Chapter 1

Introduction

1.1 Protein crystallography

X-ray crystallography plays a major role in the understanding of biological processes at a molecular level by providing atomic models of macro-molecular molecules. Typically, in protein crystallography highly purified samples at high concentrations are crystallized using vapour or liquid diffusion methods. Nowadays, large amounts of genome sequences are available and due to well-defined expression systems and efficient purification protocols, samples can be produced at scales large enough for crystallization trials for many target proteins. With the maturation of third-generation synchrotron beam lines providing high-intensity X-ray beams, cryogenic sample protection, robotic sample changing and charge-coupled-device (CCD) detectors, fast and highly automated diffraction experiments are becoming a routine matter. Crystals of only a few micrometer in size are now suitable for crystallographic analysis (Cusack *et al.*, 1998), and crystal structures of molecular assemblies as large as the 50S subunit of the ribosome have already been solved (Ban *et al.*, 2000).

However, after fifty years of extensive research two bottle-necks remain in the process of protein crystal structure determination. Firstly, although automated setups require ever-decreasing amounts of sample material and efficient sparse-matrix screens of crystallization conditions have been developed (reviewed by Stevens, 2000), obtaining suitable crystals for diffraction purposes remains a difficult process that is poorly understood (reviewed by Gilliland & Ladner, 1996). Secondly, after obtaining diffracting crystals, phases need to be determined for the measured intensities in order to reconstruct the electron density of the unit cell. Since the early days of protein crystallography the most prominent answer to this problem has been based on the incorporation of heavy atoms in the crystal, but the search for appropiate soaking solutions is often a cumbersome one. The possibilities to incorporate covalently bound heavy atoms in the protein, mainly by the incorporation of seleno-methionine using bacterial expression systems, have contributed significantly to the successful application of experimental phasing techniques. However, with the increasing requirements of post-translational modifications for the more complex target structures of nowadays, equivalents for eukaryotic expression systems are awaited anxiously. In the favourable cases where a homologous structure is available, molecular replacement can be applied to solve the phase

problem. Despite large efforts throughout the field, other (*ab initio*) phasing techniques that do not depend on the incorporation of heavy atoms, have not yet provided a generally applicable answer to the phase problem in macro-molecular crystallography.

1.2 The phase problem

In the standard crystallographic experiment a crystal is positioned in a beam of monochromatic X-ray radiation. X-rays passing through the crystal will cause the electrons of the molecules to oscillate. These oscillating charges then emit X-ray radiation of the same wavelength in all directions. A crystal consists of a periodic arrangement of molecules where repeating units, the so-called unit cells, form a three-dimensional lattice. Because of this periodicity, the waves scattered by the atoms in all unit cells will only interfere constructively in certain discrete directions. This leads to the Laue diffraction conditions (1.1):

$$\begin{aligned} \vec{a}\cdot\vec{S} &= h_1 \\ \vec{b}\cdot\vec{S} &= h_2 \\ \vec{c}\cdot\vec{S} &= h_3 \end{aligned} \tag{1.1}$$

where $\vec{a}$, $\vec{b}$ and $\vec{c}$ are the translation vectors of the crystal lattice, Laue-indices h_1, h_2 and h_3 are integers and $\vec{S}$ is called the diffraction vector.

An alternative way to describe diffraction is to consider a diffracted beam as being reflected by a plane *hkl* through the endpoints of vectors $\vec{a}/h$, $\vec{b}/k$ and $\vec{c}/l$. (Miller indices *h*, *k*, and *l* are related to the Laue indices: $h_1 = nh$, $h_2 = nk$ and $h_3 = nl$, *n* being an integer). Diffraction only occurs if the angle θ of the incident beam with the lattice plane *hkl* satisfies Bragg's law (1.2):

$$2d_{hkl}\sin(\theta) = n\lambda \tag{1.2}$$

where d_{hkl} is the distance between adjacent lattice planes *hkl* and λ is the wavelength of the incident beam; integer *n* is called the order of the reflection.

The direction of diffraction vector $\vec{S}(hkl)$ is normal to the reflecting plane *hkl* and its length $|\vec{S}(hkl)|$ is equal to $1/d_{hkl}$. The endpoints of all vectors $\vec{S}(hkl)$ form a three-dimensional lattice with translation vectors $\vec{a}^*$, $\vec{b}^*$ and $\vec{c}^*$ (with $\vec{a}^* = \vec{b}\times\vec{c}/V$, $\vec{b}^* = \vec{c}\times\vec{a}/V$ and $\vec{c}^* = \vec{a}\times\vec{b}/V$, where V is the volume of the unit cell), the so-called reciprocal lattice. This lattice allows $\vec{S}(hkl)$ to be calculated in a convenient way from (1.3):

$$\vec{S}(hkl) = h\vec{a}^* + k\vec{b}^* + l\vec{c}^* \tag{1.3}$$

The experimentally measured intensity of reflection *hkl* depends on the distribution of the electrons in the unit cell: $\rho(xyz)$ and is proportional to the square of the amplitude of structure factor $F(hkl)$ (1.4):

$$F(hkl) = V\int_x\int_y\int_z \rho(xyz)\exp\{2\pi i(hx+ky+lz)\}dxdydz \tag{1.4}$$

with *x*, *y* and *z* being fractional coordinates.

By inverse Fourier transform the electron density distribution in the unit cell is calculated from structure factors $F(hkl)$ (1.5):

$$\rho(xyz) = 1/V \sum_h \sum_k \sum_l F(hkl) \exp\{-2\pi i(hx+ky+lz)\} \tag{1.5}$$

From the electron density distribution an atomic model of the molecules in the unit cell can be constructed. However, structure factors $F(hkl)$ in equation 1.5 are complex quantities with an amplitude and a phase (1.6):

$$F(hkl) = |F(hkl)| \exp(i\varphi(hkl)) \tag{1.6}$$

From the standard monochromatic experiment, amplitude $|F(hkl)|$ can be derived from the measured intensity, but all information about phases $\varphi(hkl)$ is lost. Therefore, the electron density distribution cannot be constructed directly using equation 1.5. This problem is known as the crystallographic phase problem.

1.3 Experimental phasing

In macro-molecular crystallography it is common practice to solve the phase problem using additional experimental information (see classic texts like Drenth, 1999 for more details). In the isomorphous replacement method, protein crystals are soaked in one or more solutions containing 'heavy' atoms. In the optimal case this leads to specific binding of the heavy atoms to the protein molecules and the soaked crystals stay isomorphous to the native crystal. The resulting differences in intensity of the observed reflections are exploited to obtain phase estimates. Many heavy atoms absorb X-ray radiation at wavelengths that are typically used in protein crystallography (0.6-2.0 Å). The absorbance of X-ray fotons gives rise to anomalous diffraction. In anomalous scattering methods, the intensity differences between Friedel pair reflections *hkl* and $-h-k-l$ (the so-called Bijvoet differences) are used to calculate phase estimates. The continuous tunability of synchrotron radiation sources makes it convenient to exploit yet another signal: dispersive intensity differences between data collected at different wavelenghts. In multiple-wavelength anomalous dispersion methods (MAD, Hendrickson & Ogata, 1997), combination of the anomalous and dispersive signals allows phase determination from a single crystal. MAD-phasing has become increasingly popular due to the possibility to bio-synthetically introduce anomalous scatterers into the protein itself (reviewed by Ogata, 1998). In particular the *Eschericia coli* bacterium is used to substitute methionine for seleno-methionine. This yields anomalously scattering molecules with essentially equal structural properties compared to the native protein. Recently, density modification methods (see also below) have been applied successfully to substitute the need for dispersive signals, allowing phasing by anomalous scattering using only a single wavelength (SAD, as advocated by Wang, 1985).

1.4 Ab initio phasing

In *ab initio* phasing phase estimates are obtained from a single set of structure factor amplitudes, without using any of the experimentally determined intensity differences described above. In this case, the phase problem may be overcome by incorporation of additional, *a priori* available knowledge.

1.4.1 Direct methods

Even the simplest form of prior knowledge, the expectation that the electron density is non-negative and consists of separated atoms throughout the unit cell, leads to statistical relationships among the structure factors that can be used to solve the phase problem. In the 1950's Hauptman & Karle defined a functional form to express these relationships and thereby opened the field of direct methods. Ever since, this approach has increased its power and nowadays it is used to routinely solve thousands of structures with up to 250 non-hydrogen atoms every year. The ultimate potential of this method is still unknown; its only limitation is that it requires diffraction data up to atomic, *i.e.* $\sim$1.2 Å resolution. (reviewed by Hauptman, 1997). For the direct phasing of macromolecules, up to now, direct methods have proven of limited use. The reliability with which the phases can be estimated decreases rapidly with the number of atoms in the unit cell and in protein crystallography the requirement of diffraction data up to atomic resolution is not often met. Recent advances, where reciprocal-space phase refinement is combined with modifications in real-space, the so-called baked (Weeks *et al.*, 1993) and half-baked (Sheldrick & Gould, 1995) methods have allowed the direct phasing of small protein structures of up to 1,000 non-hydrogen atoms. Provided that some initial phasing from a substructure is available, larger structures can be solved by a combination of direct methods and density modification (Foadi *et al.*, 2002).

1.4.2 Molecular replacement

A much more common way of *'ab initio'* phasing in protein crystallography is molecular replacement. In molecular replacement (reviewed by Rossmann, 2001) the known structure of a homologous protein is used as prior information in the phasing process. Phases are obtained by correctly positioning the known model in the crystal lattice of the unkown structure. Traditionally, this problem has been broken down into two three-dimensional search problems. Using Patterson methods, first the correct orientation is determined, followed by a search for the correct translation vector. Recently also programs performing complete (Sheriff S. *et al.*, 1999) or directed (Kissinger *et al.*, 1999 and Glykos *et al.*, 2000) six-dimensional searches have been developed. Another recent development is the application of maximum likelihood to molecular replacement, which may allow positioning molecules of significantly lower homology (Read, 2001).

1.4.3 Low-resolution phasing

Several attempts have been made to solve the phase problem by first finding the protein molecular envelope in the unit cell, which encompasses phasing of only the lowest resolution

reflections. Urzhumtsev *et al.* (2000) applied various selection criteria based on histograms and connectivity of the electron density map to select the better set of phases from a large number of random trial sets. None of their criteria proved capable of unambiguously distinguising good from bad phase sets, but by making use of the statistical tendency of good phase sets to have better criterion values than bad sets, enrichment of the phase quality could be obtained. For a test case of protein G these procedures resulted in moderate phase information for reflections up to 4-5 Å resolution (Lunina *et al.*, 2000). Additionally, instead of generating random phases, these groups have tried to use large-sphere models that are placed randomly in the unit cell to generate trial phases. This so-called few atom method uses the correlation coefficient between calculated and observed structure factors as selection criterion in the enrichment process (Lunin *et al.*, 1995, 1998). A combination of the methods developed by these groups allowed phasing up to 40 Å resolution of the ribosomal 50S particle from *Thermus thermophilus* (Lunin *et al.*, 2000).

A number of other methods to solve protein structures starting from the lowest resolution reflections have been developed but none of them have come into common practice. By systematic translation of a large sphere filled with point scatterers at regular intervals and monitoring the crystallographic *R*-factor, Harris (1995) could determine the correct molecular envelope for some test cases. However, due to the limitation of a spherical search model, the method showed limited success for solvent regions with a significantly deviating shape. Subbiah (1991) used refinement of randomly distributed hard sphere point scatterers to phase the lowest resolution reflections. Although initially this method yielded solutions with equal likelihood of the point scatterers ending up in the solvent or in the protein region, later successful methods were developed to distinguish these solutions (Subbiah, 1993). Guo *et al.* (2000) applied the probabilistic approach from conventional direct methods to low-resolution phasing, avoiding the necessity of atomic resolution data by using globbic scattering factors representing multiple protein atoms. Complementation of the missing lowest-resolution reflections with calculated data appeared critical for the successful application of this method. The dependance on complete low resolution data is a common feature for most low-resolution phasing methods. Up to now, none of these methods have bridged the gap between phases providing a low-resolution molecular envelope and phases of sufficient quality to allow phase extension to medium or high resolution with density modification methods (see next section).

1.5 Phase improvement

1.5.1 Density modification

The incorporation of prior information can be used to extend phase information as obtained by the methods described above (reviewed by Abrahams & De Graaff, 1998). Protein crystals typically contain 30-70% solvent, organized in channels of unordered water molecules. In solvent flattening (Wang, 1985), the electron density is constrained towards a flat solvent region, and this real-space density modification is iterated with a phase-combination step in reciprocal space. A similar iterative procedure is used in histogram matching where prior information in the form of expected density histograms is applied as constraints on the electron density map. Similarly, knowledge of non-crystallographic symmetry (NCS) can be used to

modify the electron density by averaging over independent molecules. This NCS-averaging has proven very powerful to extend the available phase information when multiple copies of the same molecule are present in the asymmetric unit. For some viruses for example, extensive NCS-averaging has allowed *ab initio* phasing starting from simple geometric models like a hollow sphere (reviewed by Rossmann, 1995). A new development in the field of density modification is the implementation of maximum-likelihood theory in the *RESOLVE* program (Terwilliger, 2000).

1.5.2 Model building

After an electron density map has been obtained from initial phasing and density modification techniques, interpretation of this map in terms of a protein model is required. In this process prior knowledge of the amino-acid sequence as well as the known structural characteristics of protein molecules are of great importance. Therefore, visualization programs for manual model building like *O* (Jones *et al.*, 1991), make extensive use of databases of commonly observed main and side-chain conformations. Still, manual model building remains not only a time-consuming process but also a subjective one, shown to be prone to human error (Mowbray *et al.*, 1999). Major advances have been achieved in more automated ways of map interpretation. Pattern recognition methods, exploiting similar knowledge as used in manual building, have been implemented in semi-automated model-building programs like *TEXTAL* (Holton *et al.*, 2000), *RESOLVE* (Terwilliger, 2002) and *QUANTA* (Oldfield, 2000). Provided initial phases of sufficient quality, such methods have been shown to work at resolution limits of $\sim$3Å. With such limited amounts of diffraction data the generated models may suffer a rather low accuracy. Iteration of model building steps with refinement cycles (see next section), as already implemented in the *RESOLVE* program, may provide a solution to this problem.

Up till now the most widely used program for automated model building is *ARP/wARP* (Perrakis *et al.*, 1999). In *ARP/wARP*, electron density maps are interpreted in terms of free atoms, which are refined using the *ARP* procedure (Lamzin & Wilson, 1993), where (almost) unrestrained refinement is combined with atom repositioning based on various types of electron density maps. Subsequently, the warp*N*trace procedure (Perrakis *et al.*, 1999) exploits prior knowledge about oligo-peptide conformations to identify and trace possible main-chain fragments through the optimized distributions of free atoms. The resulting 'hybrid' model allows free-atom refinement to be combined with the application of standard geometric restraints, reducing the danger of overfitting the data. The main limitation of the *ARP/wARP* program is that it requires data to relatively high resolution limits since the unrestrained refinement cycles depend on a favourable observation-to-parameter ratio and the repositioning of separate atoms in the *ARP* procedure requires electron density maps of sufficient resolution. Recently, major advances have been achieved for the warp*N*trace algorithm (Morris *et al.*, 2002), currently allowing automated main-chain tracing at resolution limits of 2.5 Å.

1.5.3 Refinement

The building of a protein model is often hampered by a poor quality of the phase information or limited resolution of the diffraction data. Therefore, the initial model generally contains

errors and must be optimized. The goal of crystallographic refinement can be formulated as finding the set of atomic coordinates that results in the best fit of the observed structure factor amplitudes and the amplitudes calculated from this model. In conventional least-squares refinement (see for example Drenth, 1999), this goal has been formulated as finding the minimum of target function (1.7):

$$E_{\mathrm{X-ray}} = \sum_{hkl} w_{hkl} \left(|F_{hkl}^{\mathrm{obs}}| - k|F_{hkl}^{\mathrm{calc}}|\right)^2 \tag{1.7}$$

Where w_{hkl} is a weighting factor, k is a scale factor and the calculated structure factor amplitude $|F^{\mathrm{calc}}|$ is dependent on the parameter set of the model. Calculation of derivatives of $|F^{\mathrm{calc}}|$ towards the model parameters allows application of gradient-driven optimization techniques to minimize this function.

Major advances in refinement have been made by the formulation of maximum likelihood target functions (reviewed by Bricogne, 1997). In contrast to least-squares methods, maximum likelihood provides a statistically valid way to deal with errors and incompleteness of the model. A general approach is to represent the resolution-dependent quality of the model by the σ_A-distribution (1.8):

$$\sigma_A = \langle E^{\mathrm{obs}} \cdot E^{\mathrm{calc}} \rangle \tag{1.8}$$

where σ_A-values are calculated in resolution bins and E^{obs} and E^{calc} are observed and calculated normalized structure factors. Since the phases of E^{obs} are unknown, σ_A-values need to be estimated. For this purpose Read (1986) developed a method called *SIGMAA*. Cross-validation, initially introduced to monitor over-fitting of the data by calculation of a free *R*-factor (Brünger, 1993), plays an important role in the estimation of σ_A-values (Adams *et al.*, 1997). In cross-validation typically 5-10% of the data (the test set) is kept outside the refinement. Estimation of σ_A-values based on these test set reflections avoids serious overestimation resulting from overfitting of the data. The probability to observe E^{obs}, given E^{calc} of the model, can then be calculated by (1.9):

$$P\left(E^{\mathrm{obs}}; E^{\mathrm{calc}}\right) = \frac{1}{\pi(1-\sigma_A^2)} \exp\left(-\frac{|E^{\mathrm{obs}} - \sigma_A E^{\mathrm{calc}}|^2}{1-\sigma_A^2}\right) \tag{1.9}$$

Similar equations are derived to calculate the probability to observe structure factor amplitude $|F^{\mathrm{obs}}|$, given calculated structure factor F^{calc} and the measurement error in $|F^{\mathrm{obs}}|$. Maximum likelihood refinement aims to maximize the likelihood of measuring the set of observed structure factor amplitudes, given the calculated structure factors of the model.

A key factor in crystallographic refinement is the ratio of observations to parameters. An atomic model is only justified when data to atomic resolution is available. In protein crystallography typically resolution limits in the range of 1.5-3.5 Å are observed. To avoid over-fitting of these limited amounts of experimental data, the number of observations is effectively enlarged by the incorporation of prior geometrical knowledge. This knowledge can be expressed as real-space restraints on expected bond distances, angles and torsion angles, defining a combined target function (1.10):

$$E = E_{\mathrm{geom}} + waE_{\mathrm{X\text{-}ray}} \tag{1.10}$$

with (1.11):

$$E_{\text{geom}} = \sum_{\text{bonds}} w_{\text{bond}}(r_{\text{bond}}^{\text{ideal}} - r_{\text{bond}}^{\text{model}})^2 + \sum_{\text{angles}} w_{\text{angle}}(\theta_{\text{angle}}^{\text{ideal}} - \theta_{\text{angle}}^{\text{model}})^2 + \ldots \quad (1.11)$$

where bond distances r_{bond}, angles θ_{angle}, and other geometric parameters of the protein model like torsion angles, planarity of rings *etc.* are restrained towards their ideal values using weights *w*. Weight *wa* for the crystallographic term is chosen such that approximately equal gradient contributions from both sides of the combined target function result. If only data to moderate resolution limits ($d_{min} > 2.5$ Å) is available, constraints on bond distances and angles are justified. This limits the degrees of freedom to torsion angles, thus resulting in a further improved observation-to-parameter ratio. A torsion-angle parameterization of protein molecules has been implemented in the *CNS* program (Brünger *et al.*, 1998a, 1998b). Combined with a maximum likelihood crystallographic target function and a powerful simulated annealing optimization protocol, this makes *CNS* a preferred program for refinement of protein structures when the available data is not extending beyond 2.5 Å resolution. Multiple annealing runs starting from different initial velocities have been shown to result in optimized models that show largest spread in poorly fitted regions. Averaging over the individual solutions of this multi-start method gives a better structure factor set (Rice *et al.*, 1998). A recent development in protein structure refinement is the possibility to model anisotropic motions of complete domains by TLS-parameterization for the translation, libration and screw-rotation displacements of pseudo-rigid bodies (introduced by Schomaker & Trueblood, 1968), as implemented in the maximum likelihood refinement program *REFMAC* (Winn *et al.*, 2001).

Despite the developments mentioned above, the radius of convergence of protein structure refinement remains limited and in current practice refinement cycles still need to be iterated with time-consuming rebuilding steps where the model is improved manually by interpretation of electron density maps.

1.6 Conditional Optimization

In the described steps to obtain phase information in protein crystallography, the incorporation of prior knowledge plays a critical role to supplement the limited amounts of diffraction data. The information that is used in these steps comes from a common source: in general we know how protein molecules look like and that they form crystals with disordered solvent regions. This knowledge, embedded in the coordinates of many entries in the protein structure data base (PDB: Berman *et al.*, 2002), can be expressed in different ways. It typically depends on the quality of the available phases how much prior knowledge can be expressed in an efficient way. For example, in the absence of any phase information, limited knowledge about non-negativity and atomicity of electron density is expressed as probabilistic relationships among phases in direct methods. Given some initial phases, more specific knowledge about a flat solvent region is expressed as constraints on the electron density in density modification techniques. At the final stages of the structure determination process, when enough phase information is available for construction of a molecular model, extensive knowledge about the geometries of amino acids is expressed as restraints on bond distances and (torsion) angles in protein structure refinement.

In this thesis, a novel protein structure refinement method is presented, called conditional optimization. The conditional formalism allows expression of prior knowledge about the geometry of protein structures without the requirement of a molecular model, and thus potentially in the absence of phase information. Available knowledge about the geometries of protein fragments up to several residues long and in many possible conformations is expressed as real-space interaction functions acting on loose, unlabelled atoms. In an N-particle approach, all topological and conformational possibilities are taken into account for all combinations of loose atoms. Although other potential applications exist, this method would ultimately allow *ab initio* phasing of protein structures at medium resolution limits. Based on the assumption that combination of the defined interaction functions with a maximum likelihood crystallographic target function fully defines the system, the phase problem is rephrased as a search problem. Hereby, starting refinement from random atom distributions solving the phase problem "merely" requires an efficient search strategy to reach the global minimum of these functions. For this purpose, gradient-driven optimization methods like energy minimization and dynamics calculations are applied. The N-particle approach of the conditional formalism, combined with maximum likelihood crystallographic target functions and powerful optimization protocols, may provide a protein structure refinement method with a large radius of convergence.

1.7 Scope and outline of this thesis

In this thesis, the method of conditional optimization is presented and its potentials in crystallographic phasing are investigated. In **chapter 2** the general principles of the method of conditional optimization are presented, together with initial calculations using a simplified polyalanine test structure. These tests show that, in principle, refinement starting from random atom distributions is possible and *ab initio* phasing can be achieved using only medium resolution data. **Chapter 3** describes the development of a potential of mean force suitable for conditional optimization of protein molecules. This chapter includes test calculations with the defined force field on three small protein structures against observed diffraction data, for which a large radius of convergence was observed. In **chapter 4** the potentials of conditional optimization in automated map interpretation are explored. For three test cases at medium resolution, automated model building by conditional optimization yielded results comparable to *ARP/wARP* and *RESOLVE*. **Chapter 5** describes the application of conditional optimization to *ab initio* phasing of observed diffraction data to medium resolution. For the presented test case promising results were obtained, indicating that successful optimization of random atom distributions may be possible, although further developments are currently limited by excessive computational costs. The last chapter, **chapter 6**, gives a summary of the work described in this thesis and a short elaboration on the perspectives of conditional optimization in protein crystallography.