6 APPENDIX

6.1 Theory of X-ray crystallography

A crystal lattice can be described as a series of parallel lattice planes defined by the Miller indices h, k, l. When an incident X-ray beam hits a lattice plane under the angle θ and the Braggs condition 2 d sinθ = nλ is satisfied, it will be diffracted and the diffracted beam will have the same wavelength as the incident beam. The diffracted beam is scattered by all electrons in the crystal. Therefore the diffraction data contain the information on the location of the electrons in the unit cell. The electron density (?) at a certain position in the unit cell (x, y, z) is a function of the amplitude (|F|) and the phase (α) of all diffracted beams:

$$r\left(x, y, z\right)=\frac{1}{V}\sum_{hkl}\left|F_{hkl}\right|\cdot e^{ia_{hkl}}\cdot e^{-2p\cdot i\cdot\left(hx+ky+lz\right)}$$

The amplitudes or structure factors |F| are proportional to the square root of the intensities of the reflections and can thus be determined experimentally. However, the phase information containing the major structural information is lost during data collection on a detector. This is so called “phase problem” of crystallography. Several methods are employed to determine the lost phases:

1. Isomorphous replacement
2. Anomalous scattering
3. Molecular replacement
4. Radiation induced damage
4. Direct methods

1) Isomorphous replacement is based on the observation that a heavy atom, having a much higher concentration of electrons in a small volume, will scatter X-rays more strongly than C, N, O, S, H toms of a protein. If a heavy atom can be introduced into a protein crystal without otherwise changing the protein or the crystal lattice (i.e. the crystal remains isomorphous), the differences in the scattering intensities from this

crystal to the native crystal can be exploited to indirectly determine the missing phase angles.

The most powerful method in determining the heavy-atom coordinates is the Patterson function P (u, v, w), which is a Fourier sum without phases:

$$P(u,v,w)=\frac{1}{V}\sum_{h}\sum_{k}\sum_{l}\left|F_{hkl}\right|^{2}e^{-2p\cdot i\cdot(hu+kv+lw)}$$

The coordinates (u, v, w) display peaks at locations corresponding to vectors between atoms, whereas strong peaks in the Patterson map represent vectors between heavy atoms. As usually only few heavy atoms bind to the protein, the substructure is simple and the position of heavy atom peaks can be recognized and their coordinates determined. Additionally, unit-cell symmetry can simplify the search of the peaks, because heavy atoms bind in equivalent positions that are located on Harker sections.

2) Anomalous scattering for obtaining phases is based on the fact that heavy atoms show anomalous dispersion when irradiated by X-ray with wavelength near their absorption edge. This causes differences in the scattering intensities of Friedel pairs, which are used to find the heavy atom positions and to determine phase angles of each reflection. By MAD method (MAD = mulivawelength anomalous dispersion) one can maximize the differences between Bijvoet reflection pairs using different wavelengths.

3) Molecular replacement is based on the method where the structure factors of the target model are computed with measured intensities and with phases estimates derived from the atomic model of a search model. This search model is the structure of a protein that is sequentially homologuous to the target structure and is extracted from the Protein Data Bank (PDB). In the first step, the search model is correctly oriented (using Patterson function), and after the rotation search has been successful as shown by correlation function, the model is shifted along x, y, z coordinates through the crystal asymmetric unit (translation search). R-factors are calculated at each position and indicate the correct one.

After the search model is correctly oriented and positioned the structure factors are calculated as a sum of scattering contributions of all atoms in the unit cell:

$$F_{calc}(hkl) = \sum f_i e^{2pi\left(hx+ky+lz\right)}$$

Phase angles α_{calc} than can be calculated from the equation

$$F_{calc}(hkl) = \left|F_{hkl}\right| \cdot e^{ia_{calc}}$$

using experimental amplitudes $|F_{hkl}|$ and calculated structure factors $F_{calc}(hkl)$ and an electronic map is calculated, in which atom coordinates and B-factors of the target structure can than be refined.

4) Radiation and UV-light lead to the decarboxylation of aspartate and glutamate residues and to the break of disulfide bridges. This in turn leads to changes in reflexion intensities which can be employed phase determination similar to that using isomorphous replacement.

5) The fifth method is only feasible for structures with fewer than 1000 atoms and is a preferred method for phasing reflections from small molecules. Using direct methods, the phases are estimated from the corresponding experimental amplitudes, exploiting constraints or statistical correlations between the phases that result from the fact that the scattering density must be a positive real number. The last two methods require data with precisely estimated intensities at atomic resolutions.

6.2 Crystal mounting in a capillary tube

For X-ray scattering experiments at room temperature, crystals are transferred into quartz capillaries. The crystal is placed into the capillary filled with mother liquor and dislodged from the solution using paper tips. Excess of mother liquor around the crystal is removed using filter paper wicks and thin paper tips. The capillary is sealed with wax from both sides and mounted on the goniometer head using modeling clay.

6.3 Crystal cryo-cooling

Crystal cryocooling is usually employed to reduce radiation damage during X-ray crystallography and to reduce thermal motion of atoms, yielding lower temperature factors. A disadvantage of the method is ice formation, which disrupts the crystal lattice, since protein crystals usually have a solvent content between 40 and 60%. In practice, the ice nucleation is prevented by using cryoprotectants like glycerol, 2-methyl-2,4-pentanediol (MPD), ethylene glycol, PEG 400, oils, sugars, concentrated salt solutions and others. For flash freezing, crystals are incubated in the original mother liquor containing appropriate cryoprotectant concentration (usually 20-30%), harvested with cryo-loops and frozen by plunging the cryo-loops harbouring crystals into liquid nitrogen.

6.4 X-ray data collection

All X-ray diffraction data can be collected using oscillation method, where the crystal is rotated in the X-ray beam in steps of 0.2°-2°. The size of the oscillation step is dependent on mosaic spread, resolution, crystal unit-cell constants as well as on the divergence of the X-ray beam. After crystal symmetry and crystal orientation has been determined, the optimum oscillation angle and the rotation ranges needed to collect the dataset with the highest possible completeness in the shortest time are calculated with the MOSFLM program.

6.5 X-ray data processing

The processing of crystal diffraction data consists of several steps, which include the indexing of the diffraction pattern, refinement of crystal and detector parameters, integration of diffraction maxima, precise refinement of crystal parameters using the whole dataset, and merging and statistical analysis of the measurements related by space-group symmetry. The aim of the data processing is to reconstruct the three-dimensional image of the reciprocal space precisely from the series of two-dimensional diffraction images. The theory behind the data reduction methods is very complex and several algorithms were developed and implemented into computer software. The most widely used programs for indexing and scaling are XDS, HKL and MOSFLM.

6.6 Circular dichroism spectroscopic measurement

Circular dichroism (CD) spectroscopy is a form of light absorption spectroscopy and measures the difference in absorbance of right-circularly (ε_R) and left-circularly (ε_L) polarized light by optically active matter.

$$\Delta e = e_L - e_R$$

In a CD experiment, equal portions of left and right circularly polarized light are radiated into a protein solution. One of the two types is absorbed more than the other one, and this wavelength dependent difference of absorption is measured, yielding the CD spectrum of the probe. CD spectroscopy in the "far-UV" spectral region (190-250 nm) is sensitive to secondary structure of polypeptides and proteins containing chiral C_a atoms bound to the amide chromophore. For determination of the secondary structure content determination the recorded spectrum is deconvoluted into the individual contributions of a-helix, ß-sheet and random coil using computer programs.

The difference in absorption to be measured is very small and is usually a few 1/100ths to a few 1/10th of a percent, but it can be determined quite accurately. The raw data plotted on the chart recorder represent the ellipticity θ of the sample in millidegrees which is related to Δε by equation:

$$q = \Delta e \cdot \frac{\ln 10 \cdot 180°}{4p}$$

To be able to compare ellipticities, the values of ? are converted into normalized values (mean residue ellipticity [θ]), which is the most commonly reported unit and is measured in [degrees cm^2 dmol^{-1} residue^{-1}]:

$$[q] = 0.0001 \cdot \frac{q \cdot M_r}{c \cdot d \cdot n}$$

where θ specifies the ellipticity in millidegrees, Mr - molecular weight in g/mol, c – protein concentration in g/ml, d – path length in cm and n – number of residues. The factor 0.0001 is required to convert millidegrees to degrees and mol in dmol.

The analysis of CD spectra can yield valuable information about the secondary structure of biological macromolecules. CD is a non-destructive method and is used for verifying that the protein is in its native conformation. Furthermore, it is a valuable tool for studying the secondary structure changes upon the binding of ligands or as function of temperature, concentration of denaturing agents, pH, and others.

6.7 SAXS-Methods

Small-angle X-ray scattering (SAXS) is a useful method for studying the structure of macromolecules in solution at low resolution. In a SAXS experiment a monodisperse dilute macromolecular solution is exposed to X-rays and the scattered intensity, I(s), is recorded as a function of the momentum transfer s ($s = 4\pi\sin\theta/\lambda$, where 2θ is the angle between the incident and scattered radiation). For monodisperse solutions, the intensity after subtraction of the separately measured solvent scattering is proportional to the scattering from a single particle averaged over all orientations. In an ideal diluted monodisperse solution all particles scatter independently, and the measured intensity magnitude is merely the sum of the intensities from each individual molecule depending only on scattering angle θ between the incident and scattered beam.

The SAXS patterns directly provide parameters such as molecular mass (MM), radius of gyration (R_g), hydrated volume (V_p) and maximum particle diameter (D_{max}). The forward scattering intensity I(0) is related to its molecular mass, and the radius of gyration R_g characterizes the particle size. The size of the particle can be determined by calculating whether the intensity at zero angle ($s = 0$) from the plot ln I(s) versus s^2 (Guinier plot), which should be a linear function for monodisperse solutions. The intercept of the Guinier plot gives I(0) and the slope yields the radius of gyration.

$$I(s) \cong I(0)\exp\cdot(-\frac{1}{3}R_g^2 s^2)$$

The small angle scattering curves are rapidly decaying functions of momentum transfer (s) and are essentially determined by the particle shape. SAXS therefore contains information about gross structural features - shape, quaternary and tertiary structure. SAXS can be effectively combined with other structural, computational and biochemical methods. The most important approach for macromolecular complexes is rigid body modeling, when the structures of the individual subunits are available, and SAXS is employed to obtain the structural information about the entire complex.

As the useful part of the scattering curve contains structural information up to resolution of ~0.5 nm (Svergun 2001) and the protein structure typically consists of folded polypeptide chain composed of amino acids separated by ~0.38 nm between adjacent C_α atoms, the protein structure can be reconstructed by dummy residues (DR) represented by spherical volumes centered at $C\alpha$. The small numbers of required parameters renders possible methods for trial-and-error modeling of the three-dimensional shape of the molecule under study by *ab initio* methods. The initial model is randomly modified by simulated annealing and using Monte-Carlo-like (GASBOR, DAMMIN) or genetic-algorithm (DALAI_GA) searches to find spatial arrangements of the DRs that fit the experimental data. Some programs like GASBOR use for placing of DR the restriction that each dummy residue has two neighbors at a distance of $\approx$0.38 nm (average distance between two C_α atoms in proteins). This compactness criterion together with the limited number of DRs (equivalent to the expected number of amino acid in the asymmetric unit) allows the DR model to be built very effectively.

6.8 Homology protein structure modeling and analysis

Because of the evolutionary relationship, the domains in protein sequence that have gradually evolved can be clustered into a relatively small number of families of domains with similar sequences and structures (folds). The knowledge of the three-dimensional structure of more than 40 000 proteins makes it possible to use threading and structure modeling to predict the unknown structures of proteins based on their similarity to known protein structures. The comparative modeling consists of a three main steps: 1 – alignment of the target sequence and the sequence of the protein with known structure (template), 2 – building the model on the chosen template, 3 – assessing the model for its accuracy.

In the first step the search of representative template is done against the target sequence, whereas the primary and predicted secondary structures of the template and target sequence are compared and aligned. The fold recognition and alignment can be performed using web servers like Phyre and 3D-PSSM (http://www.sbg.bio.ic.ac.uk/3dpssm/). The obtained alignments and 3D structures of the models are then used as an input for the homology-modelling software.

A number of software and web servers for homology modelling are available. The MODELLER is one of them and available at http://salilab.org/modeller. MODELLER calculates a model by satisfaction of spatial restrains. The spatial restrains include restrains on distances and torsion angles, bond lengths and bond angles obtained from the CHARMM-22 molecular mechanics force field, statistical preferences for torsion angles, and non-bonded interatomic distances, which are obtained from a representative set of known structures. The accuracy of comparative models is judged by a variety of criteria and various model assessment scores. An interactive web service for the recognition of errors in three-dimensional structures of proteins is available under https://prosa.services.came.sbg.ac.at. PROSA is a diagnostic tool that is based on the statistical analysis of all available protein structures. The potentials of mean force compiled from the data base provide a statistical average over the known structures. Structures of soluble globular proteins whose *Z*-scores deviate strongly from the data base average are unusual, and frequently such structures turn out to be erroneous. Thus the *Z*-score of the correct homology model is expected to be within the range of scores typically found for proteins of similar size. Additionally programs like WHATIF and PROCHECK can be used to validate the geometrical quality of a model.

SWISS-MODEL is another, but automated comparative protein modeling server (http://swissmodel.expasy.org/SWISS-MODEL.html). It calculates models on the basis of sequence alignments using ProMod / ProModII algorithms (Peitsch, 1996) and energy minimization by Gromos96. The comparative protein modeling is guided by the alignment between target and template sequence, any error introduced by the alignment algorithm will have profound effects on the model. For instance, sequences sharing 40-49% identity with their template, submitted to SWISS-MODEL yielded a model deviating by less than 3 Å from their control structure for more than 60% of the sequences. This number increases to 79% for sequence identities ranging from 50 to 59%. Thus, the accurate alignment is the crucial step in homology modeling.

6.9 Channel calculations using CAVER

CAVER (http://loschmidt.chemi.muni.cz/caver/index.php) provides the calculation of pathways leading from cavities buried within a protein to outside solvent in static and dynamic protein structures. The study of these pathways is important to understand binding of inhibitors to receptors, substrate binding and product egress from enzyme active sites. The program calculates a protein grid and generates tunnel profiles (graph of cross section radius versus tunnel length) starting from the point defined by the user at traversing points in an empty space of a tunnel and preferring points that have more empty space around. Connecting of these points shows the path from to the protein surface, and the shortest path is calculated using Dijkstra's algorithm (Dijkstra, 1959). Calculated pathways can be visualized by a graphic program (for example PYMOL) dissecting anatomy and dynamics of entrance tunnels.