2 Methods

2.1 Crystallization

The main prerequisite for any crystallographic study is the crystal availability. There are several common techniques to obtain suitable crystals for measurements. For any technique applied, there are two main requirements for the protein being crystallized, purity and homogeneity.

The simplest and the oldest crystallization technique is the batch method. Supersaturated protein solution is sealed and left undisturbed till crystal formation. In the improved batch method, called microbatch, a droplet containing protein and precipitant is inserted into oil, which prevents evaporation of the droplet. If formation of nuclei and crystal growth conditions differ significantly, then seeding is recommended.

The vapour diffusion method is based on vapour equilibration between the protein containing droplet and reservoir, which leads to supersaturation causing nucleation and initial crystal growth. The vapour diffusion might be applied as either sitting or hanging drop method. The disadvantage of vapour diffusion is that protein concentration may change leading to the decreased level of supersaturation.

Dialysis crystallization allows overcoming this problem, as concentration does not change with time when molecules stay in a fixed volume. This method is based on the diffusion of molecules with the low molecular weight through a semipermeable membrane. Another advantage of this method is the possibility to work with volatile reagents and at low ionic strength conditions.

Crystallization of membrane proteins is still a non-trivial task. The main complication is that these proteins are embedded into a membrane *in vivo*, therefore this membrane part must be replaced with some amphipathic reagent, serving two functions: protecting hydrophobic areas of protein and providing sufficient solubility. The ideal candidates to fulfil the given requirements are detergents, which are surface-active molecules that self-associate and bind to hydrophobic surfaces [58]. Predominantly, alkyl-chain containing detergents with either uncharged or zwitterionic head groups with hydrophobic tail containing seven to twelve CH_2 groups that mimick the hydrophobic part of lipids forming membranes, are used to solubilize, purify and crystallize membrane proteins. The choice of

detergent is done mainly empirically, and several different detergents are recommended to try. Indeed, different detergents might yield different crystal forms [59] and even varying the length of detergent's tail might lead to well-diffracting crystals [60].

PSII crystals were grown by the microbatch method as described in [61], which yielded larger crystals in comparison to vapour diffusion, and therefore these crystals are more suitable for X-ray data collection. Bromide-substituted crystals were derived in the same way, except that *T. elongatus* cells were grown in Cl^--free medium containing Br^- instead, as described [62], and in all steps for protein preparation and purification Cl^- was replaced by Br^-, and all buffers used were supplemented with 0.5 M betaine.

2.2 Noble gas pressurization

In recent years there was increasing interest in the application of noble gases (mainly xenon (Xe) and krypton (Kr)) in protein crystallography. Binding of noble gases to protein is completely reversible [63, 64] and does not change the protein spatial structure significantly, therefore opening the way for highly-isomorphous derivatives [65, 66] which can be used for Multi-wavelength Anomalous Diffraction (MAD) or Single-wavelength Anomalous Diffraction (SAD) phasing [67], especially in case of Kr, whose *K* edge (14.3 keV) is easily accessible at all modern synchrotrons. Neon (Ne) and argon (Ar) are not used in protein crystallography due to the small number of electrons and low polarizability (see Table 1), though it was shown that Ar is might be bound to protein in principle [68]. Sites of Xe/Kr binding seem to be unique in comparison to binding sites of more traditional heavy atoms, and the extent of binding might be regulated with changes of pressure [69]. It was shown that Xe prefers highly apolar and hydrophobic regions [64], while the smaller size of Kr allows it to occupy smaller cavities, albeit with weaker interaction with protein, due to lower polarizability of Kr [70].

Moreover, noble gases might be used to explore hydrophobic sites and cavities in proteins [64] and for verification of oxygen channels [71]. Indeed, the physical properties of Xe and Kr make them particularly good analogues of oxygen due to their solubility in hydrophobic environments and van der Waals radii (Table 1) comparable to that of O_2 (Table 1). Xe and/or Kr were successfully used to identify hydrophobic cavities and possible channels for oxygen transport in many different proteins such as porcine pancreatic elastase, subtilisin, cutinase [64, 72], myoglobin, [73], cytochrome *c* oxidase [74] and urate oxidase [64].

Noble gas	N of electrons	K edge (Å)	Polarizability ($Å^3$)	Van der Waals radius (Å)
Neon	10	n.a.	0.40	1.54
Argon	18	3.8710	1.64	1.91
Krypton	36	0.8655	2.48	2.02
Xenon	54	0.3587	4.04	2.16
Oxygen	8	n.a.	0.79	2.13

Table 1. Some physical properties of noble gases. Data for oxygen are given for comparison.

Crystals of dimeric PSII from *T. elongatus* were incubated with Xe at ten, 15, 30 bar and Kr at 40 bar for two to 20 min with a Xenon Chamber (Hampton Research) and immediately flash frozen in liquid nitrogen. Incubation times of five minutes for Xe and eight minutes for Kr were found as most appropriate to achieve sufficient saturation without destroying the crystal by pressure. Approximately 50 crystals were screened for X-ray diffraction.

2.3 Cryocrystallography

Nowadays, one can hardly imagine X-ray data collection from protein crystals without cryostream. Indeed, the wide application of synchrotron radiation for X-ray diffraction experiments at high intense beamlines eliminates the possibility to collect any useful data without cooling of crystals. Going to cryogenic temperatures allows to reduce radiation damage (*vide infra*) significantly (observed back in 1970s by Haas & Rosmann [75]) and to collect more complete data with higher quality. The other advantage is the ability to transport and to store crystals at near liquid nitrogen temperatures.

The only problem is that protein crystals usually have quite high water content (50% and more) [76], thereby transformation of this water into the ice will implicitly destroy a crystal. The solution is to use a cryoprotectant, which replaces the aqueous layer around a crystal. There are two major required characteristics which a cryoprotectant must possess [77]: (i) the ability to vitrify without ice formation and (ii) be harmless to a crystal placed into it. The choice of cryoprotectant might be not an easy issue, however several rules of thumb were published [76-78]. One is to try a cryoprotectant similar to crystal growing conditions or even to introduce a cryoprotectant during crystal grow. Sugars also might be an option as well

as oils [79], which are the easiest to apply (excess water is removed from the crystal surface while in the oil). In case of a typical cryoprotectant, it should be added gradually in several steps to avoid crystal damage.

Basically there are four common techniques to transfer a crystal from mother-liquor to the cryoprotectant [77]. The first is to simply immerge a crystal fished with a loop to the cryoprotectant for a very short time. The main disadvantage is that crystal often falls out from a loop, thereby the crystal spends longer time in the cryoprotectant than planned. Therefore a cryoprotectant must be as similar as possible to the environment of a crystal. The second is called shocked transfer, where a crystal is dropped into the cryoprotectant and left there for some time. This technique might lead to cracking of a crystal. The third is the step transfer, most widely used for sensitive crystals, where a crystal is introduced to the cryoprotectant solution of lower concentration at first, and in several steps to the solution of desired concentration. The last one is dialysis and applied in the worst cases, when nothing else works. The experimental setup resembles one used in sitting or hanging drop crystallization technique.

Though the ultimate goal of cryoprotectant addition is to provide water vitrification without ice formation, it is not uncommon to observe ice rings in the X-ray diffraction pattern. There are three options one can try: (i) to exclude ice rings during data processing (*vide infra*) with software, but this is not the best option, as useful reflections are also lost; (ii) to try cryoprotectant optimization; or (iii) to try annealing – removal of a crystal from the cryostream with consequent recooling. This is usually done with a plastic card, with which cryostream is blocked. Nowadays, at some beamlines (*e.g.* PX 14.2 at BESSY) special cap is installed, which is controlled via software. The user defines the annealing time and then cryostream is automatically blocked for a given time to heat the crystal.

PEG2000 (polyethylene glycol) showed sufficient cryoprotection for PSII crystals and was routinely used in all data collections.

2.4 Synchrotron radiation

X-Rays are produced when a beam of high-energy electrons, accelerated in the vacuum, hits a target: electrons from the inner shells of metal atoms are knocked out, then vacancies are filled with the electrons from the higher orbitals and the energy excess is emitted as X-ray photons. The simplest X-Ray source is the sealed X-Ray tube, but the produced beam has not sufficient intensity for protein crystallography. Introduction of

rotating anodes solved this problem, but new and challenging projects, especially in the area of membrane protein research, demanded even higher intensities. The solution is synchrotron, where charged particles (electrons or positrons) of high energy are accelerated and kept in circular motion by magnetic fields in order to produce synchrotron radiation. Synchrotron radiation is widely used nowadays in protein crystallography [80] as it has many advantages in comparison to home-source X-Ray generators. The main benefits are: time; it is possible to collect a full dataset at a synchrotron at the range of minutes to hours instead of days; changeable wavelength, in the range of 0.7-2.5 Å, welcoming anomalous diffraction experiments; high intensity; high brilliance; high collimation.

All measurements on PSII crystals were done at several European synchrotrons: BESSY (Berlin, Germany), ESRF (Grenoble, France) and SLS (Villigen, Switzerland).

2.4.1 Radiation damage

Albeit cryocrystallography significantly reduces the problem of radiation damage, there is an evidence that damage occurs even in cryocooled crystals [81], especially at modern synchrotron sources with highly intense beam. Radiation damage leads to serious deterioration of the quality of the collected data [82] due to decreased diffraction power, increased unit-cell volume, changes of oxidation states of metals if present, increase of B-factors, disruption of disulfide bridges, decarboxylation of glutamates and aspartates, loss of hydroxyl groups in tyrosines and cleavage of sulfides in methionines [83-85].

Radiation damage might be divided into two subclasses: primary and secondary. Primary damage is the result of ionization of atoms by the X-ray photons, and can not be avoided, whereas secondary radiation damage is caused by generated secondary electrons, inducing further excitation and ionization events within the crystal [86]. The main destructors are the free radicals [87] – species carrying one or more unpaired electrons. They are short-lived and tend to react quickly, either colliding with the other reactive species, thereby terminating radical propagation or with neutral species, which become free radicals in turn.

During data collection, incident X-ray photons interact not only with protein molecules, but also with solvent (mostly water). Such kind of interaction produce a variety of radicals [88]:

$$H_2O \xrightarrow{\text{Ionizing radiation}} H_2O^{+\bullet} + e^- \text{ (ionization)}$$

$$H_2O \xrightarrow{\text{Ionizing radiation}} H_2O^{\bullet} \text{ (electronic ionization)}$$

$$H_2O^{+\bullet} + H_2O \rightarrow H_3O^+ + {}^{\bullet}OH$$

$$e^- + nH_2O \rightarrow e^-_{aq}$$

$$H_2O^{\bullet} \rightarrow H^{\bullet} + {}^{\bullet}OH$$

$$e^-_{aq} + H^+ \rightarrow H^{\bullet}$$

It should be noted that the hydroxyl radical is extremely labile and the generated electrons are mobile at 100K, with electrons tunnelling even at 5K [89].

One technique to reduce secondary radiation damage is the addition of radioprotectants or radical scavengers [88]. The latter react with externally generated radicals, thereby protecting protein from secondary damage, while radioprotectants either react with secondary radicals or repair already damaged residues [86]. There are two general mechanisms of protection [88]: (i) a radioprotectant donates electrons and the radical reaction is terminated or (ii) a radioprotectant catches mobile electrons and forms relatively stable products.

Among proved radioprotectants / radical scavengers are nicotinic and nitrobenzoic acids [90], ascorbate [88, 91] and quinone [88].

PSII is very prone to radiation damage, especially in the area of the unique Mn_4Ca cluster, which undergoes significant reduction of oxidation states under the influence of radicals [92]. Therefore the addition of radical scavengers is desirable for data collection with PSII crystals.

2.5 Data collection

In the case of poorly diffracting and radiation sensitive crystals, the optimized strategy for data collection is a must. The ultimate goal is a complete dataset with high redundancy, low R-factors and highest possible resolution. Although this is a rare case, some steps might be taken to improve the quality of collectable data.

It is widely accepted that the most important parameter of a collected dataset is its completeness. Data from a diffracting crystal are sets of Miller indices plus associated intensities with their standard uncertainties [93]. The completeness of indices strongly depends on crystal lattice geometry and setup, whereas the completeness of intensities is

influenced by exposure time, numerous characteristics of X-Ray source and detector and others.

The typical way to describe interactions between a crystal and X-rays is to draw the Ewald construction. It is drawn as sphere with radius 1/λ (λ - wavelength of X-rays), centred on the beam of X-rays, intersecting the reciprocal lattice of the crystal. In case any reciprocal-lattice point lies on the surface of the Ewald sphere, Bragg law (λ=2dsinθ) is fulfilled and diffraction is observed.

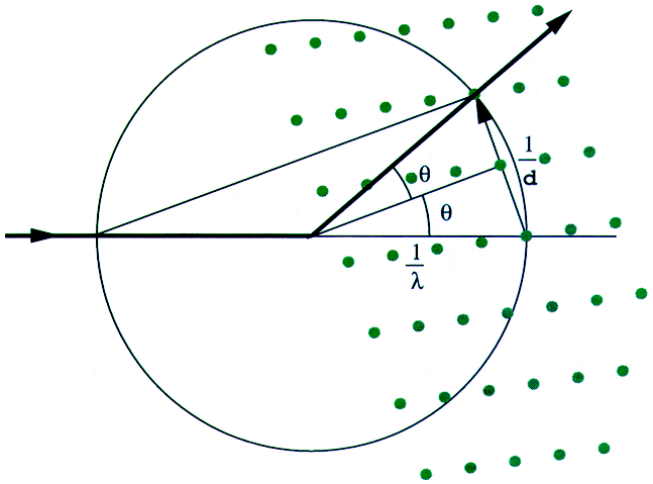

Figure 12. The Ewald construction. λ is the wavelength of X-rays, *d* is the interplanar distance, θ is the scattering angle. Adapted from [93].

It is obvious that in order to collect data containing many reflections (reciprocal points lying on the Ewald sphere), either the position of spots or the radius of sphere must be changed [93]. In the first approach termed rotation method, the radius of the Ewald sphere (wavelength of X-rays) is constant and the crystal is rotated (usually along one axis), whereas in the Laue method, the crystal is not rotated and white radiation (continuous wavelength spectrum) is used.

The highly recommended strategy with the rotation method is to collect initially two images separated by 90 degrees and inspect them by eye as well as with software.

Visually one can determine the resolution limit and, based on this observation, adjust the crystal-to-detector distance. But one should keep in mind that modern software might take into account even weak reflections, barely visible by eye, so it is recommended to position the detector closer, but only slightly, as the signal-to-noise ratio is better with longer distances. The reflection profiles (diffraction spots) must be thoroughly examined; they should have defined shape, regular nature and one peak per spot (no satellites which are signs

for crystal splitting into several parts). Anisotropy of diffraction and mosaicity of crystal might also be estimated. Exposure time might also be adjusted after evaluation of spots intensities. If the spots are weak, longer exposure times might be recommended, but experimentator must keep in mind that radiation damage will also increase, therefore the run for higher resolution by an increase of exposure time should be avoided. The rule of thumb is that it is better to collect complete dataset with somewhat lower intensities than only a partial dataset with higher intensities (leading to better quality in theory). Moreover, longer exposure time might lead to an increased number of overloads in the image.

When the user is satisfied with the results, the next step is to determine the proper rotation angle and total rotation range. There are two approaches how to collect data: wide-slicing, when images are collected with a rotation angle of 0.5° or wider, resulting in decreased amount of partial reflections; and fine-slicing, when an angle of 0.1° or even less is used, resulting in the decline of fully recorded reflections and a spread of each reflection over several images. The advantage of fine slicing is that it provides more accurate intensity integration by construction of three-dimensional profiles and minimizes background. The main limitation of fine slicing is that it cannot be used with detectors with significant dead-time (imaging plates) and results in increased overall data collection time (the crystal is exposed for longer time and therefore the radiation damage might be worse). In wide slicing, there is a limitation for the maximal angle of rotation (rotation range) as overlapping of reflections must be avoided. The maximum permitted rotation angle $\Delta\phi$ can be calculated as follows [93]:

$$\Delta\varphi = 180d/\pi a - \eta,$$

Where η is the angular width of a reflection (due to mosaicity and beam divergence), a is the length of the primitive unit-cell dimension along the direction of the X-ray beam, d is the high-resolution limit and $180/\pi$ is used to convert radians to degrees. The value of the rotation angle is of great importance especially for crystals with large unit cell (*e.g.* PSII crystals). For such type of crystals the alignment of the longest axis parallel to the spindle axis is recommended to avoid overlapping of reflections.

The last step is to calculate the total rotation range to collect a complete dataset. The data are complete if all reflections (or their symmetry mates) in the asymmetric part of the reciprocal space have crossed the Ewald sphere. The optimal strategy to collect a complete

dataset within minimal time is usually evaluated by software, but several simple considerations might be taken into account for initial evaluation. Collecting 180° (or 360° for anomalous data) will definitely give the maximum completeness, but knowing the symmetry of the crystal it is possible to minimize data collection time accordingly (see Table 2).

Point group	Rotation range, °
1	180
2	90
222*	90
4	90
422	45
3	60
32	30
6	60
622	30
23	60
432	35

Table 2. Minimum rotation range to collect a complete dataset for given point group. * PSII crystals belong to the 222 point group.

All PSII X-ray datasets were collected with the wide slicing technique with 0.5° rotation, the total rotation range was estimated with XPLAN procedure in XDS [94]. The exposure time was varied according to the intensity of the reflections, typical range was 5-30 seconds per image depending on the intensity of the beamline.

If a crystal decays rapidly during X-ray data collection, another part of the same crystal is used to collect a partial dataset, and all partial datasets are merged and scaled on the first one at the final stage.

2.6 Data processing

After data have been collected, they must be processed. A typical dataset usually contains more than one hundred images, collected from different crystal orientations. Data

processing consists of several steps: indexing of diffraction spots, refinement of parameters of the detector, integration of diffraction spots, scaling, precise refinement of unit cell parameters and merging of measurements. There are several computer programs, which significantly alleviate this problem and allow processing of data collected literally with any detector. The most widely used are MOSFLM [95], HKL-2000 [96] and XDS [94]. MOSFLM and HKL-2000 have an intuitive and friendly user-interface, but the latter is license-paid software. XDS has no graphical user interface, but is easy to use after some training – all commands are user-defined in the input file, which is fed to the program. The program produces several comprehensive tables, and after analysis the user makes a further decision. XDS runs several subroutine procedures:

XYCORR calculates tables of spatial corrections for each detector pixel, which are used in subsequent steps;

INIT creates three tables, used later to classify pixels in the data images as background or belonging to a diffraction spot;

COLSPOT determines strong diffraction spots in the subset of the data images and stores them in special file;

IDXREF takes experiment parameters defined by the user (or automatically written out at some beamlines) and strong diffraction spots found by the COLSPOT procedure to find the orientation, metric, and symmetry of the crystal lattice. It uses the refined metric parameters of the reduced cell (primitive triclinic) to test each of the 44 possible lattice types and reports the likelihood of being correct and the conventional cell parameters for each lattice type;

DEFPIX determines shadowed pixels (*e.g.* by hardware) and marks them as untrusted to exclude from processing;

XPLAN calculates the best strategy for data collection. It estimates the completeness of new reflection data, expected to be collected for each given starting angle;

INTEGRATE determines the intensity of each reflection predicted to occur in the rotation data images. The integration is done in two passes. In the first pass, spot templates are generated by superimposing profiles of fully recorded strong reflections, and all grid points with a defined value are taken as elements of the integration domain. To allow for variations of their shape, profile templates are generated from reflections located at nine regions of equal size covering the detector surface and additional sets of nine to cover equally-sized batches of images. The actual integration is carried out in the second pass by profile fitting with respect to the spot shape determined in the first pass;

CORRECT applies correction factors to intensities and standard deviations of all reflections after integration, determines the space group if unknown and refines the unit cell constants, reports the quality and completeness of the data set, and saves the final integrated intensities. Data quality as a function of resolution is described by the agreement of intensities of symmetry-related reflections and quantified by the R-factors: R_{sym}, and the more robust indicator, R_{meas} [97]. These R-factors as well as the intensities of all reflections with indices of type h 0 0, 0 k 0, and 0 0 l and those expected to be systematically absent provide important information for identification of the correct space group. Clearly, large R-factors or many rejected reflections (MISFITS) or large observed intensities for reflections expected to be systematically absent suggest that the assumed space-group or the indexing is incorrect. Finally this procedure analyzes the distribution of reflection intensities as a function of their resolution and reports outliers from the Wilson plot. Often these outliers arise from ice rings in the data images, which might be removed by request.

The next stage is scaling, the process of rescaling all measured intensities to the one common scale. The program XSCALE in XDS package serves for this purpose, which can also scale different datasets (useful for scaling MAD datasets collected from one crystal) and can correct reflections individually for radiation damage [98]. The output file might be converted into the conventional mtz file with either XDS subprogram XDSCONV or with program POINTLESS included in the CCP4 Package [99].

2.7 Structure solution

In the X-Ray data analysis the structure solution usually means solving the phase problem, which is described by the fact that only intensities but not the phases (essential for calculation of electron density) can be measured in X-ray diffraction experiments.

There are several phasing techniques; a short description is given below.

2.7.1 Isomorphorous Replacement

The oldest method introduced in 1956 by Perutz and Kendrew to determine phases is based on the idea of introducing heavy atoms that serve as markers, into the crystal without disturbing its order as the main requirement is isomorphism between the native crystal and its heavy-atom derivatives. The usual technique is to soak a crystal in a solution containing heavy-atoms. The typical candidates are quicksilver, platinum, gold and lead.

In single isomorphous replacement (SIR), the contribution of a heavy-atom is calculated as isomorphous difference between amplitudes of reflections measured from a native crystal and derivative crystal:

$$|F_H| = |F_{PH}| - |F_P|$$

The obtained difference is used to obtain approximate positions of heavy-atoms, which in turn allows calculating of initial phases α_H, used for the calculation of native phases α_P. The problem is that these phases are ambiguous and proper estimation of phase probabilities is needed. The phase probability distribution is described by the Hendrickson-Lattman coefficients [100] HLA, HLB, HLC, HLD, which are used to calculate initial maps based on the best-estimated phase:

$$P(\alpha) = N\exp(A\cos(\alpha) + B\sin(\alpha) + C\cos(2\alpha) + D\sin(2\alpha)),$$

where α is the true phase; A, B, C, D are Hendrickson-Lattman coefficients; N is the normalization constant.

In multiple isomorphous replacement (MIR) the phase ambiguity might be overcome. Here, the phase probabilities are derived by multiplying individual phase probabilities [101].

2.7.2 Anomalous diffraction

Serious limitations of isomorphous replacement were overcome with the introduction of Multi-wavelength Anomalous Diffraction (MAD), as this method provides exact isomorphism (data are collected from the same crystal). In MAD, the anomalous signal is collected from anomalous scatterers, which might be either intrinsic (metal cations in metalloproteins) or exogenous (*e.g.* selenomethionines or heavy-atom derivatives) [102, 103]. A heavy atom absorbs X-rays of specified wavelength. In the result of such absorption, Friedel's law is broken and opposite (F^+ and F^-) reflections are no longer identical. This phenomenon is called anomalous scattering (anomalous dispersion) and allows extraction of phase information. Usually a MAD experiment is carried out at three different wavelengths: at minimum (f′) and maximum (f″) at the absorption edge of a given anomalous scatterer and additionally at a remote wavelength, where anomalous scattering is neglected.

Modern hardware and software allow resolving of the phase problem even with data collection only at the absorption peak – single-wavelength anomalous diffraction (SAD) [104]. SAD is also used to determine positions of heavy atoms in proteins.

Application of SAD allowed determination of bromide-binding position in PSII and also positions of noble gas molecules (see section 3 for detailed discussion).

2.7.3 Molecular replacement

Molecular replacement (MR) is a powerful phasing technique, which might be used if the studied protein has noticeable homology to already solved structures. It is assumed that proteins with high homology should show similar tertiary structure; therefore phases from a known structure might be used to determine the phases of a studied protein. But to extract phases, the known structure must be superimposed with the studied molecule in its unit cell. The idea is that trial positions and orientations of the known model in the unit cell of the studied protein must be found and structure factors (F_{calc}) calculated and compared with measured (observed) structure factors (F_{obs}). The best match between F_{calc} and F_{obs} gives the most appropriate position and orientation of the molecule and phases α_{calc} might be calculated and taken as initial phases for structure determination. In reality such search might be very time-consuming, especially for large unit cells, therefore it is usually done in two major steps: a search of the best orientation and a search of the best position. Independent search for best orientation is done by application of the Patterson function (in this particular case it is termed rotation function [105]), which is used to reconstruct two three-dimensional Patterson maps of search and target models. The correlation between these two maps is monitored, and the best value is taken as a candidate for the next step. The obtained solution is used for search of the best position (translation function). Quality of solution is monitored with correlation coefficients and R-factors.

As a rule of thumb, there is no need to perform the search with all collected data, a range between 3 and 5 Å resolution is generally used. Indeed, low-resolution data contain significant contribution from the solvent, whereas high-resolution data might be too sensitive for a model search.

MR is extremely useful in studies of conformational changes in proteins, *e.g.* by ligand binding, as both native and derivative structures are almost identical and phases from the native structure might be easily applied for the same protein that has been co-crystallized with a ligand. With rapidly increasing amount of deposited structures in the Protein Data

Bank [106] this method might be recommended as a first choice if the sequence of a studied protein shows some homology (currently it is possible to solve structure even with 20-25% of homology, sometimes with additional truncation of the model into separate domains and using poly-alanine models or using a combination with other phase information) as the experimentator needs only to collect the dataset from the native crystal.

There are several stand-alone MR packages: AMORE [107], MOLREP [108], PHASER [109] (all also included in standard CCP4 package [99]) as well as pipelines BALBES [110], MrBump [111] (combining search for the best model with actual MR) available for scientific community.

The structural model of PSII described in this work was obtained by MR with a structure of PSII at 3.0 Å as search model with PHASER software [109].

2.8 Model Building and refinement

After initial phases are derived and first maps are calculated, the phases must be improved to obtain interpretable maps to construct a model of studied protein. The most common procedure employed for phase improvement is the so-called density modification (or solvent flattening). At this stage, the electron density in the unit cell is divided into two regions: with high density and assigned to protein and with lower density assigned to solvent. In this way, the contrast is enhanced and some random fluctuations in solvent region are diminished. The other way to improve electron density maps is to apply noncrystallographic symmetry (NCS) if present. Such symmetry averaging leads to improved signal-to-noise ratio and therefore better maps are obtained. At the point, when a map is interpretable, the protein chain must be traced. In case of high resolution (better than 1.5 Å), algorithms for automatic building might be employed [112], but in the area of membrane crystallography, where typical resolution is 2.5-3.5 Å, this option is excluded. Depending on resolution and actual size of protein, building of model and interpretation of electron density might take considerable amount of time.

After each round of building, the model must be refined to include new phase information and to monitor reasonableness of the constructed model in terms of chemistry (bond lengths, bond angles, torsion angels, planar groups, etc) and physics (absence of atomic overlaps, presence of energetically-favored interactions, etc) [113].

By definition, the crystallographic refinement is the optimization of a function of a set of observations by changing the parameters of a model [114].

Several algorithms have been adjusted for refinement of macromolecules: from typical least-square minimization to the sophisticated maximum likelihood, based on the Bayesian method.

In the least-square refinement [115], the target function is minimized in terms of minimizing the squares of differences between expected structural factors ($|F_{calc}|$) and observed ones ($|F_{obs}|$):

$$E = \sum_{hkl} \omega_{hkl} (|F_O| - |F_C|)^2_{hkl}$$

where E is the target function and ϖ is a weighting factor.

The most serious limitation of this algorithm is that the least-square solution might be captured in a local minimum, i.e. the given procedure will find the nearest minimum of a function from the starting point; therefore parameters of the starting model must be near the global minimum to provide a correct solution. The depth of search is defined by the radius of convergence.

This limitation is overcome in the simulated annealing method [116], implemented in the program complex CNS [117]. Model is "heated" to give it more freedom and to escape form the local minimum (as the radius of convergence is increased and energetical barriers are not forbidding) and then allowed to "cool down" to find the lowest energy state in agreement with diffraction data.

In Bayesian methods [118] rather ensembles of models are refined to avoid incidental loss of solution at any stage of refinement, which is possible with other methods, where only best model is considered. For each model in the ensemble prior probability is assigned, which is later compared with the probability distribution of the observed amplitudes, and then the likelihood of each model is determined. Later the posterior probability is calculated, and the best models are used to generate new ensembles of models, but of higher likelihood. Here, the target of refinement is the maximization of either likelihood or posterior probability.

The set of parameters that are used to minimize target function includes the so-called B-factor (which describes thermal motions of atom in the lattice) and site occupancy in addition to atom positions (x, y, z). To perform a reliable and stable refinement, the amount of observations must be greater than the number of parameters. The typical way to increase this ratio is the introduction of constraints (parameters are fixed during refinement) and the addition of restraints (so-called prior knowledge, *e.g.* standard values for bond angles and

bond lengths taken from high-resolution small-molecule structures [119] which serve as the target values in the refinement). Constraints reduce the number of parameters, whereas restrains on the contrary increase the number of observations. Occupancy in the most of cases is treated as constraint, except for water and ligand molecules, which might have partial occupancy. In contrast, the refinement of B-factors decreases the ratio between observations and parameters, because one parameter is needed if a B-factor is considered as isotropic (spherical) and six parameters are required for anisotropic (ellipsoid) B-factors. Therefore, only high-resolution data might be used for the refinement of anisotropic B-factors. But it is possible to overcome this physical problem with TLS (translation, libration, screw) parameterization [120], which describes movements of groups of atoms as being a rigid unit, thereby greatly decreasing number of parameters. One TLS group is described by 20 parameters (six values of the translation tensor, six values of the libration tensor and eight values of the screw rotetion tensor), but one such group might be assigned to a domain, for example.

The success of a refinement is monitored by so-called R-factor, describing how well the amplitudes ($|F_{calc}|$) calculated from the current model correlate with the experimental amplitudes ($|F_{obs}|$):

$$R = \frac{\sum \| F_{obs} | - | F_{calc} \|}{\sum | F_{obs} |}$$

In addition, R_{free}-factor [121] is used for cross-validation and calculated based on a small fraction of data (usually ~5%) which is not used in the refinement and thus unbiased by the model.

All PSII models were refined by simulating annealing with maximum likelihood target using the CNS program [117]. In order to improve the observations-to-parameters ratio, NCS was applied and B-factors were treated as isotropic. All model building procedures were performed with COOT software [122].

2.9 Structure Validation

After each step of refinement the structure must be validated in order to check the presence of errors in the model. The most frequently occurring errors during manual building

are related to the conformation of the polypeptide chain, which must comply with a common set of rules describing steric collisions between main and side chains. Such errors might be easily detected with a Ramachandran plot [123], in which torsion angles ϕ (the angle of rotation around C_α-N bond) and ψ (the rotation angle around C_α-C bond) (Fig. 13) are plotted against each other for all residues. There are common zones of the plot, which correspond to the left-handed α_L- and the right-handed α_R-helices as well as to β-strands. All residues should ideally be in favourable, allowed or generally allowed regions of the plot, although some outliers might occur at strained section of the model.

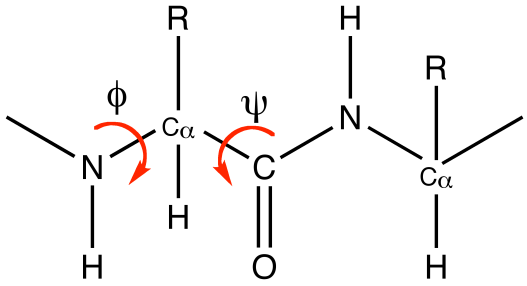

Figure 13. The backbone conformational angles in proteins (see text for details).

The overall geometry (bond distances and bond angles) is usually checked by comparison with standard values obtained from small molecules [124].

Validation packages, such as WHAT IF [125] and PROCHECK [126] allow to highlight not only issues described above, but also perform a thorough check of van der Waals contacts, planarity of aromatic groups, chirality, nomenclature, torsion angles and other parameters.

The derived model of PSII underwent validation in COOT [122] at every step of model building as well as thorough check with PROCHECK [126] before atomic coordinates were deposited in the protein data bank with codes 3BZ1 and 3BZ2 (each monomer of the PSII dimer in a separate file).

2.10 Theoretical calculations of channels

It is obvious that all proteins are not rigid-body solid objects and have a rather complex shape determined by pockets, clefts, cavities, channels and protrusions. In many proteins catalytic sites are buried deeply and connected by dedicated channels / tunnels with bulk solvent. Therefore exploration of such transport pathways is indeed an important point in

revealing catalytic mechanisms. These channels might play a role not only in the delivery to, but also could be responsible for selectivity of substrate from the active site [127].

In PSII, water is the substrate of the unique Mn_4Ca cluster and the side product (O_2) is generated in the reaction of water photolysis:

$$2H_2O \rightarrow O_2 + 4H^+ + 4e^-,$$

therefore calculation of transport pathways to and from the Mn_4Ca is of great interest.

The recently developed program CAVER [128] allows fine, accurate and reliable calculation of channels in proteins of any size. It has a robust search algorithm that calculates the geometrically most favourable pathways. The starting point is defined by the user (usually it is the binding site) and the program constructs virtual objects on a discrete grid space. The atoms of the protein are treated as rigid spheres of appropriate van der Waals radii, and grid nodes, which coincide with protein spheres, are excluded from further calculations. In such a manner the shortest low-cost path is calculated by application of the Dijkstra algorithm [128]. The output might be visualized with Pymol software [129], using the maximum radii for each node of the path; the smallest radii represents the channel gorge.

The below-described model of PSII at 2.9 Å resolution including all cofactors except for Mn_4Ca (which was used as the starting point for calculations of channels) and all other ions (non-haem Fe^{2+}, Ca^{2+}, Cl^-) (due to limitations of the software) was subjected to extensive calculations with CAVER with a grid size (the distance between two nodes) of 0.4 Å (very fine grid). Additionally, manual placement of putative water molecules in channels was performed with respect to all geometric parameters (hydrogen bond lengths, angles between donor and acceptor) using program COOT [122]. The estimated maximal coordinate errors were calculated by program SFCHECK [130].