technique is represented by an instrument in the Jülich Centre for Neutron Science (JCNS) and because of its similarity to X-ray crystallography it will be given some space in this lecture.

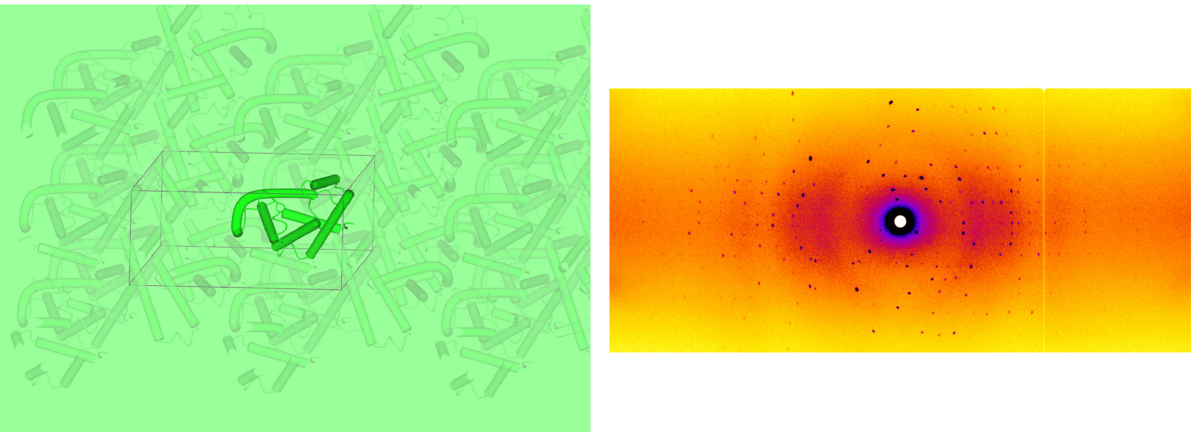

Fig. 7: *Real space arrangement of myoglobin molecules in a crystal of space group P2$_1$ (on the left) versus diffraction pattern (right) of a myoglobin crystal.*

For both techniques X-ray and neutron protein crystallography a single crystal of the protein of interest is required since the scattering of one protein molecule is very weak. Only using very bright X-ray sources in the future (e. g. XFEL) one might be able to gain enough information out of many single protein molecules in solution, averaging many exposures and orientations. But in general a crystal has to be grown, especially large ones in case of neutron crystallography since the neutron luminosity of modern sources is much smaller than the X-ray flux reached by synchrotron sources. To grow sufficiently large crystals is a big challenge in the case of many proteins, especially membrane proteins. Here, one has to adjust a large parameter set of protein concentration, pH condition, salt concentration, percipitant concentration and type just to name a few. The crystal then serves as a noiseless amplipier of the diffraction signal. But the arrangement of proteins in a crystals brings in another advantage, since the orientational averaging can be avoided, which is always present in the solution phase. Fig. 7 shows on the left the regular arrangement of myoglobin molecules in a crystal lattice. The unit cell of the monoclinic lattice (space group P2$_1$) is indicated by black lines. It bears two myoglobin molecules in one unit cell. The picture on the right shows a diffraction pattern recorded with the instrument BioDiff on a myoglobin crystal. The crystal is rotated by ca. 0.5° while recording one diffraction pattern. In order to map the reciprocal space completely one has to put the crystal in many different orientations into the beam and record diffraction patterns as mentioned above. Fortunately, crystal symmetry helps that some orientations are equivalent to each other and need not be recorded.

3.1 An X-ray protein crystallography beamline

Synchrotron beamlines provide extremely high photon flux for X-ray protein crystallography. Due to the high demand from the structural biology community, often more than one protein crystallography beamline is operated at a synchrotron. Those beamlines are optimized for special wavelengths and focal diameters. Here as an example the beamline BL14.2 is used to elucidate a typical X-ray protein crystallography experiment.

Figure 8 shows a schematic view of the beam path of beamline BL14.2. The beam paths to the other beamlines BL14.1 and BL14.3 have been omitted for clarity. A supraconducting magnet-structure interacts with the electron orbit of the storage ring and creates the so called synchrotron radiation. This radiation is used as a white light X-ray source for the three beamlines. A double crystal monochromator is used to select a very narrow energy band (2 eV at 9 keV) from the broad spectrum of the X-ray source. The mechanics of the double crystal monochromator keeps the out-going beam path constant when changing the wavelength. Focussing mirrors and collimators in the beam path ensure an efficient photon transport from the source to the sample and a small beam size at the sample position of 150 µm x 100 µm (height x width, FWHM). The Rayonix MX-225 detector has a pixel size of 37 µm. Without on chip binning one frame amounts to 6144 x 6144 pixels. The exposure time per frame is typically between 3 to 10 seconds.

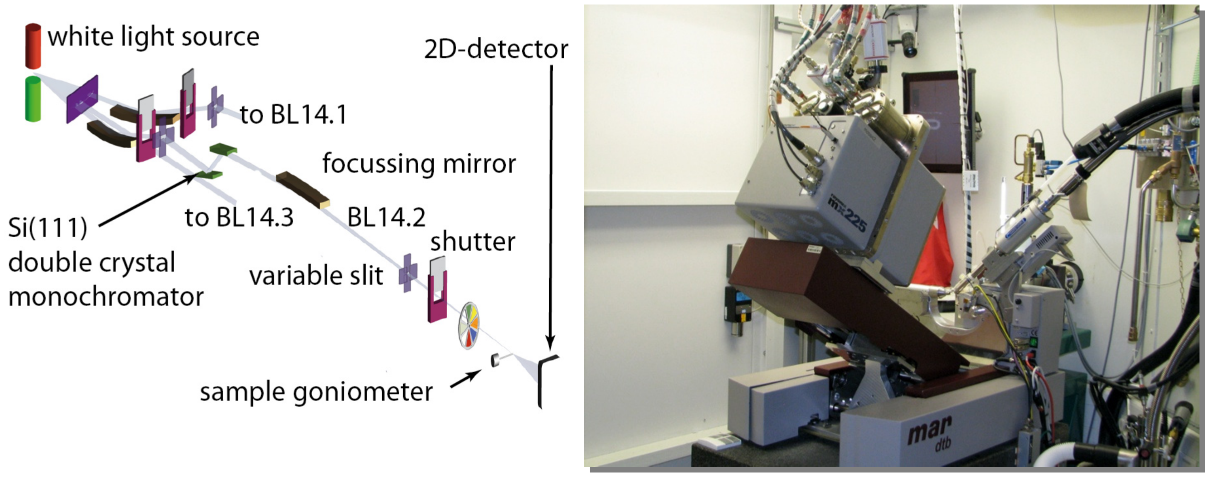

Fig. 8: *Schematic view of the beamline 14.2 of BESSY beginning with the magnet structure which is used as a white light X-ray photon source. On the right a picture taken from the experimental hutch of beamline 14.1 is shown. The beam enters from the right and the sample goniometer is mounted horizontally from the left. A cryostream sample environment to stabilize the sample temperature is discernible pointing towards the sample from the top right corner. Pictures kindly provided by Dr. Uwe Müller, BESSY.*

Typically the sample crystals are kept at liquid nitrogen temperatures to avoid radiation damages. To record a full data set takes about 10-30 minutes. The largest diagonal of a typical protein crystal ranges between 10 to 500 µm

3.2 A neutron protein crystallography instrument

Since X-rays are scattered from the electrons in the crystal and neutrons from the nuclei, hydrogen atoms are hardly seen in X-ray crystallography experiments. Only at very high resolutions of 1 Å or less there is a chance to observe hydrogen atom positions. This resolution is often not within reach because of the crystal quality. Here neutron protein crystallography must be employed to retrieve the hydrogen atom positions. Moreover, neutron scattering can distinguish between different isotopes, especially between hydrogen and deuterium. Whereas from X-ray crystallography the electron density in the unit cell of the crystal can be calculated, neutron protein crystallography yields the nuclear scattering length density, which is a signed quantity. In fact, the coherent scattering length of hydrogen is negative and the one from deuterium is positive (cf. Fig. 9).

Nucleus	atomic number	scattering length [10^{-12} cm]
^{1}**H**	1	**-0.378**
^{2}**H**	1	**0.667**
^{12}**C**	6	**0.665**
^{15}**N**	7	**0.921**
^{16}**O**	8	**0.581**

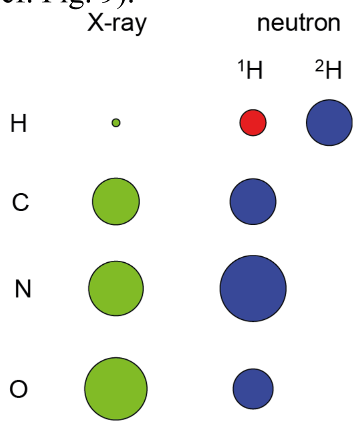

Fig. 9: *The table on the left lists scattering lengths of selected atoms of biological relevance. On the right there is a comparison of X-ray scattering cross section with scattering lengths from neutron scattering. The circles are scaled to match at the carbon atom.*

A major drawback of the method neutron protein crystallography is the required crystal size. Due to the much smaller neutron flux as compared to X-ray flux the crystals required for a neutron crystallography study must be much larger as compared to X-ray crystallography. Here, often crystal diagonals of 1 mm and more have to be reached.
As an example of a neutron diffraction instrument optimised for protein crystallography the instrument BioDiff at the FRM II shall be introduced. It is a collaboration between the Forschungszentrum Jülich (FZJ) and the Forschungs-Neutronenquelle Heinz Maier-Leibnitz (FRM II).

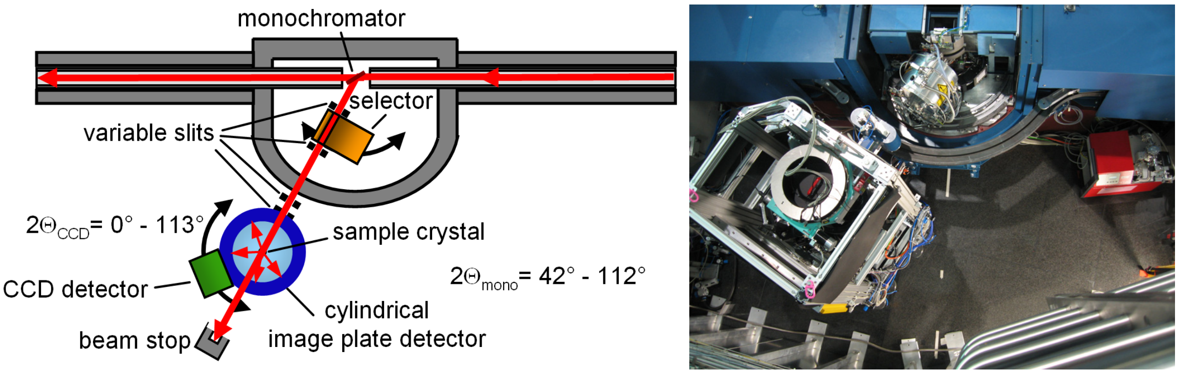

Fig. 10: *Schematic view of the BioDiff instrument (left) and a picture taken from a similar view point with the biological shielding removed (right).*

Figure 10 shows a schematic view of the instrument from the top and a corresponding picture when the biological shielding has been removed. The neutron beam from the cold source of the FRM II reactor enters from the right. By Bragg reflection from a pyrolytic graphite crystal (002-reflex) neutrons are taken out from the white neutron spectrum of the neutron guide NL1 and pass a first boron carbide adjustable slit and a velocity selector. The velocity selector acts as a λ/2 filter. Together with the pyrolytic graphite crystal it forms a monochromator with a Δλ/λ of ca. 2.5 %. Behind the velocity selector the beam passes a second variable slit and the main instrument shutter, named γ-shutter. Additionally, a boron carbide neutron shutter is placed directly after the monochromator crystal for a faster shutter operation. Before entering the detector drum of the image plate detector through a Zirconium window a collimator made out of two manually exchangeable boron carbide apertures with fixed diameters between 3 mm and 5 mm shape the beam to fit to the sample size. At present the sample is usually contained in a glass tube (either a thin walled capillary or a NMR-tube for larger crystals). It is fixed to a standard goniometer which is mounted upside-down from the sample stage on top of the instrument. After passing the sample the main neutron beam exits the detector drum through a second Zirconium window and hits finally the beam stop which consists of a cavity of 4 mm thick boron carbide plates surrounded by a 13 cm thick wall of lead shielding bricks. The cylindrical image plate detector is covering roughly half of the total 4π solid angle. It is 200 mm in diameter and 450 mm in height. It can be read out with three different resolutions of 125 µm, 250 µm and 500 µm. As an alternative, one can lower the image plate detector and swing in a neutron sensitive scintillator which is imaged onto a CCD-chip. This CCD-camera set up serves as a second detector. In particular it is used for a fast alignment of the sample crystal with respect to the neutron beam.

With the image plate detector the diffraction pattern shown in Fig. 7 on the right has been recorded. In fact, a complete crystallographic data set on a myoglobin crystal has been recorded allowing for the calculation of a nuclear scattering length density map. The exposure time was 17 minutes per frame and the crystal was rotated by 0.5° during exposure. 331 frames were recorded before the crystal was manually rotated by ca. 90° in the capillary and another set of 243 frames were recorded. Altogether ca. 9 days of beam time were necessary to record the complete data set. The achieved resolution with sufficient completeness was 1.7 Å. The required time to record this data set was much longer as the 30 minutes from X-ray diffraction.

3.3 Some general aspects of diffraction by a protein crystal

Having recorded a complete data set on a crystal some data treatment is necessary in order to calculate meaningful atom positions. Here only a brief introduction can be given more details can be found in text books [8-9].

Assuming a number of n atoms per unit cell the structure factor of a single unit cell can be written as (see previous lectures):

$$F(\mathbf{S}) = \sum_{j=1}^{n} f_j \exp(2\pi i \mathbf{r}_j \mathbf{S}) \qquad (1)$$

Here $\mathbf{r}_j$ denote the atom position of atom j and $\mathbf{S}$ is the scattering vector perpendicular to the plane which reflects the incident beam. f_j can be seen here either as the scattering length of atom j in the neutron scattering case or the atomic scattering factor in case of X-ray diffraction. One can generalize this approach by switching form the summation to an integration to yield:

$$F(\mathbf{S}) = \int_{unitcell} \rho(\mathbf{r}) \exp(2\pi i \mathbf{r}\mathbf{S}) d^3\mathbf{r} \qquad (2)$$

where $\rho(\mathbf{r})$ is the electron density distribution or the scattering length density respectively. Since a crystal consists of AxBxC unit cells, the structure factor of the crystal can be composed as

$$F_{cryst.}(\mathbf{S}) = F(\mathbf{S}) \sum_{u=0}^{A} \exp(2\pi i u \mathbf{a}\mathbf{S}) \sum_{v=0}^{B} \exp(2\pi i v \mathbf{b}\mathbf{S}) \sum_{w=0}^{C} \exp(2\pi i w \mathbf{c}\mathbf{S}) \qquad (3)$$

The vectors $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ denote basis vectors of the unit cell. For an increasing number of unit cells the sums can be represented by delta functions leading to the Laue conditions for the structure factor being non-zero:

$$\mathbf{a}\mathbf{S} = h, \mathbf{b}\mathbf{S} = k, \mathbf{c}\mathbf{S} = l \qquad (4)$$

This means that one only gets constructive interference, when the scattering vector is perpendicular to planes in the crystal which can be denoted by the index vector $\mathbf{h} = hkl$. For this reason the diffraction pattern of a single crystals shows distinct peaks, the so called Bragg peaks. The Bragg law can be easily derived from equation 4. Figure 11 shows the Ewald sphere construction. It is a tool to construct the direction of the diffracted beam. The Ewald sphere has its origin at the position of the crystal. Its radius is the reciprocal wavelength used in the scattering experiment. The origin of the reciprocal space lattice is placed at the intersection of the sphere with the incident beam direction. Whenever the orientation of the reciprocal space is such that another point of the reciprocal space lies on the Ewald sphere a diffracted beam results in the direction of line running from the centre of the Ewald sphere through that point.

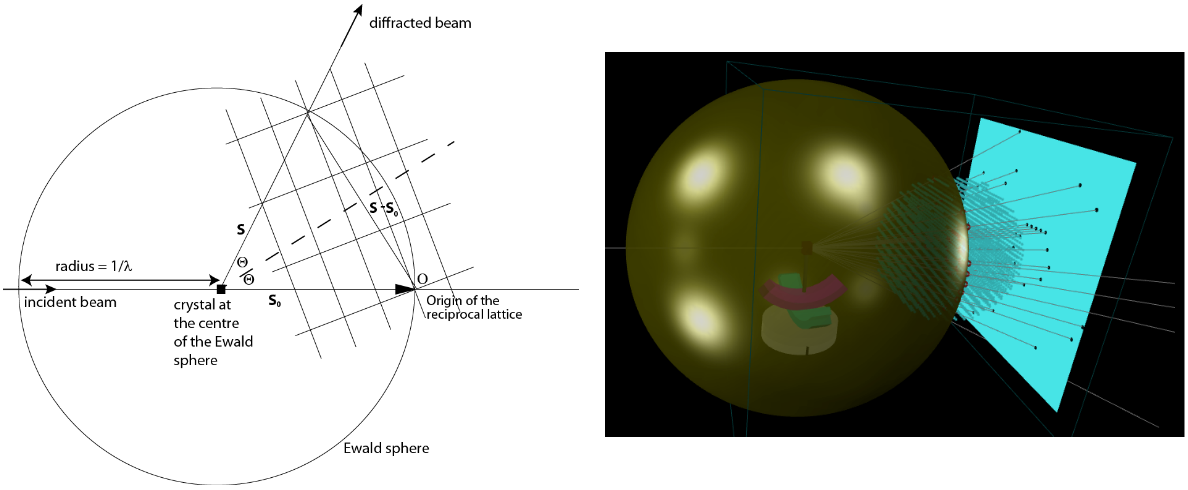
diffracted beam
s - s₀
s
Θ
Θ
radius = 1/λ
incident beam
crystal at the centre of the Ewald sphere
s₀
O
Origin of the reciprocal lattice
Ewald sphere

Fig. 11: *Ewald sphere: On the left the schematic shows how to construct the Ewald sphere. On the right an Ewald sphere (golden colour) construction is shown in three dimensions. The blue square represents a flat two dimensional detector.*

When the crystal is rotated the reciprocal space rotates with it resulting in other lattice points to cause diffracted beams. In practice the incident beam is not strictly monochromatic but has a wavelength distribution which causes the Ewald sphere to be elongated to form a spherical shell of a certain thickness. This increases the number of diffracted beams observed. The beam divergence adds also to its thickness.

So, the positions of the diffracted beams on the detector only depend on the reciprocal lattice. The structure of the protein inside the unit cell is encoded in the amplitude and phase of the structure factor.

To obtain the electron density or the nuclear scattering length density one has to perform the inverse Fourier transformation:

$$\rho(\mathbf{r}) = \frac{1}{V} \sum_{\mathbf{h}} F(\mathbf{h}) \exp(-2\pi i \mathbf{r}\mathbf{h}) \tag{5}$$

Here V is the volume of the unit cell. Unfortunately only the modulus squared of the structure factor is measured as intensity on the detector. The phase information is lost which is known as the phase problem of crystallography.

There are several solutions to the phase problem which are only applicable for the X-ray diffraction case:

- isomorphous replacement: Several crystals of the same crystal structure have to be available for this method. First a crystallographic data set is recorded on an untreated crystal. Then crystals are soaked in at least two different heavy atom salt solutions. In the best case, the different heavy atom ions occupy different regular positions in the unit cell. From these (at least) two crystallographic data sets recorded on the heavy atom treated crystals phase information can be retrieved which is then used to determine the phases of the data set of the untreated crystal.

- anomalous dispersion: Often it is possible to replace one distinct methionine amino acid with an artificial selenomethionine one. The selenium atom has a suitable absorption edge on which anomalous scattering can be performed by tuning the wavelength of the beamline to the anomalous regime. Crystallographic data sets are then recorded at different wavelengths from with the phase information can be calculated. In some cases this approach can also be adopted for sulfur atoms present in naturally occuring cysteine residues.
- molecular replacement: From the primary structure one can search the protein data base (pdb) for proteins with a similar amino acid sequence. If one finds enough fragments which seem to be sufficiently homologous to the unknown structure one can use those fragments for the calculation of initial phases. In further refinement steps these phases can be improved further. Since the number of unique structures entered in the protein data base is growing this method is increasingly favoured over other methods.

The phase problem of the neutron data sets is solved by using the X-ray structure and the molecular replacement technique.

3.4 Model building and refinement

With the data treatment one has now arrived at a contour map $\rho(\mathbf{r})$ be it either a nuclear scattering length density or an electron density. Now the information on the primary structure of the protein is used and either manually or employing software first the backbone is coarsely fitted into the contour map. Then from this model new amplitudes and phases of the structure factor are calculated using eq. 1. The modulus squared of the structure factor is again compared with the data. One could now think of a least square based fitting procedure to find the optimum arrangement of the protein atoms in the unit cell. In practice however maximum likelihood and simulated annealing molecular dynamics simulations are used because those are superior to the least square approach in terms of overcoming local minima. In these molecular dynamics simulations a lot of stereochemical information is used as restraints for example known bond lengths of CC single bonds or bond angles. The agreement between model and observed contour map is often measured by calculating a so called R-factor:

$$R = \frac{\sum_{\mathbf{h}} \left\| F_{obs}(\mathbf{h}) \right| - \left| F_{calc}(\mathbf{h}) \right\|}{\sum_{\mathbf{h}} \left| F_{obs}(\mathbf{h}) \right|} \qquad (6)$$

The index "obs" denotes the observed structure factors and the index "calc" the calculated structure factors from the model. The value of the R-factor lies in the limits between 0 and 1. A good agreement between model and measured data leads to an R-factor of about 0.2. R-factors of 0.5 and above are indicative for a random agreement between model and data.

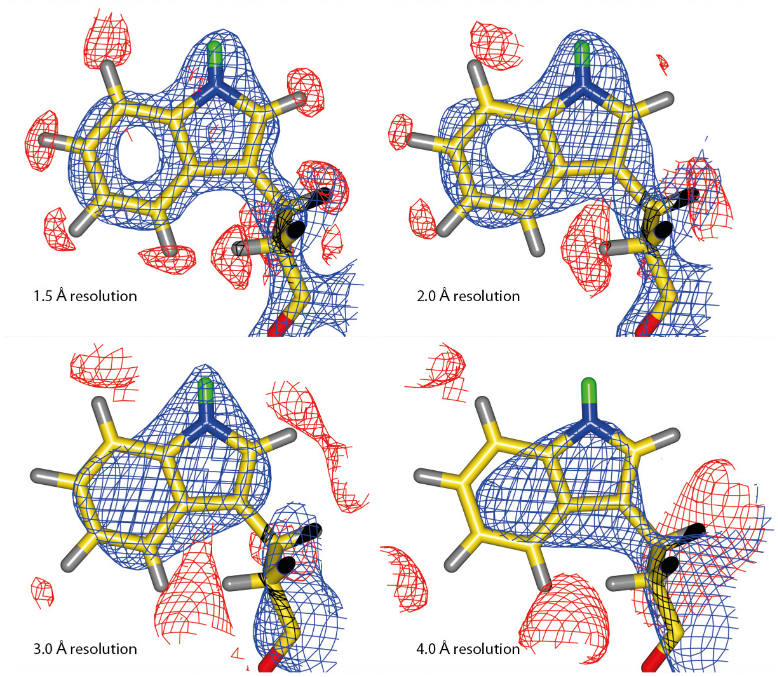

Fig. 12: *The side chain of the amino acid tryptophan no. 7 of myoglobin measured with neutron diffraction at different resolutions. The contour level of the shown nuclear map is +1.5σ (blue) and -1.75σ (red). Exchanged (liable) hydrogen atoms (green) and C- (yellow), N- (blue) and O-atoms (red) appear on a positive contour level. Only not liable hydrogen atoms are seen on the negative contour level.*

But even a good R-factor does not guarantee that the model fits to the data. In fact, it is possible in special cases to obtain a reasonably low R-factor when using the amino acid chain in the wrong direction as a model [10]. Here, Brünger et al. [11] have suggested to divide the measured Bragg reflections into two subset one working set denoted by "A" and one test set denoted by "T". With the working set the fitting procedure is performed, whereas the test set only serves to control the model quality by calculating the R_{free} factor.

$$R_{free} = \frac{\sum_{\mathbf{h}\in T} \left| \left|F_{obs}(\mathbf{h})\right| - \left|F_{calc}(\mathbf{h})\right| \right|}{\sum_{\mathbf{h}\in T} \left|F_{obs}(\mathbf{h})\right|} \tag{7}$$

The test set usually consists of 5 to 10 % of all structure factors, uniformly distributed over the resolution range.

The R-factor and R_{free} factor should not differ too much from each other. In general, it is good practice to always look at the resulting model and its fit to the calculated map after each refinement step. Ramachandran plots can also be used to judge whether the amino acid backbone adopts a reasonable fold. With decreasing resolution (cf. Fig. 12) of the data it becomes more and more difficult to find the right orientations of side chains or even errors in the registry of the protein backbone can occur, whereby for example one amino acid is left out.

4 A case study: Water network around myoglobin

This example is chosen since it nicely shows the interplay between X-ray and neutron crystallography. Myoglobin has been used quite frequently as an example throughout this lecture. Its function is to take over the oxygen molecules from the blood heme proteins in the red blood cells and to transport and store it within the muscle cells. Therefore its binding affinity to oxygene must be stronger than that of the hemoglobin. In order to perform the transport tasks myoglobin has to be highly soluble and movable within the context of a muscle cell. Let alone therefore it is interesting to study the water network surrounding of myoglobin. Since it is a joint neutron and X-ray diffraction study the crystal under investigation has been soaked in D_2O in order to exchange all liable hydrogen atoms with deuterium atoms.

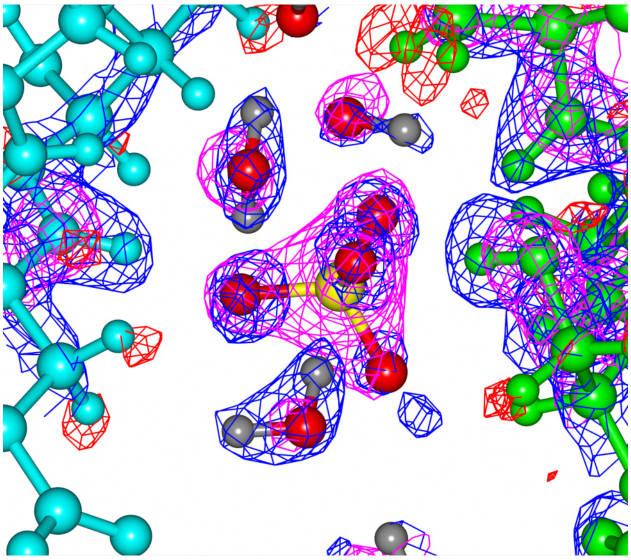

Fig. 13: *Water network in the contact region between two myoglobin molecules in the crystal. In grey colour on the left amino acids 51 to 52 of one myoglobin molecule are shown. On the right amino acids 58 to 60 (from top to bottom) are depicted in green. In the centre of picture a sulfate ion SO_4^- is seen with the sulfur atom shown in yellow, the oxygen atoms shown in red. The deuterium atoms of the water molecules are coloured grey. The X-ray map is shown in magenta at a contour level of +2.7s. The nuclear map is shown at a contour level of -1.75s in red and at +2.3s in blue. Data taken from ref. [3]. (The picture is similar to the one shown in [12] or [13] since it is based on the same data.)*

Figure 13 shows a contact region between two myoglobin molecules. In the centre a sulfate ion is observed. Here, in the nuclear map the central sulfur atom is hardly seen because of its small scattering length. The oxygen atoms of the sulfate ion and of the water molecules are readily observed by both techniques. The deuterium atoms of the water molecules are discernible only in the nuclear map. When the water molecule is fixed by hydrogen bonds, all three atoms can be observed. These water molecules exhibit a triangular shape in the nuclear map. In case it is free to rotate around the OH-axis the nuclear scattering length of the rotating deuterium atom is distributed over a large volume and is therefore averaged out. Those water molecules are denoted as "short ellipsoidal" by Chatake et al. [12]. The long ellipsoidal water molecules are fixed at both deuterium atoms but only the oxygen can rotate freely around the DD-axis. In the fourth case only the oxygen atom is observed. In this case the orientation of the water molecule is not fixed, only the oxygen atom is held in place.

This classification of water molecules helps to judge the flexibility of the water network around proteins. It can also be used to classify observed water molecules and their hydrogen bonding pattern in molecular crystals in general.

Acknowledgement

The author would like to thank Christian Felder for software support, Dr. Weiergräber for providing his lecture notes and transparencies of the 2011 spring school and Prof. Georg Büldt for valuable comments. The author is indebted to Dr. Uwe Müller from BESSY for his timely help on the X-ray beamline chapter. The author is especially grateful to Dr. Andreas Ostermann for supplying some of the Figures and for carefully reading the manuscript.

References

[1] L. Stryer, *Biochemie* (Spektrum Akad. Verlag, Heidelberg Berlin New York, 1991).

[2] M. P. Blakeley *et al.*, Proc. Natl. Acad. Sci. U. S. A. **101**, 16405 (2004).

[3] A. Ostermann *et al.*, Biophys. Chem. **95**, 183 (2002).

[4] O. H. Weiergräber, in *Macromolecular Systems in Soft and Living Matter*, edited by J. K. G. Dohnt *et al.* (Forschungszentrum Jülich GmbH, Institute of Complex Systems, Jülich, 2011).

[5] W. Humphrey, A. Dalke, and K. Schulten, Journal of Molecular Graphics **14**, 33 (1996).

[6] P. Llinas *et al.*, EMBO J. **24**, 1655 (2005).

[7] C. B. Anfinsen, Comparative Biochemistry and Physiology **4**, 229 (1962).

[8] J. Drenth, *Principles of Protein X-Ray Crystallography* (Springer Science+Business Media, LLC, New York, 2007), p. 332.

[9] N. Niimura, and A. Podjarny, *Neutron Protein Crystallography - Hydrogen, Protons, and Hydration in Bio-macromolecules* (Oxford University Press, Oxford, New York, 2011).

[10] G. J. Kleywegt, and T. A. Jones, Structure **3**, 535 (1995).

[11] A. T. Brunger, Nature **355**, 472 (1992).

[12] T. Chatake *et al.*, Proteins-Structure Function and Genetics **50**, 516 (2003).

[13] T. Chatake *et al.*, Journal of Synchrotron Radiation **11**, 72 (2004).

Recommended Textbooks

on proteins in gerneral:

L. Stryer, *Biochemie* (Spektrum Akad. Verlag, Heidelberg Berlin New York, 1991)

on X-ray crystallography:

J. Drenth, *Principles of Protein X-Ray Crystallography* (Springer Science+Business Media, LLC, New York, 2007)

on neutron protein crystallography:

N. Niimura, and A. Podjarny, *Neutron Protein Crystallography - Hydrogen, Protons, and Hydration in Bio-macromolecules* (Oxford University Press, Oxford, New York, 2011)