Conditional Optimization

An *N*-particle formalism for protein structure refinement

Conditionele Optimalisatie

Een *N*-deeltjes formalisme voor eiwitstructuurverfijning

(met een samenvatting in het Nederlands)

PROEFSCHRIFT

TER VERKRIJGING VAN DE GRAAD VAN DOCTOR AAN DE
UNIVERSITEIT UTRECHT OP GEZAG VAN DE
RECTOR MAGNIFICUS, PROF. DR. W.H. GISPEN,
INGEVOLGE HET BESLUIT VAN HET COLLEGE VOOR
PROMOTIES IN HET OPENBAAR TE VERDEDIGEN OP
MAANDAG 16 JUNI 2003 DES OCHTENDS OM 10.30 UUR

door

SJORS HENDRIK WILLEM SCHERES

geboren op 5 augustus 1975, te Roermond

Promotor: prof. dr. P. Gros
verbonden aan het Bijvoet Centrum voor Biomoleculair Onderzoek,
Faculteit Scheikunde der Universiteit Utrecht

Dit proefschrift werd mede mogelijk gemaakt door financiële steun van NWO
(Jonge Chemici 99-564)

ISBN: 90-393-3352-1

Ter nagedachtenis aan Jan Kroon

Reproductie: Drukkerij Labor, Utrecht

Omslag
voorzijde: Hooiberg bij Huub vanne Waal in Stramproy, Limburg.
achterzijde: Naaldvormig eiwitkristal, gegroeid in de ruimte (bron: NASA).

Contents

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.

Chapter 2

Conditional Optimization: a new formalism for protein structure refinement [1]

Abstract

Conditional Optimization allows unlabelled, loose atom refinement to be combined with extensive application of geometrical restraints. It offers an *N*-particle solution for the assignment of topology to loose atoms, with weighted gradients applied to all possibilities. For a simplified test structure, consisting of a polyalanine four-helical bundle, this method shows a large radius of convergence using calculated diffraction data to at least 3.5 Å resolution. It is shown that, with a new multiple-model protocol to estimate σ_A-values, this structure can be successfully optimised against 2.0 Å resolution diffraction data starting from a random atom distribution. Conditional Optimization has potentials for map improvement and automated model building at low or medium resolution limits. Future experiments will have to be performed to explore the possibilities of this method for *ab initio* phasing of real protein diffraction data.

[1] Sjors H.W. Scheres & Piet Gros (2001) *Acta Cryst. D***57**, 1820-1828

2.1 Introduction

A critical step in crystallographic protein-structure determination is deriving phase information for the measured amplitude data. Direct calculation of phases or phase improvement depends on the use of prior information about the content of the unit cell. The simplest form of information, *i.e.* non-negativity and atomicity, is sufficient when diffraction data is available to very high resolution (Bragg spacing $d < 1.3$ Å). The methods of *Shake-and-Bake* (Weeks *et al.*, 1993) and *Half-baked* (Sheldrick and Gould, 1995) solve protein structures using near-atomic resolution by combining phase refinement in reciprocal space and an elementary form of density modification in real space, *i.e.* atom positioning by peak picking in the electron density map. Alternatively, for approximate phasing of low-resolution diffraction data, prior information about connectivity and globbicity of protein structures has been applied using few-atom models (Lunin *et al.*, 1998; Subbiah, 1991). More typically, in protein crystallography structure determination uses initial phases that are derived by either experimental methods (reviewed by Ke, 1997; Hendrickson and Ogata, 1997) or through the use of a known homologous structure (reviewed by Rossmann, 1990). Improvement of these initial phase estimates may be achieved by including prior knowledge of e.g. flatness of the electron density in the bulk solvent region or non-crystallographic symmetry among independent molecules by the technique of density modification (reviewed by Abrahams and De Graaff, 1998). At the last stage, *i.e.* in protein-structure refinement, the prior knowledge of protein structures is used in the form of e.g. specific bond lengths, bond-angles and dihedral angles (reviewed by Brünger *et al.*, 1998a). In these processes of phase improvement the prior knowledge is essential to supplement the limited amount of information available when the resolution of the diffraction data is insufficient.

Here, we focus on the application of the prior knowledge of protein structures, *i.e.* the arrangement of protein atoms in polypeptide chains with secondary structural elements. This information is most easily expressed in real space using atomic models. Optimization of these models against the available X-ray data and the geometrical restraints is, however, complicated by the presence of many local minima. Therefore, the refinement procedures have limited convergence radii and optimization depends on iterative model building and refinement. Probably, the search problem is greatly reduced when using loose atoms instead of polypeptide chains with fixed topologies (see Isaacs and Agarwal, 1977, for an early use of loose atom refinement). However, in the absence of a topology the existing methods cannot apply the available geometrical information. As a compromise the *ARP/wARP* method (Perrakis *et al.*, 1999) uses a hybrid model of restrained structural fragments and loose atoms. This has allowed structure building and refinement in an automated fashion, when data to $\sim$2.3 Å resolution and initial phase estimates are available. Critical in this process is the information content that allows approximate positioning of loose atoms and subsequent identification of structural fragments. A procedure, in which more information can be applied to loose atoms, may depend less on the resolution of the diffraction data and the quality of the initial phase set.

Here, we present a new formalism that allows conditional formulation of target functions in structure optimization. Using this formalism, we can express the geometrical information of protein structures in terms of loose atoms. Our approach overcomes the problem that, in general, a chemical topology cannot be assigned unambiguously to loose atoms. We con-

sider all possible interpretations, based on the structural similarity between the distribution of loose atoms and that of given protein fragments. Weighted geometrical restraints are applied in the optimization according to the extent by which the individual interpretations could be made. In effect, the formalism presented here yields an N-particle solution to the problem of assigning a topology to a given atomic coordinate set. Thereby, the method of conditional optimization combines the search efficiency of loose atoms with the possibility of including large amounts of geometrical information. The information expressed, using the conditional formalism, includes structural fragments of protein structures from single bonds up to secondary structural elements. We show that for a simple test case this method yields reliable phases when starting from random atom distributions.

2.2 Conditional Formalism

In the conditional formalism we describe a protein structure by linear elements, which are non-branched sequences of atoms occurring in the protein structure. A protein structure contains various types of these linear elements with characteristic geometrical arrangements of the atoms (one example of such a type is the typical arrangement of the atoms *CA*-*C*-*N*-*CA* in a peptide plane). Using simple geometric criteria, we express the structural resemblance of a set of loose atoms to any of the expected structural elements in a protein structure. The amino acid sequence and predicted secondary structure content determine the types of elements that we expect for a given protein. The geometrical arrangements of these types can be deduced from known protein structures. The best arrangement of loose atoms, corresponding to the minimum of the target function, is a distribution with exactly the expected number of structural elements present as given by the protein sequence and expected secondary structure.

We define a linear structural element as a non-branched sequence of atoms $ij\ldots pq$ of L bonds long, containing $L+1$ atoms. A linear structural element of atoms $ij\ldots pq$ of length L is composed of two linear sub-elements $ij\ldots p$ and $j\ldots pq$, both of length $L-1$ (see Figure 2.1). We define conditions C, which are continuous functions with $C=[0,1]$, assigned to each of these elements. Conditions C reflect the degree to which a geometrical criterion is fulfilled associated with forming a specific type of element from its two sub-elements. When considering only distance criteria, the conditions C become pair-wise atomic interaction functions (see Figure 2.2).

A linear element of length L is then described by a joint condition JC, which is a product of conditions C according to the binary decomposition of the linear element into its sub-elements. Thus, the $(L+1)$-particle function $JC_{i\ldots q}$ for a linear structure consisting of atoms $i\ldots q$ forming L bonds is expressed in a (binomial) product of $L(L+1)/2$ pair-wise functions.

Figure 2.1 shows an example of a binary combination of four atoms i, j, k and l resembling a peptide plane. A peptide plane is composed of six types of linear elements: bonds *CA*-*C*, *C*-*N* and *N*-*CA*, bond-angles *CA*-*C*-*N* and *C*-*N*-*CA* and peptide plane *CA*-*C*-*N*-*CA*. For each type of element a pair-wise interaction function C^{type} is assigned. The resemblance of the four atoms to a peptide plane can then be expressed by the following multiplication of functions C^{type} yielding joint condition $JC_{ijkl}^{CA-C-N-CA}$, which depends on all six inter-atomic

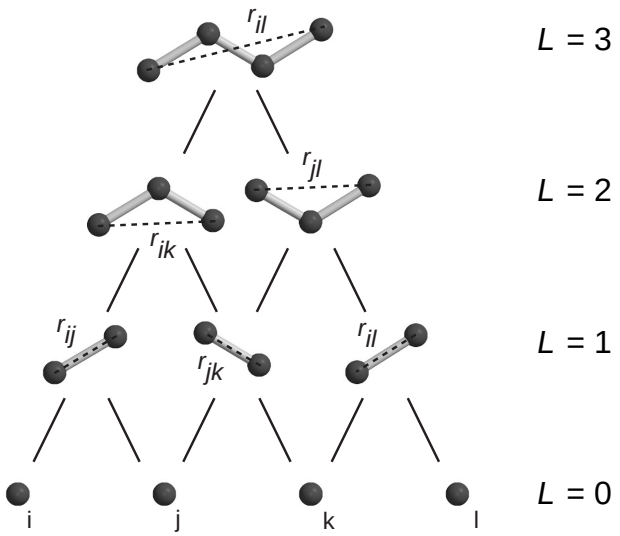

Figure 2.1: *Formation of a peptide plane by binary combinations of four loose atoms, three bonds and two bond-angles. For each binary combination of two sub-elements of length $L-1$ into one element of length L, a condition is assigned. These conditions represent geometrical criteria, e.g. depending on the inter-atomic distance between the two outer atoms of an element. The resemblance of four atoms i, j, k, and l to a peptide plane is given by multiplying the conditions into a joint condition, as defined in Equation 2.1.*

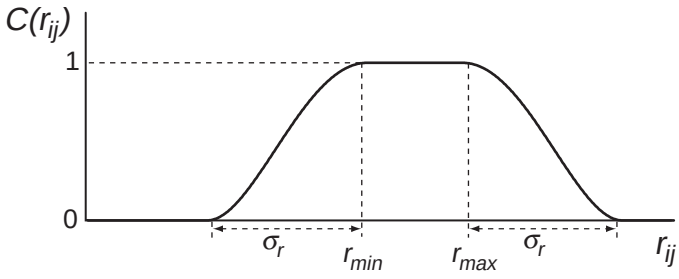

Figure 2.2: *Conditions $C(r_{ij})$ are defined by an optimal range of distances from $r_{\min}$ to $r_{\max}$ and a fourth-order polynomial slope with a width of σ_r:$C(r_{ij}) = 0$ for $r_{ij} \leq r_{\min} - \sigma_r$; $C(r_{ij}) = \{1 - [(r_{\min} - r_{ij})/\sigma_r]^2\}^2$ for $r_{\min} - \sigma_r < r_{ij} < r_{\min}$; $C(r_{ij}) = 1$ for $r_{\min} \leq r_{ij} \leq r_{\max}$; $C(r_{ij}) = \{1 - [(r_{\max} - r_{ij})/\sigma_r]^2\}^2$ for $r_{\max} < r_{ij} < r_{\max} + \sigma_r$; $C(r_{ij}) = 0$ for $r_{ij} \geq r_{\max} + \sigma_r$.*

distances r_{ij}, r_{jk}, r_{kl}, r_{ik}, r_{jl} and r_{il}:

$$\begin{aligned} JC_{ijkl}^{CA-C-N-CA} &= C^{CA-C}(r_{ij})C^{C-N}(r_{jk})C^{CA-C-N}(r_{ik})C^{C-N}(r_{jk}) \\ &\quad \times C^{N-CA}(r_{kl})C^{C-N-CA}(r_{jl})C^{CA-C-N-CA}(r_{il}) \end{aligned} \tag{2.1}$$

Generalized forms of joint conditions for linear elements of $L = 2$ and $L \geq 3$ are shown in Equation 2.2 and 2.3, respectively. An element of length L of a specific *type* is formed by combination of its two sub-elements of *subtype-A* and *subtype-B*, both of length $L-1$.

$$JC_{ijk}^{\text{type}} = C^{\text{subtype}-\text{A}}(r_{ij})C^{\text{subtype}-\text{B}}(r_{jk})C^{\text{type}}(r_{ik}) \tag{2.2}$$

$$JC_{ij\ldots pq}^{\text{type}} = JC_{ij\ldots p}^{\text{subtype}-\text{A}} JC_{j\ldots pq}^{\text{subtype}-\text{B}} C^{\text{type}}(r_{iq}) \tag{2.3}$$

where JC_{ijk}^{type} is the joint condition of linear element ijk of length $L = 2$. $C^{\text{subtype}-\text{A}}(r_{ij})$, $C^{\text{subtype}-\text{B}}(r_{jk})$ and $C^{\text{type}}(r_{ik})$ are pair-wise conditions defined for the terminal atoms i and j, j and k, i and k of elements ij, jk and ijk with lengths L of 1, 1 and 2 respectively; $JC_{ij\ldots pq}^{\text{type}}$, $JC_{ij\ldots p}^{\text{subtype}-\text{A}}$ and $JC_{j\ldots pq}^{\text{subtype}-\text{B}}$ are joint conditions of linear elements $ij\ldots pq$, $ij\ldots p$, and $j\ldots pq$ of lengths L, $L-1$ and $L-1$ respectively, and $C^{\text{type}}(r_{iq})$ is a pair-wise condition defined for the terminal atoms i and q of elements $ij\ldots pq$ of length L.

To describe a complete protein structure, we define target functions expressing the expected occurrence of linear structural elements. For each type of linear element of length L a target function E^{type} is defined, see Equation 2.4.

$$E^{\text{type}} = w^{\text{type}}\left(TC^{\text{type}} - \sum_{ij\ldots pq} JC_{ij\ldots pq}^{\text{type}}\right)^2 \tag{2.4}$$

where w^{type} is a weighting factor and TC^{type} is the expected sum of joint conditions for this particular type of element of length L in the target structure, and where the summation runs over all combinations of $L+1$ atoms $ij\ldots pq$. The total target function E for a given protein structure is then given by the summation of over all expected types (Equation 2.5):

$$E = \sum_{\text{type}} E^{\text{type}} = \sum_{\text{type}} w^{\text{type}}\left(TC^{\text{type}} - \sum_{ij\ldots pq} JC_{ij\ldots pq}^{\text{type}}\right)^2 \tag{2.5}$$

Since the joint conditions $JC_{ij\ldots pq}^{\text{type}}$ are expressed as products of continuous and non-negative functions C, the derivatives with respect to inter-atomic distances for non-zero joint conditions may be computed according to Equation 2.6.

$$\frac{\partial}{\partial r_{kl}} \sum_{ij\ldots pq} JC_{ij\ldots pq}^{type} = \sum_{\ldots k\ldots l\ldots} \frac{nJC_{\ldots k\ldots l\ldots}^{\text{type}}}{C^{\text{subtype}}(r_{kl})} \frac{\partial C^{\text{subtype}}(r_{kl})}{\partial r_{kl}} \tag{2.6}$$

where the summation on the right-hand side runs over linear elements $\ldots k\ldots l\ldots$, which form a subset of linear elements $ij\ldots pq$ that contain both atoms k and l; C^{subtype} is a condition contributing to $JC_{\ldots k\ldots l\ldots}^{\text{type}}$ depending on the interatomic vector r_{kl}, and n is the power of C^{subtype} in the binomial distribution of $JC_{\ldots k\ldots l\ldots}^{\text{type}}$. Equation 2.7 shows the derivative of the target function given in Equation 2.4.

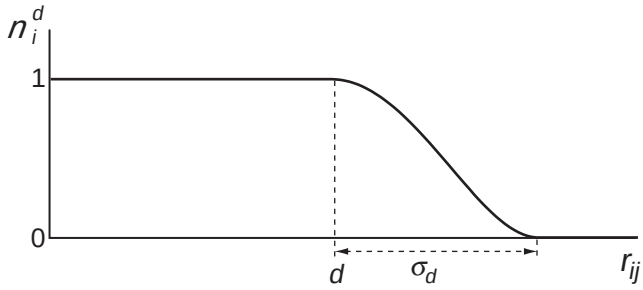

Figure 2.3: *Neighbouring atoms j around atom i are counted using a continuous function* n_i^d*:* $n_i^d(r_{ij}) = 1$ *for* $r_{ij} \leq d$*;* $n_i^d(r_{ij}) = \{1-[(d-r_{ij})/\sigma_d]^2\}^2$ *for* $d < r_{ij} < d+\sigma_d$*;* $n_i^d(r_{ij}) = 0$ *for* $r_{ij} \geq d+\sigma_d$ *The total number of neighbours,* $N_i^d = \sum_j n_i^d(r_{ij})$*, is used to calculate a neighbour-condition* $C_i^0(N_i^d)$*. Given an optimal range for the number of neighbouring atoms* $N_{\min}$ *to* $N_{\max}$ *and a width* σ_N *for the fourth-order polynomial slope, this condition can be calculated using the functional form as described in Figure 2.2.*

$$\begin{aligned}\frac{\partial E^{\text{type}}}{\partial r_{kl}} &= -2\sum_{\ldots k\ldots l\ldots} w^{\text{type}}\left(TC^{\text{type}} - \sum_{ij\ldots pq} JC^{\text{type}}_{ij\ldots pq}\right) \\ &\quad \times \frac{nJC^{\text{type}}_{\ldots k\ldots l\ldots}}{C^{\text{subtype}}(r_{kl})}\frac{\partial C^{\text{subtype}}(r_{kl})}{\partial r_{kl}} \\ &= G^{\text{type}}_{kl}\frac{1}{C^{\text{subtype}}(r_{kl})}\frac{\partial C^{\text{subtype}}(r_{kl})}{\partial r_{kl}} \end{aligned} \tag{2.7}$$

where G^{type}_{kl} is the sum of gradient coefficients from all linear elements depending on $C^{\text{subtype}}(r_{kl})$. Equation 2.7 shows that the effective weight on a gradient for a particular subtype depends on the extent to which this particular subtype-element is incorporated into larger structural elements. Total gradients can be calculated efficiently, because in the summation over all types of linear elements (see Equation 2.5) gradient coefficients G^{type}_{kl} can be precalculated for all subtypes, so that for each interacting pair of atoms *kl* only a summation over the subtypes needs to be performed.

The formulation given above is not restricted to pair-wise, distance functions. We have extended the description of protein structures with conditions for packing densities and chirality. For all atoms *i* atomic conditions C_i^{atomtype} (L = 0) are defined, depending on the expected number of neighbouring atoms around an atom of a specific *atomtype* (see Figure 2.3). Thereby, linear elements of a single bond (L = 1) are then described by a joint condition (Equation 2.8):

$$JC^{\text{type}}_{ij} = C_i^{\text{atomtype-A}} C_j^{\text{atomtype-B}} C^{\text{type}}(r_{ij}) \tag{2.8}$$

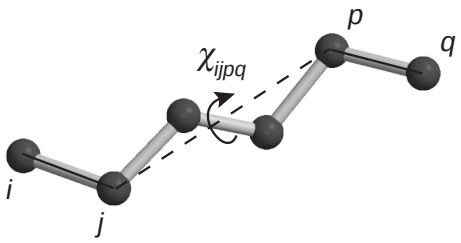

Figure 2.4: *A dihedral angle χ_{ijpq} is defined for the four outermost atoms i, j, p and q of any linear element $ij\ldots pq$ of length $L \geq 3$. Given an optimal value χ_{opt} for this dihedral angle, a condition $C_{\chi}^{\text{type}}(\chi_{ijpq})$ can be defined as: $C_{\chi}^{\text{type}} = \{1 - [(\chi_{\text{opt}} - \chi_{ijpq})/\pi]^2\}^2$*

Conditions C_{χ}^{type} are defined that describe the chirality of linear structures $ij\ldots pq$ with $L \geq 3$ (see Figure 2.4). Thereby, Equation 2.3 becomes Equation 2.9:

$$JC_{ij\ldots pq}^{\text{type}} = JC_{ij\ldots p}^{\text{subtype-A}} JC_{j\ldots pq}^{\text{subtype-B}} C^{\text{type}}(r_{iq}) C_{\chi}^{\text{type}}(\vec{r}_i, \vec{r}_j, \vec{r}_p, \vec{r}_q) \tag{2.9}$$

where chirality condition C_{χ}^{type} depends on positional vectors $\vec{r}_i$, $\vec{r}_j$, $\vec{r}_p$ and $\vec{r}_q$.

2.3 Experimental

2.3.1 Implementation

The formalism as described in the previous section has been implemented as a non-bonded routine in the *CNS* program (Brünger *et al.*, 1998b). A slight modification of Equation 2.4 is used for the target functions:

$$E^{\text{type}} = \frac{\left(TC^{\text{type}} - \sum_{ij\ldots pq} JC_{ij\ldots pq}^{\text{type}}\right)^2}{TC^{\text{type}}} - TC^{\text{type}} \tag{2.10}$$

By dividing by TC^{type} the pseudo-potential energy function depends linearly on the size and complexity of the system. Energies E^{type} range from zero (when e.g. none of the joint conditions is fulfilled) to -TC^{type} (all joint conditions fulfilled).

To compute all non-zero joint conditions a binary tree is generated starting from the atom-pair list. Joint conditions (see Equations 2.8 and 2.9) are computed for all defined types moving from the bottom layer, *i.e.* atoms ($L = 0$), 'upwards' to higher levels of bonded conditions ($L \geq 1$). Energies are computed, see Equations 2.5 and 2.10, when all joint conditions are known. Gradients are computed moving 'downwards' from the defined top level to the bottom layer (see Equation 2.7). The gradient coefficients G^{type} are computed by summation while moving downwards through the binary tree. For each node in the tree the gradient is computed once.

The number of interactions equals the total number of nodes, which is in the order of the number of atoms, N_{atoms}, times the number of types, M_{types} (where the number of types

are summed over all defined conditional layers L; for a simple all-helical poly-alanine model $M_{\text{types}} = 71$, when defining $L = 9$ conditional layers). The full binary tree with (non-zero) joint conditions is stored in memory at each pass. M_{types} is a fixed number given the complexity and the number of conditional layers defined. Thus, the order of the algorithm is $O(N) = N$.

2.3.2 Test case

A target structure was built starting from the published coordinates of a four-helix bundle *Alpha-1* crystallized in space group P1 with unit cell dimensions $a = 20.846$, $b = 20.909$, $c = 27.057$ Å, $\alpha = 102.40^o$, $\beta = 95.33^o$ and $\gamma = 119.62^o$ (PDB-code: 1BYZ; Privé *et al.*, 1999). All 48 amino acids of this peptide were replaced by alanines and all atomic B-factors were set to 15 Å^2. The structure-factor amplitudes were taken from calculated X-ray data to 2.0 Å resolution.

Two types of starting models were generated for testing purposes. First, scrambled starting models with increasing coordinate errors were made by applying random coordinate shifts of increasing magnitude to all atoms in the unit cell. For these starting structures a minimum inter-atomic distance of 1.4 Å was enforced. Second, random atom distributions were made by randomly placing 264 atoms in the unit cell, while enforcing a minimum inter-atomic distance of 1.8 Å. All atoms in the starting structures were given equal labels and carbon scattering factors were assigned to all of them.

2.3.3 Refinement protocols

The refinement protocols for optimization starting from the scrambled models and random models are given in Figures 2.5 *a* and *b*. These optimization protocols include standard procedures: overall *B*-factor optimization and weight determination for the X-ray restraint followed by maximum likelihood optimization by either energy minimization or dynamics simulation. Table 2.1 contains the set of parameters defining the conditional force field; target values for packing densities and inter-atomic distances were determined from their distributions in several high-resolution structures in the Protein Data Bank. Up to 9 layers of bonded conditions have been defined, corresponding to linear elements up to e.g. $C^{\alpha}(i)$ to $C^{\alpha}(i+3)$. During the optimization, width σ_r of the conditional functions was adjusted according to the estimated coordinate error (ε_r) derived from the estimated σ_A-values: $\sigma_r^{'} = \sigma_r + \varepsilon_r L^{1/2}$. Atomic *B*-factors were assigned using an exponentially decreasing function depending on the number of neighbours N_i^d within a shell d (+σ_d) of 4.3 (+0.7) Å: $B_i = 150\exp(-0.1N_i^d)$, with a minimum value of 15 Å^2. The time step in these calculations was 0.2 fs and during the dynamics calculations the temperature was coupled to a temperature bath (T_{bath} = 300 K).

Two aspects were tested for optimization starting from scrambled models: *i.* the effect of resolution by using data truncated at 3.5, 3.0, 2.5 and 2.0 Å resolution and *ii.* the effect of the number of conditional layers L, three, six or nine. For each test condition three trials were performed using different random starting velocities. A randomly selected 10% of the reflections were excluded from refinement and used for calculation of R_{free} (Brünger, 1993) and cross-validated σ_A-estimates (Read, 1986; Pannu and Read, 1996).

$d+\sigma_d$:	**1.6+0.5 Å**			**2.6+0.7 Å**			**3.6+0.7 Å**			**4.3+0.7 Å**			**5.0+0.7 Å**		
atomtype	$N_{\min}$	$N_{\max}$	σ_N	$N_{\min}$	$N_{\max}$	σ_N	$N_{\min}$	$N_{\max}$	σ_N	$N_{\min}$	$N_{\max}$	σ_N	$N_{\min}$	$N_{\max}$	σ_N
N	1.0	2.0	4.0	6.5	9.5	8.0	10.0	16.0	8.0	10.0	25.0	8.0	10.0	32.0	8.0
CA	3.0	3.0	4.0	6.7	7.1	8.0	8.5	11.5	8.0	10.0	25.0	8.0	15.0	32.0	8.0
C	3.0	3.0	4.0	6.0	8.0	8.0	9.0	15.0	8.0	10.0	25.0	8.0	17.0	33.0	8.0
O	1.0	1.0	2.5	3.5	6.5	8.0	7.0	19.0	8.0	10.0	25.0	8.0	15.0	33.0	8.0
CB	1.0	1.0	2.5	3.0	4.0	8.0	5.0	9.5	8.0	6.5	19.5	8.0	9.0	29.0	8.0

Table 2.1: **Conditional force field for alanines in a helical conformation.** (*a*) *Parameters* $N_{\min}$,$N_{\max}$ *and* σ_N *for atom types N, CA, C, O and CB, defining the atomic conditions for five neighbour shells with different* $d+\sigma_d$ *(see Figure 2.3).*

(*b*) Parameters $r_{\min}$, $r_{\max}$, σ_r and χ_{opt} see Figures 2.2 and 2.4, describing the bonded conditions for all types of linear elements with $L = [1,9]$.

Layer	type (L)	subtype-A ($L-1$)	subtype-B ($L-1$)	$r_{\min}$ [Å]	$r_{\max}$ [Å]	σ_r [Å]	χ_{opt} [^{o}]
***L*=1**	N-CA	N	CA	1.43	1.51	0.05	
	CA-C	CA	C	1.51	1.55	0.05	
	C-O	C	O	1.21	1.27	0.05	
	C-N	C	N	1.31	1.35	0.05	
	CA-CB	CA	CB	1.51	1.57	0.05	
***L*=2**	N-C	N-CA	CA-C	2.41	2.53	0.08	
	CA-O	CA-C	C-O	2.35	2.45	0.08	
	CA-N	CA-C	C-N	2.39	2.49	0.08	
	C-CA	C-N	N-CA	2.39	2.49	0.08	
	O-N	O-C*	C-N	2.21	2.31	0.08	
	O-O	O-C*	C-O	2.10	2.30	0.08	
	N-CB	N-CA	CA-CB	2.39	2.55	0.08	
	CB-C	CB-CA*	CA-C	2.43	2.61	0.08	
***L*=3**	N-O	N-C	CA-O	3.43	3.61	0.15	138
	N-N	N-C	CA-N	2.71	2.93	0.15	-42
	CA-CA	CA-N	C-CA	3.75	3.87	0.15	178
	C-C	C-CA	N-C	2.91	3.15	0.15	-62
	O-CA	O-N	C-CA	2.69	2.85	0.15	-2
	CB-O	CB-C	CA-O	3.15	3.47	0.15	-98
	CB-N	CB-C	CA-N	3.01	3.37	0.15	82
	C-CB	C-CA	N-CB	3.63	3.79	0.15	174
***L*=4**	N-CA	N-N	CA-CA	4.11	4.33	0.20	138
	CA-C	CA-CA	C-C	4.29	4.53	0.20	122
	C-O	C-C	N-O	3.69	4.05	0.20	62
	O-C	O-CA	C-C	2.81	3.15	0.20	-58
	C-N	C-C	N-N	3.13	3.47	0.20	-90
	CA-CB	CA-CA	C-CB	4.77	4.99	0.20	-10
	CB-CA	CB-N	CA-CA	4.31	4.71	0.20	-110
	O-CB	O-CA	C-CB	4.17	4.35	0.20	174
***L*=5**	N-C	N-CA	CA-C	4.59	4.85	0.25	82
	CA-O	CA-C	C-O	5.17	5.51	0.25	142
	O-O	O-C	C-O	3.17	3.71	0.25	14
	CA-N	CA-C	C-N	4.19	4.59	0.25	14
	O-N	O-C	C-N	3.21	3.71	0.25	-114
	C-CA	C-N	N-CA	4.31	4.67	0.25	-6
	N-CB	N-CA	CA-CB	4.81	5.19	0.25	-46

	CB-C	CB-CA	CA-C	5.31	5.61	0.25	-166
	CB-CB	CB-CA	CA-CB	5.13	5.67	0.25	66
***L*=6**	N-O	N-C	CA-O	5.63	5.97	0.30	158
	CA-CA	CA-N	C-CA	5.27	5.69	0.30	78
	N-N	N-C	CA-N	4.13	4.53	0.30	30
	C-C	C-CA	N-C	4.37	4.75	0.30	22
	O-CA	O-N	C-CA	4.11	4.69	0.30	-50
	CB-O	CB-C	CA-O	6.11	6.51	0.30	-86
	CB-N	CB-C	CA-N	5.47	5.81	0.30	146
	C-CB	C-CA	N-CB	4.99	5.53	0.30	-94
***L*=7**	CA-C	CA-CA	C-C	5.25	5.77	0.35	90
	C-N	C-C	N-N	3.63	4.07	0.35	-14
	N-CA	N-N	CA-CA	5.13	5.61	0.35	86
	O-C	O-CA	C-C	3.83	4.39	0.35	-30
	C-O	C-C	N-O	5.43	5.85	0.35	98
	CB-CA	CB-N	CA-CA	6.65	7.05	0.35	-166
	CA-CB	CA-CA	C-CB	5.57	6.27	0.35	-18
	O-CB	O-CA	C-CB	5.05	5.81	0.35	-138
***L*=8**	N-C	N-CA	CA-C	5.43	5.85	0.40	130
	O-O	O-C	C-O	4.73	5.26	0.40	22
	CA-O	CA-C	C-O	6.37	6.91	0.40	130
	CA-N	CA-C	C-N	4.27	4.85	0.40	42
	C-CA	C-N	N-CA	4.33	4.87	0.40	22
	O-N	O-C	C-N	2.99	3.65	0.40	-70
	CB-CB	CB-CA	CA-CB	7.05	7.66	0.40	130
	N-CB	N-CA	CA-CB	5.05	5.77	0.40	22
	CB-C	CB-CA	CA-C	6.71	7.15	0.40	-122
***L*=9**	N-O	N-C	CA-O	6.65	7.09	0.45	166
	CA-CA	CA-N	C-CA	4.85	5.55	0.45	74
	C-C	C-CA	N-C	4.67	5.17	0.45	90
	N-N	N-C	CA-N	4.61	5.07	0.45	110
	O-CA	O-N	C-CA	3.47	4.09	0.45	-38
	CB-O	CB-C	CA-O	7.79	8.27	0.45	-58
	CB-N	CB-C	CA-N	5.71	6.31	0.45	-114
	C-CB	C-CA	N-CB	3.97	4.81	0.45	-22

* For types O-C and CB-CA the same parameters were used as for types C-O and CA-CB, respectively.

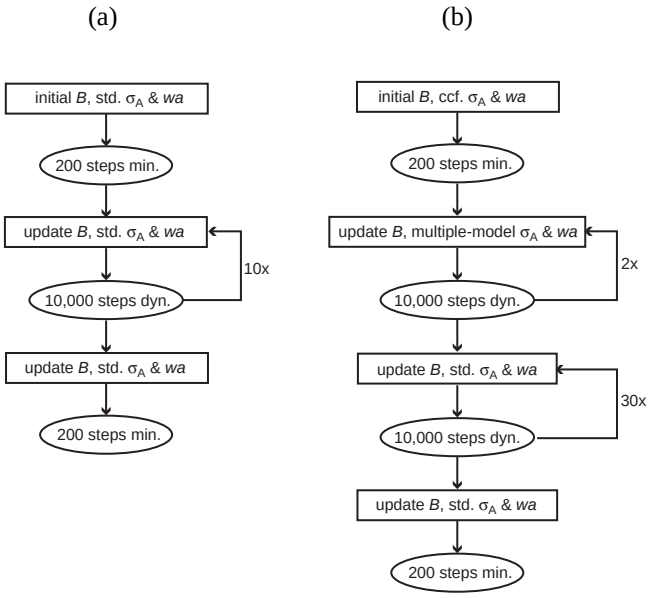

Figure 2.5: *Refinement protocols for a) scrambled models and b) random atom distributions. Conditional energy minimization (min.) and dynamics simulation (dyn.) are alternated with overall isotropic temperature-factor optimization (B), determination of the weight for the X-ray term in the target function (wa) and estimation of σ_A's using the standard SIGMAA procedure (std.), our modified procedure (multiple-model) or correlation coefficients between the observed and calculated normalised structure factors up to 5 Å resolution (ccf.).*

For optimization starting from randomly placed atoms all X-ray data to 2.0 Å resolution were included. Compared to the optimization of scrambled models, three modifications were made: alternative protocols were defined for estimating σ_A-values and for handling the "test set" reflections and to allow faster sampling, T_{bath} was set to 600 K. Standard σ_A-estimates are based on the correlation coefficient between observed and calculated normalized structure factors, $|E^{\mathrm{obs}}|$ and $|E^{\mathrm{calc}}|$ (Read, 1986). For random atom distributions and structures very far away from the correct answer the bin-wise correlation coefficients on normalized structure factors yield spuriously high values. We used a multiple-model approach to obtain estimates of the phase error $\varphi^{\mathrm{obs}} - \varphi^{\mathrm{calc}}$ in the theoretical values for σ_A: $\sigma_A = \langle |E^{\mathrm{obs}}||E^{\mathrm{calc}}| \cos(\varphi^{\mathrm{obs}} - \varphi^{\mathrm{calc}}) \rangle$ (Srinivasan and Parthasaraty, 1976). Starting from the coordinate set corresponding to F^{calc}, four dynamics runs of 1,000 steps each were performed at an elevated temperature of 900 K using different random starting velocities (yielding structure factors sets F^i). From the resulting four models, we compute the average structure factor F^{ave} and figure-of-merit m^{ave} ($m^{\mathrm{ave}} = |F^{\mathrm{ave}}| / \langle |F^i| \rangle$). By rewriting $(\varphi^{\mathrm{obs}} - \varphi^{\mathrm{calc}}) = (\varphi^{\mathrm{obs}} - \varphi^{\mathrm{ave}}) + (\varphi^{\mathrm{ave}} - \varphi^{\mathrm{calc}})$ and assuming $(\varphi^{\mathrm{obs}} - \varphi^{\mathrm{ave}}) \approx m^{\mathrm{ave}}$ we can estimate σ_A. For a range of test structures far away from the known answer these estimates had a reasonable correlation to the theoretical values as calculated using known phases φ^{obs} of the test cases. The second feature deviating from normal crystallographic refinement pro-

tocols was the handling of the test set reflections. A conventional test set comprising 7% of all reflections was used to calculate R_{free} and to estimate cross-validated σ_A-values according to Pannu and Read (1996) in the later stages of refinement. Additionally, another 7% of the reflections were taken out of the refinement. After every 1,000 steps, the selection of these 7% was modified. As a result, the reflections used in the crystallographic target function changed every 1,000 steps, resulting in a "tacking" behaviour during refinement minimizing the chance of stalled progress due to local minima in the crystallographic target function.

Calculations were performed on a Compaq XP1000 workstation with 256 Mb of computer memory and a single 667 MHz processor. The CPU-time needed was about 4 hours for 100,000 steps of optimization.

2.4 Results

2.4.1 Refinement of scrambled models

Six scrambled models with coordinate errors of 1.0, 1.2, 1.4, 1.6, 1.8 and 2.0 Å root mean square deviation (r.m.s.d.) respectively were generated. The dependence of the method on the number of conditional layers was tested performing a series of refinements using three, six or nine layers. The resulting amplitude-weighted phase errors are shown in figure 2.6. Three layers of conditions are not enough to give significant phase improvement. Using six layers, scrambled models with r.m.s.d.'s up to 1.4 Å could be improved significantly. Adding another three layers of conditions led to a small increase in the success rate. Figure 2.7 shows the phase improvement for the refined 1.4 Å r.m.s.d. structure with the lowest free *R*-factor, using three, six or nine layers of conditions. Figure 2.8 shows an initial model with a coordinate error of 1.4 Å r.m.s.d and the refined structure with the lowest free *R*-factor using nine layers of conditions. This structure is representative for all successful runs: the four helices are clearly visible although some are not completed, contain breaks in the main chain or the N-C direction is reversed. For structures with a coordinate error larger than 1.4 Å r.m.s.d., refinement did not yield improvement of the phases. This coincides with the observation that for models with large errors, the *SIGMAA* procedure (Pannu and Read, 1996) gave spurious estimates for the σ_A-values (results not shown).

The dependence on the high-resolution limit of the diffraction data was tested by refining the 1.0 Å r.m.s.d. model using data truncated at various resolution limits. Calculations were performed using six or nine layers of conditions. The resulting phase improvements are shown in figure 2.9. All runs using data to a resolution of 3.0 Å were successful. When using only 3.5 Å data all three runs using six layers of conditions failed, while using nine layers of conditions resulted in a success rate of two out of three.

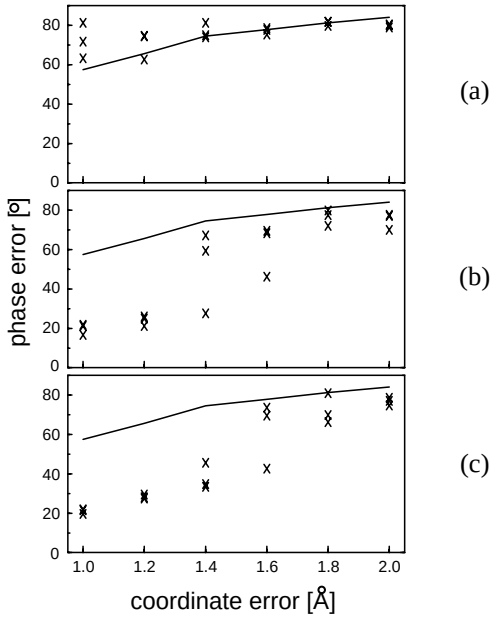

Figure 2.6: *Optimizations of scrambled models with different initial coordinate errors against 2.0 Å resolution diffraction data. Overall amplitude-weighted phase errors are shown for the starting models (solid lines) and the refined structures (crosses) using a) three, b) six and c) nine layers of conditions, where each run was performed three times starting from different random velocities.*

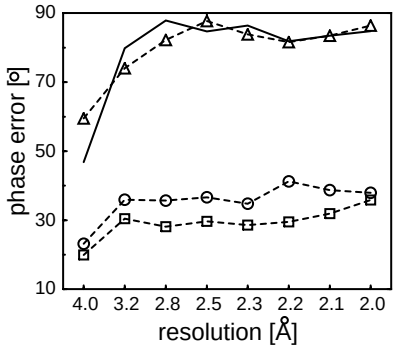

Figure 2.7: *Optimizations of a scrambled model with an initial coordinate error of 1.4 Å r.m.s.d. against 2.0 Å resolution diffraction data. Amplitude-weighted phase errors per resolution shell are shown for the initial model (solid line) and the refined models (dashed lines) using three (triangles), six (squares) and nine (circles) layers of conditions, corresponding to the runs with the lowest overall amplitude-weighted phase error in Figure 2.6.*

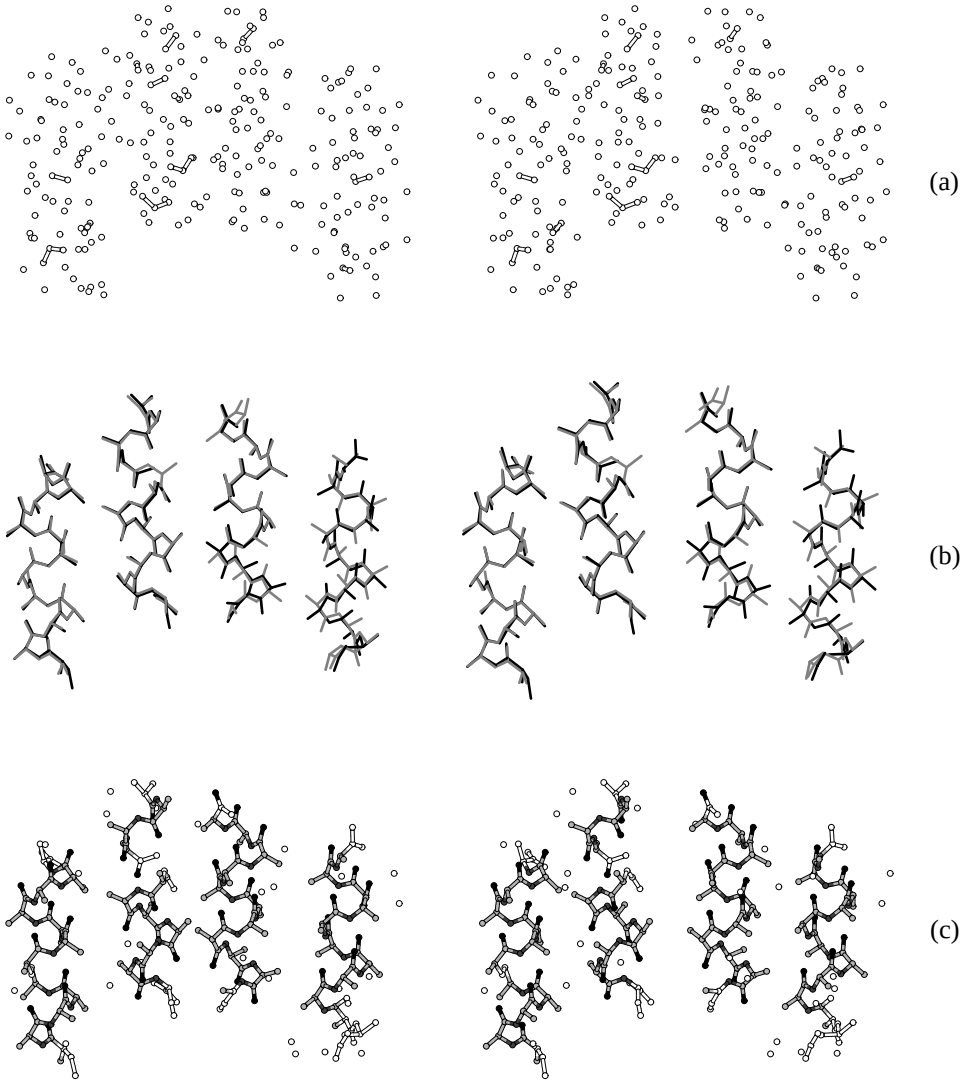

Figure 2.8: *Stereo views of a) the initial, scrambled model with a coordinate error of 1.4 Å r.m.s.d. b) its refined structure superimposed on the target structure and c) the same structure in ball-and-stick representation with automatic assignment of atom types based on the scores of joint conditions (white = unassigned, light grey = carbon, dark grey = nitrogen and black = oxygen). Atoms within 1.8 Å inter-atomic distance are connected.*

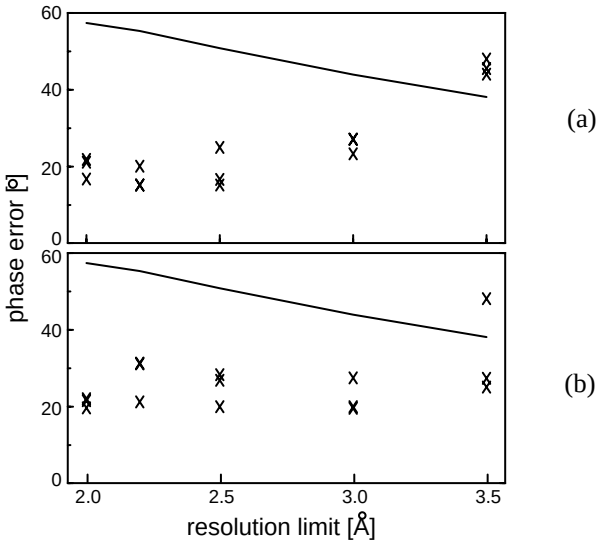

Figure 2.9: *Optimizations of a scrambled model with an initial coordinate error of 1.0 Å r.m.s.d. against diffraction data with different high-resolution limits. The overall amplitude-weighted phase errors are shown for the initial model (solid lines) and the refined structures (crosses) using a) six and b) nine layers of conditions, where each run was performed three times starting from different random velocities.*

2.4.2 Refinement of random atom distributions

Sixteen different random atom distributions were refined according to the protocol in figure 2.5b. One run was abandoned, because standard σ_A-estimates could not be obtained by the *SIGMAA*-procedure after the initial 20,000 steps. Of the remaining fifteen models, six yielded a final amplitude-weighted phase error of smaller than 50^o for data up to 2.0 Å resolution. This corresponds to a success rate of one out of three. For these successful runs a condensation into four rod-like structures was observed during the initial stages of the refinement process, thereby establishing a choice of origin for the triclinic cell. Subsequent dynamics optimization lead to the formation of helical fragments that were expanded into near-complete α-helices. Figure 2.10 shows a clear correlation between the phase errors and the overall free *R*-factor obtained for the final models. The structure with the lowest free *R*-factor is shown in figure 2.11[2].This structure clearly shows the four α-helices and resembles the results obtained from the refinement of the scrambled models. The errors in the model include chain breaks, incomplete helices and chain reversals.

[2]A movie, showing the formation of the four helices starting from a random atom distribution is available on: http://journals.iucr.org

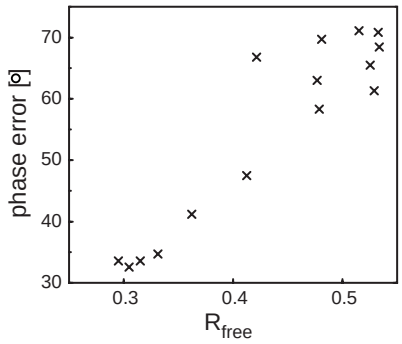

Figure 2.10: *Scatter plot of the amplitude-weighted phase error vs. the free R-factor for the fifteen final models that were obtained starting from random atom distributions.*

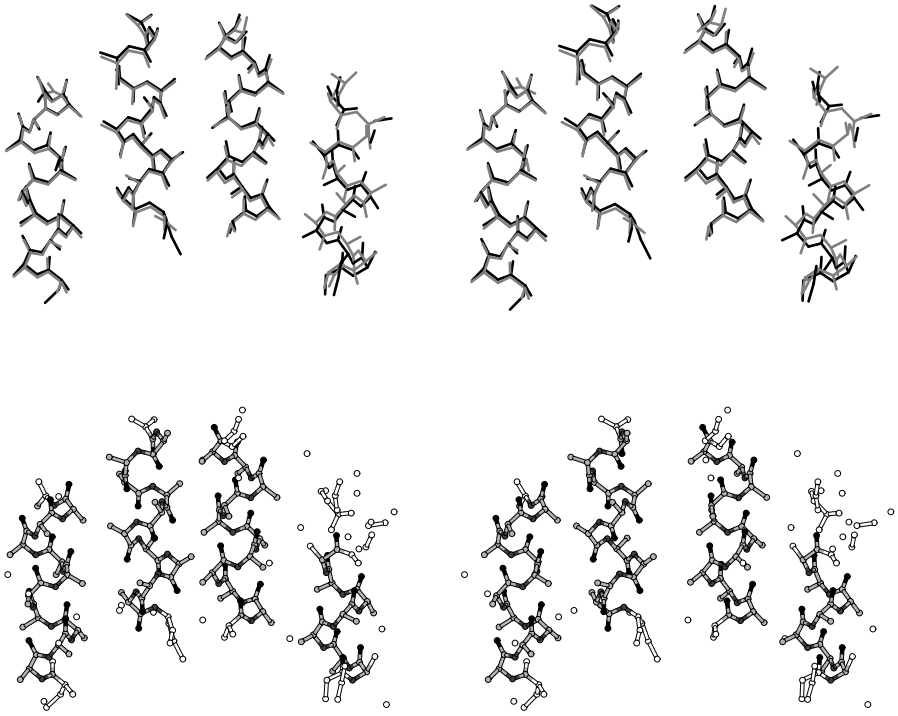

Figure 2.11: *Stereo views of a) a successfully refined structure starting from a random atom distribution superimposed on the target structure and b) the same structure in ball-and-stick representation with automatic assignment of atom types based on the scores of joint conditions (white = unassigned, light gray = carbon, dark gray = nitrogen and black = oxygen). Atoms within 1.8 Å inter-atomic distance are connected.*

2.5 Discussion

We introduced a new method for optimization of protein structures that overcomes the necessity of a fixed topology for defining geometrical restraints. This *N*-particle approach offers a 'restrained topology', where weighted gradients over all possible assignments are applied to loose atoms. We tested this method using calculated data and a very simple test case consisting of four poly-alanine helices with 244 non-hydrogen atoms in total. Optimizations starting from scrambled models show that the method works successfully with diffraction data of at least 3.0 to 3.5 Å resolution and with six or nine layers of conditions, corresponding to linear structural elements of the length of two and three peptide planes, respectively. Moreover, we have shown that our test structure can be optimized successfully starting from randomly distributed atoms when using 2.0 Å resolution diffraction data. Important for successful optimization of random starting models was estimation of reasonable σ_A-values for very bad models using a multiple-model procedure. For trials with different random starts a success rate of one out of three was observed. The free *R*-factor readily distinguished correct solutions from false ones. To our knowledge, we have presented the first method that in principle allows an *ab initio* optimization of atomic models under conditions relevant for protein crystallography (*i.e.* at medium resolution).

In our experiments we used, however, calculated data without a bulk solvent contribution and a small and very simple test case. Calculations against real protein diffraction data will require a model for the bulk solvent and the conditional force field will have to be expanded to target functions that also include the structurally more variable β-sheets, loop regions and side chains. In analogy with the hybrid model of the *ARP/wARP* program, constrained assignments of recognizable structural elements may be included in the optimization process in order to improve the rate of convergence by e.g. correcting errors like chain breaks and reversals. The efficiency of our approach for larger and more complex systems will have to be demonstrated. Due to the possibility to use prior information extensively, conditional optimization may offer a powerful alternative for phase improvement, both when initial phase estimates are available and in *ab initio* structure determination.

Acknowledgements

We gratefully thank Drs. Bouke van Eijck, Jan Kroon (deceased, 3 May 2001), Wijnand Mooij and Titia Sixma for stimulating discussions. We also thank Drs. Alexandre Bonvin and Bouke van Eijck for carefully reading this manuscript. This work is supported by the Netherlands Organization for Scientific Research (NWO-CW: Jonge Chemici 99-564).

Chapter 3

Development of a force field for conditional optimization of protein structures [1]

Abstract

Conditional optimization allows the incorporation of extensive geometrical information in protein structure refinement, without the requirement of an explicit chemical assignment of the individual atoms. Here, a potential of mean force for the conditional optimization of protein structures is presented that expresses knowledge about common protein conformations in terms of inter-atomic distances, torsion angles and numbers of neighbouring atoms. Information is included for protein fragments up to several residues long in α-helical, β-strand and loop conformations, comprising the main chain and side chains up to the γ-position in three distinct rotamers. Using this parameter set, conditional optimization of three small protein structures against 2.0 Å observed diffraction data shows a large radius of convergence, validating the presented force field and illustrating the feasibility of the approach. The generally applicable force field allows the development of novel phase improvement procedures using the conditional optimization technique.

[1] Sjors H.W. Scheres & Piet Gros (2003) *Acta Cryst. D***59**, 438-446

3.1 Introduction

During the standard crystallographic diffraction experiment, information about the phases of the observed reflections gets lost. In order to obtain a molecular model describing the crystal content, this information must be regained. In protein crystallography, this process typically has been divided into well-separated steps: phase determination by experimental methods or molecular replacement, phase extension by density modification and iterative cycles of model building and refinement. Nowadays it is realized that these steps are coupled more tightly than thought before (Lamzin *et al.*, 2000) and programs have been developed that link these steps in an automated way. For example, the *(RE-)SOLVE* package (Terwilliger, 1999) links structure solution, density modification and model building and the *ARP/wARP* program (Perrakis *et al.*, 1999) links density modification, model building and refinement. Due to the typically low observation-to-parameter ratio in protein crystallography, the incorporation of additional information in this process is critical. We have presented a method, called conditional optimization, in which extensive prior stereo-chemical information may be formulated in terms of loose atoms (Scheres & Gros, 2001). With initial, simplified test calculations we showed that a structure can be obtained using 2.0 Å diffraction data without any prior phase information by this approach. Thus, in principle the entire process from phasing to refinement can be expressed in a single step. However, these tests were performed with calculated diffraction data of a highly simplified structure of four poly-alanine α-helices, which can be described by a very limited parameter set defining the expected geometries. Here, we present a parameter set for conditional optimization of the far more complex structures that are protein molecules.

In the conditional formalism, we express geometrical knowledge by the definition of interaction functions, termed conditions. These conditions depend on expected numbers of neighbouring atoms, inter-atomic distances and torsion angles within protein molecules. Conditions are continuous functions ranging from zero to one, and show similarities with the knowledge-based interaction functions as defined by Sippl (1995). Conformations of protein fragments up to several residues long are described by joint conditions, which are products of conditions describing a set of geometrical features of a protein fragment. In principle, (joint) conditions could be defined for all possible conformations in protein molecules, but this would require a vast amount of interaction functions exceeding available computing power. Therefore, we have defined conditions describing the most common conformations observed in the protein structural database (PDB, Berman *et al.*, 2002) for main-chain atoms and side-chain atoms up to the γ-position.

With the defined parameter set, we show that a large radius of convergence can be obtained for conditional optimization of three small protein structures against 2.0 Å observed diffraction data.

3.2 Mean-force potential for protein structures

3.2.1 Brief review of the conditional formalism

In conditional optimization, we express prior knowledge about protein structures without explicitly assigning chemical identities to the atoms. Instead, we take all possible assign-

ments into account by using an N-particle approach. We define conditions $C = [0,1]$, which are continuous interaction functions based on optimal values for the inter-atomic distances, torsion angles and numbers of neighbouring atoms in protein structures. We describe protein structures as a collection of linear elements (of length L), which are non-branched sequences of $L+1$ atoms. Figure 3.1 shows a common fragment present in protein structures and a schematic representation of the conditions that describe a linear element *N-CA-C-N-CA*, which depend on the number of neighbouring atoms per atom, inter-atomic distances and torsion angles.

As discussed in more detail before (Scheres & Gros 2001), a linear element of length L is composed of in total $L(L+1)/2$ linear (sub-)elements of length $l \leq L$. Multiplication of all conditions corresponding to these (sub-)elements, gives the so-called joint condition *JC*.

For a linear combination of $L+1$ atoms i, j, $\ldots$, p and q, the joint condition $JC^{\text{type}}_{ij\ldots pq}$ describes to what extent the conformation of the atoms resembles a defined target conformation of a particular type of linear element. A minor change was made to the conditional formalism as presented before. Originally, joint conditions were defined as binomial multiplications of the individual conditions, according to the binary combination of all (sub-) elements. In the current implementation, the resulting higher powers of individual conditions in the expression of joint conditions have been removed, and joint conditions are defined as the multiplication of all corresponding individual conditions. Consequently, the factor n in the calculation of derivatives (see formulas 2.6 and 2.6 in chapter 2) reduces to one. In figure 3.1 the atoms i, j, k and l resemble a linear element of type *N-CA-C-N*. The corresponding joint condition for *N-CA-C-N* is determined by the individual conditions as given in (3.1):

$$\begin{aligned} JC^{N-CA-C-N}_{ijkl} = \quad & C^{N}_{nb}(n_i)C^{CA}_{nb}(n_j)C^{C}_{nb}(n_k)C^{N}_{nb}(n_l)C^{N-CA}(r_{ij}) \\ & \times C^{CA-C}(r_{jk})C^{C-N}(r_{kl})C^{N-CA-C}(r_{ik}) \\ & \times C^{CA-C-N}(r_{jl})C^{N-CA-C-N}(r_{il})C^{N-CA-C-N}_{\chi}(\chi_{ijkl}) \end{aligned} \tag{3.1}$$

The function $JC^{N-CA-C-N}_{ijkl}$ will take on the value one, when the configuration of atoms i, j, k and l, with numbers of neighbouring atoms n, inter-atomic distances r and dihedral angle χ, matches all individual conditions. This implies that these atoms have adopted a *N-CA-C-N* conformation.

A protein structure can be described by the sum of its linear elements. Therefore, we define a least-squares target function, as given in (3.2), that depends on the expected number of conformations present in the target structure.

$$E = \sum_{\text{type}} E^{\text{type}} = \sum_{\text{type}} w^{\text{type}} \left(TC^{\text{type}} - \sum_{ij\ldots pq} JC^{\text{type}}_{ij\ldots pq} \right)^2 \tag{3.2}$$

where, TC^{type} is the expected sum of joint conditions for the types of linear elements in the target structure, and w^{type} is a weighting factor. The first summation runs over all types of linear elements (of various lengths L^{type}) that have been defined. The second summation runs over all possible combinations of $L^{\text{type}}+1$ atoms $ij\ldots pq$. The minimum of this target function corresponds to a set of atoms with the expected number of linear elements in their expected types of conformations. Derivatives of this target function can be calculated with

respect to all atomic coordinates. This allows the application of gradient-driven optimization techniques, which we termed conditional optimization.

3.2.2 The general parameter set

For the potential of mean force, we defined 18 atom types ($L = 0$) described by the expected numbers of neighbouring atoms within four neighbour shells with increasing radii (corresponding to typical distances of respectively bonds, angles, torsion-angles and Lennard-Jones interactions). We did not take into account glycine C^{α} and proline N, which have deviating numbers of nearest neighbouring atoms. By combination of the atom types, we defined 26 bond types ($L = 1$) and these combine to form 43 types of angles ($L = 2$). For longer fragments ($L > 2$) separate conformations observed in α-helices, β-strands and loops were defined. For α-helices we defined conditions up to $L = 12$, for β-strands up to $L = 9$ and for loops up to $L = 7$ taking into account the structural variability of the secondary structure elements. For loops, separate conditions were defined for conformations corresponding to the A and B-region of the Ramachandran plot, but conformations corresponding to the L-region were not taken into account. For linear elements comprising two subsequent loop residues, separate conditions were defined for the possible combinations of φ/ψ-angle rotations AA, AB, BA and BB. For side-chain atoms up to the γ-position we defined conditions according to the three preferred χ_1-rotamer conformations; in α-helices only the two commonly observed χ_1-rotamers were defined. No distinction was made between the atoms at the γ-position of different amino acids except for the cysteine S^{γ}-atom; consequently, C^{γ} or O^{γ}-atoms were treated equally. Side chain atoms beyond the γ-position were only defined up to $L = 2$, omitting information with respect to their rotamer conformations, which drastically reduced the number of possible combinations of defined conformations.

To determine minimum and maximum values for the condition parameters, distributions of observed numbers of neighbouring atoms, inter-atomic distances and torsion angles were calculated for the high-resolution protein structures in the *SCAN3D* database of the *WHATIF* program (Vriend *et al.*, 1994). The observed numbers of neighbouring atoms were calculated for twenty protein structures in this data base, comprising in total approximately 24,000 protein atoms. Observed inter-atomic distances and torsion angles were calculated from oligo-peptides that were extracted using the *SCAN3D* structural annotation. Oligo-peptides in a helical conformation were extracted as seven subsequent residues with an H (helix) assignment, β-strands were extracted as five subsequent residues with an S (strand) assignment and loops as five subsequent residues, with a T (turn) or C (coil) assignment for the middle three residues. Backbone conformations with annotated torsion angles $-180^{\circ} < \varphi < 0^{\circ}$ and $-110^{\circ} < \psi < 50^{\circ}$ were termed A and conformations with $-180^{\circ} < \varphi < 0^{\circ}$ and $50^{\circ} < \psi < 180^{\circ}$ were termed B. Only β-strands with five subsequent residues in the B-conformation were taken into account. For the middle residue of the extracted oligo-peptides a distinction between the three χ_1-rotamers g^-, t and g^+ was made based on its value as annotated in the database: respectively: $-120^{\circ} < \chi_1 < 0^{\circ}$, $120^{\circ} < \chi_1 < 240^{\circ}$ and $0^{\circ} < \chi_1 < 120^{\circ}$. Table 3.1 shows the total numbers of extracted oligo-peptides in the different conformations that were used to determine the corresponding condition parameters.

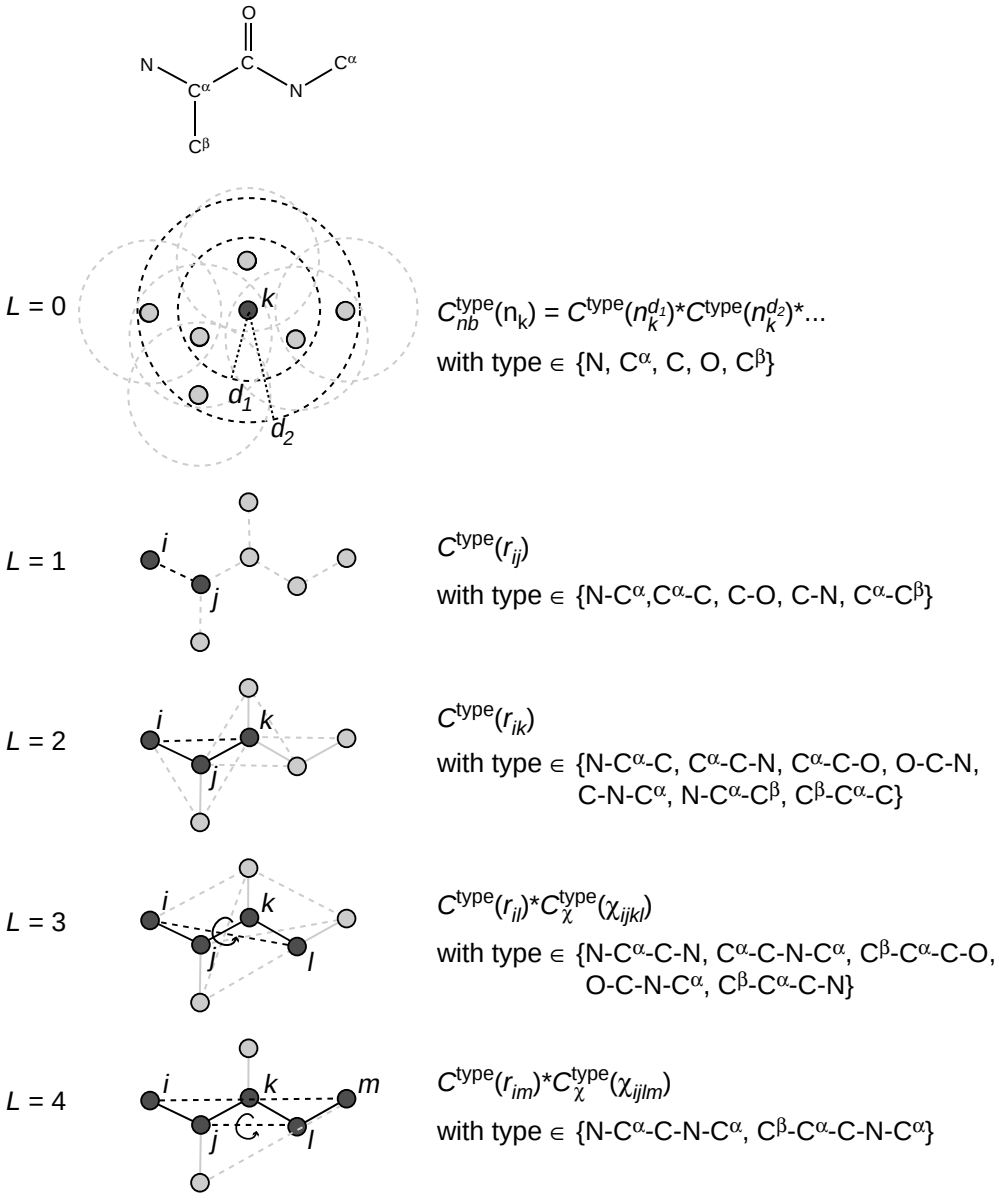

Figure 3.1: *Schematic representation of the conditions defining a linear element N-CA-C-N-CA. Shown are a protein fragment containing this linear element and schematic representations of the conditions involved. These conditions depend on number of neighbouring atoms n (for two shells with radii d_1 and d_2), inter-atomic distances r and torsion angles χ. The interactions present in this fragment are shown in dashed lines for $L = [0,4]$. For convenience bonds are shown in solid lines. For each layer L a single example is highlighted in black and conditions applicable are given on the right-hand side.*

Secondary structure	No. of peptides	$g^-(\chi_1)$	$t(\chi_1)$	$g^+(\chi_1)$
α	1999	910	609	0
β	2332	883	721	285
l^{AA}	1590	515	190	312
l^{AB}	2180	924	199	457
l^{BA}	1822	541	613	191
l^{BB}	2562	930	546	440

Table 3.1: *Number of penta and hepta-peptide configurations used for defining the force field parameters. Shown are the number of hepta-peptide configurations extracted in α-helical (α) conformation, the number of penta-peptides for β-strand (β) and loop (l^{AA}, l^{AB}, l^{BA} and l^{BB}) conformations and the number of χ_1-rotamer conformations, g^-, t or g^+, observed for the middle residues of the penta and hepta-peptides.*

For each condition type, the minimum and maximum values of the condition parameter were set so to comprise 90% of the conformations as observed in the *SCAN3D* database. Histograms were made of the observed numbers of neighbouring atoms, inter-atomic distances or torsion angles. Bin widths were chosen such that the top of each histogram reached at least 50 hits, except for distributions with less than 200 hits, where the top should reach at least 20 hits. For each histogram a frequency cut-off value was chosen such that 90% of all hits lie within the interval ranging from the first to the last bin for which the number of hits exceeds this cut-off value. For this interval condition C corresponds to one. The widths of the slopes (see figure 3.2) were set to 0.05 Å at layer $L = 1$ up to 0.75 Å at layer $L = 12$ for distance conditions; widths of neighbour conditions were set to respectively 1.5, 4.8, 12.7 and 26.7 neighbouring atoms for the four shells with increasing radii; widths of torsion-angle conditions were set to $(360^\circ - \chi_{\max} + \chi_{\min})/2$, thus providing a continuous function for the entire range of torsion angles. As an example, figure 3.2 shows the histograms of observed distances and torsion angles and the resulting conditions for a linear element of $C^\alpha(i)$ to $C^\alpha(i+4)$ in an α-helical conformation.

The complete conditional parameter set that was obtained as described above has been submitted as supplementary material and is available from the IUCr electronic archive. A summary of the numbers of all defined conditions is given in table 3.2.

3.2.3 Protein-specific force fields

The force field parameters as defined in the previous section represent geometric expectations of common conformations as observed in many protein structures. To define the expectations for a specific protein, a sub-set is extracted from this general parameter set, specific for that particular protein. Based on the known amino acid sequence and estimated fractions of α-helical, β-strand and loop content, occurrences of all types of linear elements are determined and used to calculate expected sums of joint conditions TC^{type}. In this calculation, we also take into account contributions from reminiscent conformations that give non-zero values for JC^{type}. For differentiation of loops into A and B-conformations and differentiation of χ_1-rotamers, the expected fractions are set to the observed relative occurrences of these conformations in the *SCAN3D* database. The target functions corresponding to these types are

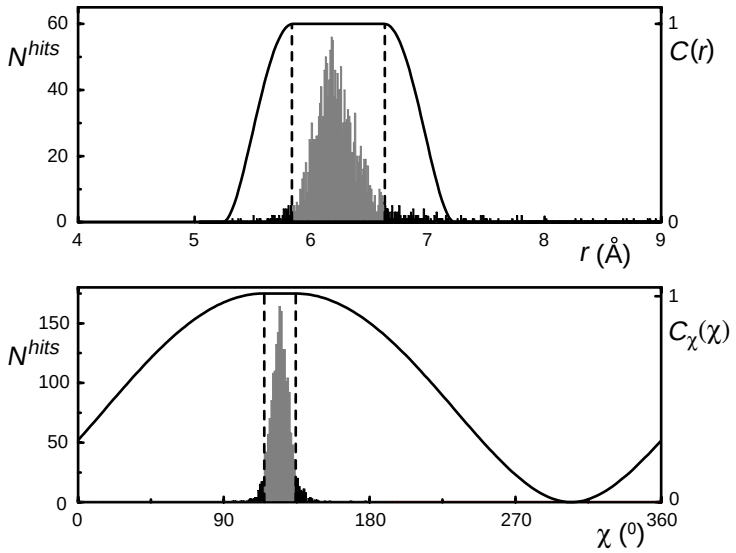

Figure 3.2: *Observed distributions N_{hits} and defined conditions C for (a) inter-atomic distance r and (b) C_χ for torsion angle χ between the outermost atoms of a linear element comprising atoms $C^\alpha(i)$ until $C^\alpha(i+4)$ in an α-helical conformation. Minimum and maximum values for which $C = 1$ are set to comprise 90% of the observed conformations (grey).*

grouped (3.3):

$$E^{\text{group}} = \Big(\sum_{\text{group}} g^{\text{type}} \big(TC^{\text{type}} - \sum_{ij\ldots pq} JC^{\text{type}}_{ij\ldots pq} \big) \Big)^2 \tag{3.3}$$

where, g^{type} (with $\sum_{\text{group}} g^{\text{type}} = 1$) corresponds to the relative occurrence of each group member and the summation runs over all types that are part of the group.

3.3 Experimental

Three small protein structures were selected for testing purposes: human hyperplastic discs protein (PDB-code: 1I2T), erabutoxin (PDB-code: 3EBX) and turkey ovomucoid third domain (PDB-code: 1DS3); see table 3.3. These represent examples of an all-α helical, an all-β sheet and a mixed α/β-fold, respectively. Published diffraction data sets were truncated at 2.0 Å resolution. All three data sets were nearly complete up to this resolution limit. For the ovomucoid third domain test case, five of the lowest resolution reflections were marked as probable measurement errors and these reflections were removed from the reflection file. For these reflections an almost-zero intensity was observed, while their calculated intensities were significantly higher.

Two aspects of conditional dynamics using the presented force field were tested: the stability of structures when starting with correct coordinates and the optimization behaviour

Layer	No. of different topologies	secondary structure differentiation	χ_1-rotamer differentiation	No. of conditions
$L=0$	18	-	-	72
$L=1$	26	-	-	26
$L=2$	42	-	-	42
	1	α,β,l	-	3
$L=3$	2	-	-	2
	2	α,β,l	-	6
	4	α,β,l^A,l^B	-	16
	4	-	g^-,t,g^+	12
$L=4$	4	α,β,l	-	12
	4	α,β,l^A,l^B	-	16
	3	α,β,l^A,l^B	g^-,t,g^{+*}	33
$L=5$	9	α,β,l^A,l^B	-	36
	3	α,β,l^A,l^B	g^-,t,g^{+*}	33
$L=6$	4	α,β,l^A,l^B	-	16
	4	α,β,l^{AA},l^{AB},l^{BA},l^{BB}	-	24
	4	α,β,l^A,l^B	g^-,t,g^{+*}	44
$L=7$	4	α,β,l^A,l^B	-	16
	4	α,β,l^{AA},l^{AB},l^{BA},l^{BB}	-	24
	1	α,β,l^A,l^B	g^-,t,g^{+*}	11
	2	α,β,l^{AA},l^{AB},l^{BA},l^{BB}	g^-,t,g^{+*}	34
$L=8$	9	α,β	-	18
	3	α,β	g^-,t,g^{+*}	15
$L=9$	8	α,β	-	16
	4	α,β	g^-,t,g^{+*}	20
$L=10$	8	α	-	8
	3	α	g^-,t	6
$L=11$	9	α	-	9
	3	α	g^-,t	6
$L=12$	8	α	-	8
	4	α	g^-,t	8
Total				592

Table 3.2: *Number of conditions defined in the general parameter set for conditional optimization. Conditions are defined for linear elements of different length (L) and different chemical topologies. In addition, the defined conditions differentiate between distinct conformations of secondary structure elements,* α, β, l^{AA}, l^{AB}, l^{BA} *and* l^{BB}, *and* χ_1*-rotamer conformations* g^-, t *and* g^+.
* *For helices only conditions for* χ_1*-rotamer* g^- *and* t *were defined.*

for structures away from the correct answer. To test the stability of structures corresponding to the correct answer, equilibrium runs were started from the deposited protein coordinates. These optimizations comprised 5,000 steps of dynamics preceded and followed by

PDB-code	No. of atoms: protein/total	2° structure content	space group	$d_{\min}$ (Å)	No. reflections* ($d > 2$Å)
1I2T	472/602	α	$P2_12_12_1$	1.04	4662 (7)
3EBX	475/590	β/loop	$P2_12_12_1$	1.4	3690 (0)
1DS3	378/426	α/β/loop	$P2_1$	1.65	2938 (13)

Table 3.3: *Characteristics of the three test cases, human hyperplastic discs protein (PDB-code 1I2T), erabutoxin (PDB-code 3EBX) and turkey ovomucoid third domain (PDB-code 1DS3).*
** The number of missing reflections is given between brackets.*

200 steps of minimization using conditional optimisation implemented in *CNS* (Brünger *et al.*, 1998). We used the maximum-likelihood crystallographic target function (MLF; Pannu & Read, 1996) with σ_A-values estimated by Read's procedure [Read, 1986] based on 10% of free reflections (Brünger, 1993). Reflections for cross validation were selected randomly from reflections with a Bragg spacing $d < 10$ Å. To test the optimization behaviour for structures away from the correct answer, optimization runs were performed starting from scrambled models with a root-mean-square (r.m.s.) coordinate error of 1.5 Å. For each test case twelve different starting models were generated by applying random coordinate shifts to all protein atoms. The scrambled models were refined according to the protocol as shown in figure 3.3. For each cycle of phase-restrained maximum-likelihood refinement (MLHL; Pannu *et al.*, 1998), target phases were obtained from the average structure factor F^{ave} of all twelve individual structure factor sets F^i. (In the presented test cases, averaging the structure factors of the twelve starting models yielded phase errors of $\sim 70^\circ$ for data up to 2 Å resolution. Phase errors of similar magnitude would result from a single model with an r.m.s. random coordinate error of $\sim$1.1 Å.) Resolution-dependent figures of merit were calculated from the reflections in the test set as $m'_a = \sum_{i=1}^{N} F^i / \sum_{i=1}^{N} |F^i|$ and extrapolated to $N \to \infty : m_a = \sqrt{(N(m'_a)^2 - 1)/(N-1)}$. σ_A-Estimates were calculated from these cross-validated figures of merit m_a, because the standard routine to estimate σ_A-values gave spurious results for these structures with large errors and small numbers of reflections in the test set. Values for weights *wa* on the X-ray restraint as determined with standard routines showed a strong variation over the twelve different structures. One common value for each cycle was determined by exploiting a relationship with the sum of figure of merit over all reflections ($wa \propto 1/\Sigma m$), as observed during initial calculations with models of varying quality (results not shown).

For each test case equal atom labels, 'X', were given to all protein atoms and carbon scattering factors were assigned to all of them. Water and other non-protein atoms were not included in the calculations. Atomic *B*-factors were assigned based on the number of neighbouring atoms as described before (Scheres & Gros, 2001). Standard routines were used for scaling and bulk solvent correction. To avoid negative atomic *B*-factors after scaling, the inverse scaling was applied to $|F^{\text{obs}}|$ rather than scaling $|F^{\text{calc}}|$. Dynamics calculations were performed with a time step of 0.2 fs and the temperature was coupled to a bath of 600K. All calculations were performed on four, 667 MHz single-processor Compaq XP1000 workstations with at least 1.2 Gb of computer memory.

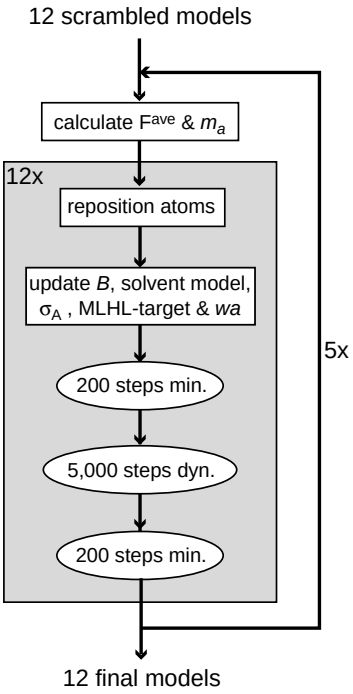

Figure 3.3: *Refinement protocol for the optimization starting from twelve scrambled structures. Prior to every optimization cycle, average structure factor* F^{ave} *and figures of merit* m_a *were calculated from the 12 individual structure factor sets* F^i*. At every cycle a small amount of atoms was repositioned for each structure, based on its* $m_a|F^{\mathrm{obs}}|\exp(i\varphi^{\mathrm{ave}}) - D|F^{\mathrm{calc}}|\exp(i\varphi^{\mathrm{calc}})$ *difference map. All atoms at density levels lower than -2.5*σ *and their neighbouring atoms (within 1.8 Å distance) with density lower than -1.5*σ *were selected for repositioning. These atoms were repositioned at the highest positive peaks of the difference map, with a minimum inter-atomic distance constraint of 1.2 Å and a triangulation constraint prohibiting the formation of a triangle of three bonded atoms. For each model, overall isotropic B-factor optimization, bulk solvent correction, estimation of* σ_A*-values based on figures of merit* m_a *and calculation of weight wa on the X-ray term of the target function (MLHL) were performed. Every optimization cycle comprised 5,000 steps of dynamics calculations (dyn.) preceded and followed by 200 steps of energy minimization (min.) for each of the twelve structures.*

3.4 Results & discussion

3.4.1 Stability of correct structures

The first evaluation of the defined force field concerns the stability of correct protein structures in conditional optimization. Figure 3.4 shows equilibrated structures after dynamics calculations started from the deposited coordinates of all three test cases. The mean phase errors of these structures increase from $< 20^\circ$ up to $\sim 30^\circ$ (see table 3.4), but still the corresponding electron-density maps are easily interpretable. Errors that are introduced during these runs can be attributed to conformations for which no or limited conditions were de-

fined. In the all-α case a single main-chain break occurs in a turn next to a proline residue. The all-β case shows three main-chain breaks that concern two residues with a conformation in the L-region and one glycine in a conformation outside any of the three common regions of the Ramachandran plot. For the mixed α/β case also three main-chain breaks are observed related to conformations outside the A and B-region of the Ramachandran plot. For all three test cases side chains beyond the γ-position are unstable and atoms at the δ, ε, ζ and η positions of the side chains are displaced from their correct positions during equilibration. Since unstable parts in the protein structures coincide with conformations that were poorly or not defined, extension of the parameter set to describe these conformations may lead to better modelling of the target structure at the expense of more computing power.

test case	protein-specific force field	Δφ (°) before equilibration	Δφ (°) after equilibration	Δφ (°) before optimization	Δφ (°) after optimization	CPU-time (h)
1I2T	100% α	19	27	71	28	10
3EBX	100% β	19	32	70	45	12
1DS3	25% α, 25% β, 50% loop	14	27	71	45	18

Table 3.4: *Results from equilibrium and optimization runs using the presented force field for conditional optimization. For each of the three test cases human hyperplastic discs protein (PDB-code 1I2T), erabutoxin (PDB-code 3EBX) and turkey ovomucoid third domain (PDB-code 1DS3), the secondary-structure content, α-helix, β-sheet and loop, used to define the protein-specific force fields are given in percentages. Amplitude-weighted ($|F^{obs}|$) mean phase errors before and after the equilibrium and optimization runs are given. Phase errors are calculated with respect to phases of the structures deposited in the PDB. In the case of optimization starting from 12 models, phase errors are given for the averaged structure factors. CPU-times are given that were required for each of the 12 models in these optimization runs.*

3.4.2 Searching behaviour in optimization

A second requirement for the presented force field is a favourable searching behaviour in the optimization of structures (far) away from the defined minimum. Optimization runs were performed for all three test cases, starting from twelve scrambled structures with coordinate errors of 1.5 Å r.m.s.d. Figures 3.5, 3.6 and 3.7 show optimized structures and map improvements for respectively the all-α, all-β and mixed α/β test cases. Corresponding phase improvements and CPU-times required for these runs are given in table 3.4. For the all-α test case, optimization converges readily towards the global minimum. Subsequent refinement cycles yield significant improvement of the electron-density map and phase information over the whole resolution range. Errors in the optimized structures coincide with conformations that were also unstable during equilibration. For the all-β and mixed α/β cases, optimization converges less readily, but still considerable phase improvement is obtained. Besides the parts unstable during equilibration, most of the errors in the optimized structures are observed in the loop regions and include missing and false main-chain connections. For some of the β-strands we observe also inadvertent reversal of the chain direction.

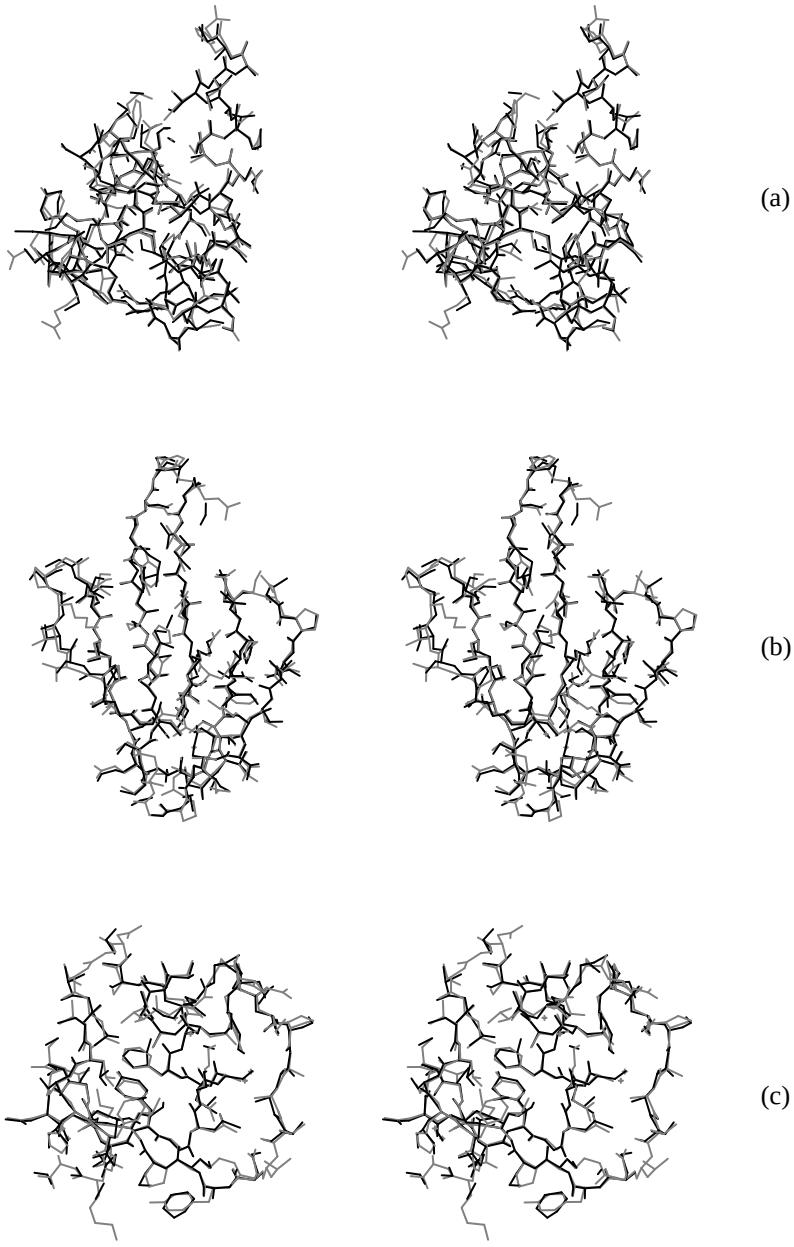

Figure 3.4: *Stereo-views of equilibrated structures in black superimposed on the target structures in gray of the (a) all-α helical case, 1I2T, (b) all-β sheet, 3EBX, and (c) mixed α/β, 1DS3, test cases.*

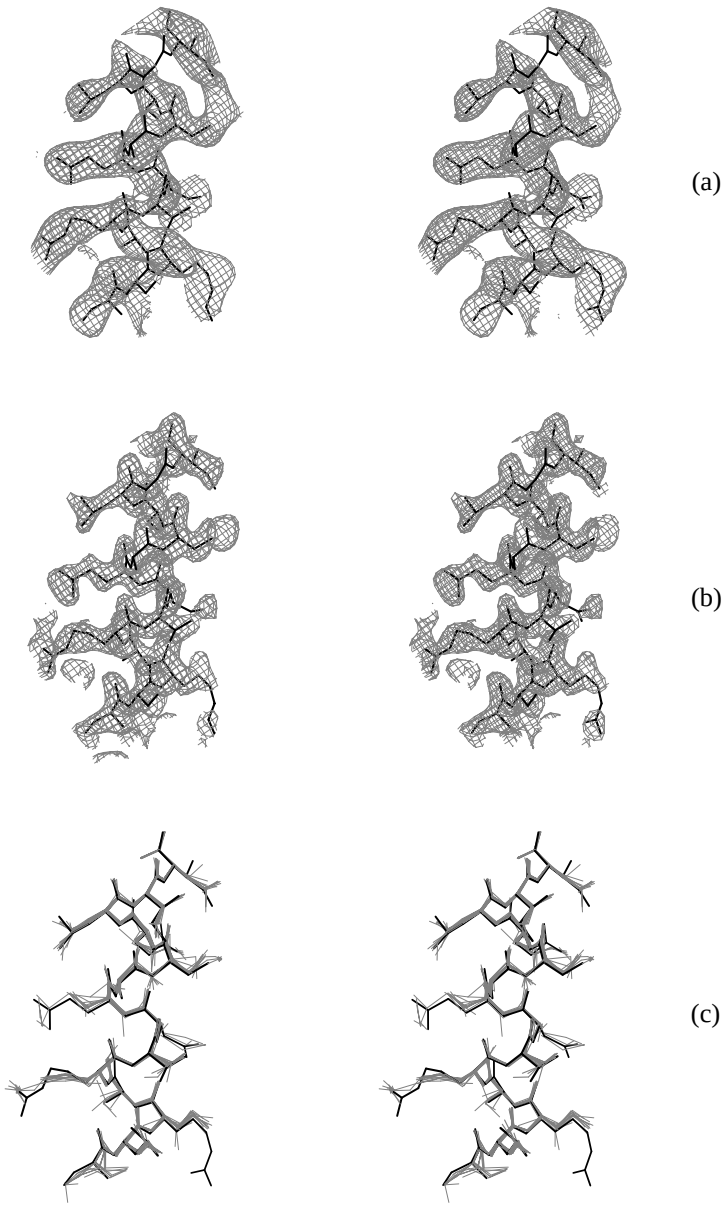

Figure 3.5: *Electron-density maps and models obtained by conditional optimization of 1I2T using twelve scrambled models. Shown are stereo-views of part of the* $m_a|F^{\mathrm{obs}}|\exp(i\varphi^{\mathrm{ave}})$*-electron density maps obtained before (a) and after (b) optimization, and the twelve final structures obtained in gray (c). The target structure of 1I2T is superimposed in black.*

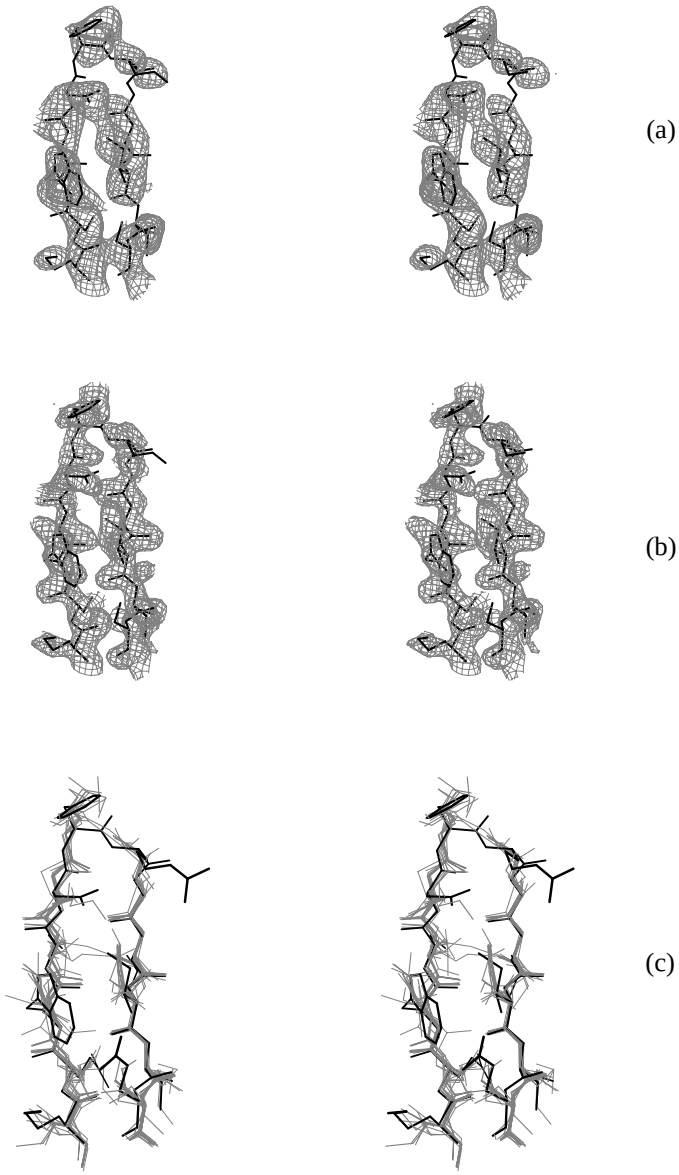

Figure 3.6: *Electron-density maps and models obtained by conditional optimization of 3EBX using twelve scrambled models. Shown are stereo-views of part of the* $m_a|F^{\mathrm{obs}}|\exp(i\varphi^{\mathrm{ave}})$*-electron density maps obtained before (a) and after (b) optimization, and the twelve final structures obtained in gray (c). The target structure of 1I2T is superimposed in black.*

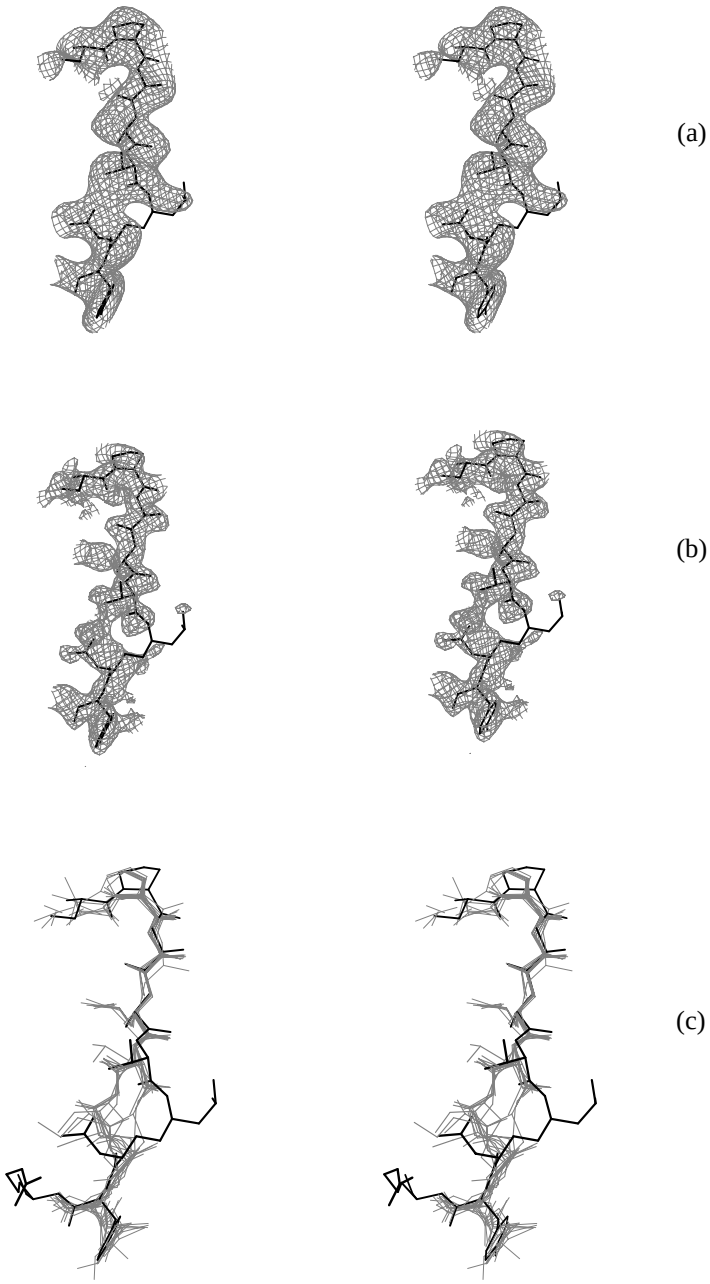

Figure 3.7: *Electron-density maps and models obtained by conditional optimization of 1DS3 using twelve scrambled models. Shown are stereo-views of part of the* $m_a|F^{\mathrm{obs}}|\exp(i\varphi^{\mathrm{ave}})$*-electron density maps obtained before (a) and after (b) optimization, and the twelve final structures obtained in gray (c). The target structure of 1I2T is superimposed in black.*

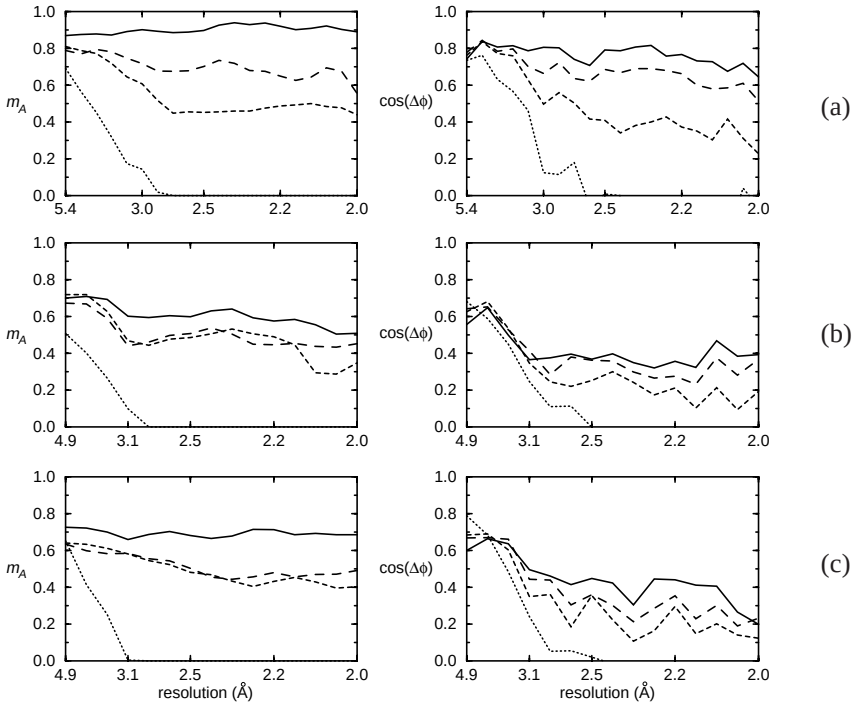

Figure 3.8: *Estimated figures of merit m_a and average cosine of the true phase error, $\langle\cos(\Delta\varphi)\rangle$, for the (a) all-α helical, 1I2T, (b) all-β sheet, 3EBX, and (c) mixed α/β, 1DS3, test cases. Values are shown as calculated for the initial, scrambled structures (dotted lines), structures after optimization cycle 1 (dashed lines) and cycle 2 (long-dashed lines) and for the final optimized structures (solid lines).*

The all-α helical test case performs significantly better in the conditional optimization than the all-β sheet and mixed α/β test cases. This difference may be attributed to various reasons: *i* for the all-α test case estimated figures of merit m_a are in good agreement with the mean cosine of the phase error, while for the all-β and mixed α/β test cases significant over-estimation is observed (see figure 3.8); *ii* the all-α test case has a higher solvent content ($\sim 50\%$) than the other two cases (both $\sim 35\%$), resulting in a significantly larger number of reflections up to 2.0 Å resolution (see table 3.3); *iii* the information content of the used force field is higher for the all-α test case than for the all-β and mixed α/β test cases; *iv* the proteins from the all-β and mixed α/β test cases contain more conformations that are not accounted for in the used force fields.

3.5 Conclusions

We introduced a potential of mean force for conditional optimization of protein structures. The interaction functions in this force field describe protein fragments in α-helical, β-strand and loop conformations of up to respectively four, three and two residues long. Distinct in-

teraction functions for the three preferred χ_1-rotamers describe corresponding geometries for side chains up to the γ-position. Notably, we omitted glycine and proline residues, main-chain conformations involving the L-region of the Ramachandran plot and torsion angles (and higher order information) for side-chain atoms beyond the γ-position, due to increasing computational costs. We tested the parameter set in conditional optimization of three small protein structures using 2.0 Å observed diffraction data. Dynamics runs starting from the deposited coordinates show that the definition of the global minimum is correct for the defined main-chain conformations and for side chains up to the γ-position. Breaks are observed for main-chain conformations outside the A and B-region and for side chains beyond the γ-position that were not or poorly defined. A more precise definition of these conformations in the force field could improve the optimization behaviour. However, inclusion of the omitted elements would give rise to a large increase of the number of possible combinations, increasing the computational cost dramatically.

Optimization starting from twelve structures with 1.5 Å r.m.s.d. random coordinate shifts showed excellent convergence for the α-helical hyperplastic discs protein. Considerable phase improvement was obtained as well for the β-sheet protein erabutoxin and the ovomucoid third domain with mixed α/β fold, but the optimized structures contain more errors, typically chain reversals for β-strands and incorrect formation of loops. The applied multiple-model procedure proved crucial for these optimizations, since with the limited numbers of available test set reflections standard procedures to estimate phase quality failed for starting models with such large errors. In contrast to the all-α helical case, significant over-estimation of the phase quality was observed for the all-β sheet and mixed α/β test cases. This over-estimation coincides with the more difficult convergence in the optimization runs of the all-β and mixed α/β test cases, which may indicate the importance of further improvement of this procedure.

Our results illustrate that a large radius of convergence may be obtained by conditional optimization of protein molecules with observed diffraction data to medium resolution. The coordinate errors of our starting models were generated in a completely random way and such favourable error distributions are hard to obtain when starting from a single electron-density map. In addition, we used truncated data, which also may have contributed favourably to the optimization behaviour. Still, the significant reduction in phase errors, from $\sim 70^\circ$ to 45° or better, is promising. The presented, generally applicable potential of mean force allows development of phase improvement and automated model-building procedures using conditional optimization, as well as investigation of the efficacy of this approach in *ab initio* phasing of protein structures.

Acknowledgements

This work is supported by the Netherlands Organization for Scientific Research (NWO-CW: Jonge Chemici 99-564).

Chapter 4

The potentials of conditional optimization in automated model building

Abstract

We have applied conditional optimization to automated model building for three test cases with data to medium resolution and good experimental phases. Compared to *ARP/wARP* and *RESOLVE*, conditional optimization yielded models of comparable phase quality, for which most of the α-helical and β-strand segments were modelled. The main difference in the results obtained was a poor modelling of loops and turns by conditional optimization. This might be improved by incorporation of more loop conformations in the conditional force field. Although iteration of conditional optimization with discrete model building steps may provide a more efficient procedure, here we only tested the potentials of conditional optimization alone. Further development of the procedures and the optimization of hybrid models with explicit assignments of atom labels may provide a method suitable for automated model building at lower resolutions and in maps of lower quality. This may then justify the large amounts of computer memory and CPU-time required for conditional optimization of protein structures.

4.1 Introduction

In protein crystallography, once the phase problem has been solved the electron density map needs to be interpreted in terms of a molecular model. This task is far from straightforward and has been shown to be prone to human error (Mowbray *et al.*, 1999). Even for a skilled crystallographer manual model building can take up to weeks of time in front of a computer graphics. In the last few years, a number of programs have been developed that aim to automate this process. By far the most widely used approach in automated model building up till now is the *ARP/wARP* program (Perrakis *et al.*, 1999). In this program a powerful combination of loose-atom refinement and recognition of protein fragments is implemented. Because of the coupling of refinement with the process of model building, the resulting models are typically highly accurate and the program does not depend strongly on the quality of the initial phase information. The major limitation of this program lies in a strong dependence on the availability of large amounts of diffraction data. Firstly, because in the automated refinement procedure (*ARP*, Lamzin & Wilson, 1993) atoms are re-positioned in electron density maps, which should be of sufficiently high resolution. Secondly, because the (almost) unrestrained loose-atom refinement in *ARP* depends intrinsically on a favourable observation-to-parameter ratio. The *ARP* refinement cycles are iterated with discrete model building steps, where the *warpNtrace* algorithm recognizes protein main-chain fragments in the refined distribution of individual atoms. Application of stereo-chemical restraints on the recognized protein fragments increases the observation-to-parameter ratio. Therefore, refinement of a hybrid model consisting of auto-built protein fragments and the remaining loose atoms is less dependent on the number of reflections available. Recently, major advances in the *warpNtrace* algorithm have been reported (Morris *et al.*, 2002), currently allowing main-chain tracing at 2.5 Å resolution.

Several alternative approaches simulate the process of manual building, where fragments of secondary structure elements are positioned in the electron density map. For this task various pattern recognition techniques have been implemented in programs like *TEXTAL* (Holton *et el.*, 2000), *QUANTA* (Oldfield, 2000) and *RESOLVE* (Terwilliger, 2002). A major advantage of these methods is that, in principle, they can be applied to electron density maps of medium to low resolution. A disadvantage is a strong dependence on the quality of the initial phase information. For electron density maps of limited quality or resolution, the positioned fragments may suffer from a low accuracy. Iterative application of these model building techniques with protein structure refinement may provide a solution to this problem. Such an approach has been implemented in the latest version of *RESOLVE* (Terwilliger, 2002). In this program secondary structure elements are positioned in the electron density map using phased translation functions (Cowtan, 1998). Subsequently, the positioned fragments are extended into loop regions and the primary sequence is docked on the constructed models. A final model is obtained by iteration of these building steps with maximum-likelihood density modification and restrained protein structure refinement in *REFMAC* (Murshudov *et al.*, 1997).

With the method of conditional optimization (Scheres & Gros, 2001), we introduced a technique that allows loose-atom refinement without the intrinsic need for high-resolution diffraction data. In conditional optimization, the number of observations from the diffraction experiment is supplemented with extensive geometrical information, without the require-

ment of an explicit chemical assignment of the unlabelled atoms. Initial test calculations with this method showed a large radius of convergence, allowing successful refinement starting from random atom distributions for an artificial test case. The introduction of a force field describing commonly observed protein conformations (Scheres & Gros, 2003) allowed application to protein molecules, for which a large radius of convergence was observed using observed diffraction data. The large radius of convergence and the limited dependence on high-resolution data raised expectations for the application of conditional optimization to automated model building in electron density maps of limited resolution. In analogy to *ARP/wARP* and *RESOLVE*, the most powerful approach would probably be an iterated process of refinement cycles and discrete model-building steps. However, rather than providing a ready-to-use solution, we chose to first test the potentials of conditional optimization alone in the model building process. Therefore, in the calculations presented here, no other pattern recognition or model building techniques were applied than conditional optimization itself. Still, for the three test cases presented conditional optimization yielded models of comparable quality as *ARP/wARP* and *RESOLVE*, in which most of the α-helices and β-strands were built.

4.2 Experimental

Three protein structures were selected for testing purposes: the A3-domain from human von Willebrand Factor (vWF-A3, Huizinga *et al.*, 1997), outer-membrane protein NspA from *Neisseria meningitidis* (Vandeputte-Rutten *et al.*, in preparation) and the C-terminal domain of leech anti-platelet protein (LAPP, Huizinga *et al.*, 2001). All three structures were solved in our laboratory and initial models were built manually using the graphics program *O* (Jones *et al.*, 1991). Main characteristics of these test cases are given in table 4.1. The structure of vWF-A3 was solved at 2.35 Å resolution in space group $P2_12_12_1$ by multiple-wavelength anomalous disperson methods (MAD) using the anomalous contribution from four selenomethionine residues. An initial model built in the excellent experimental electron density map was transformed to space group $P2_1$ of the native crystals and subsequent refinement was carried out using native data up to 1.8 Å resolution. Here, automated model building was performed using only the MAD-data to 2.4 Å resolution. The structure of NspA was solved by single-wavelength anomalous diffraction (SAD) at 4 Å resolution using the strong anomalous signal from two gold atoms, which were bound to the protein by soaking the crystal in a solution containing $Au(CN)_2$. Subsequent density modification and phase extension of a native data set to 2.6 Å resolution yielded an easily interpretable electron density map. This map was used for automated model building here. For LAPP, three heavy-atom sites were identified in a K_2PtCl_4 derivative and the structure was solved at 3.1 Å resolution using single isomorphous replacement with anomalous scattering (SIRAS). Density modification by solvent flattening and three-fold non-crystallographic symmetry (NCS-) averaging was used for phase extension of a low-resolution native data set to 3.0 Å. The resulting electron density map was of high quality and allowed construction of an initial model. A final model was obtained by refinement against a high-resolution native data set to 2.2 Å resolution. Here, the 3.0 Å map after phase extension was used for automated model building.

Automated model building was performed in separate cycles of conditional optimiza-

	vWF-A3	**NspA**	**LAPP**
space group	$P2_12_12_1$	$R32$	$P4_322$
Z	1	1	3
no. residues/ molecule	183	155	88
solvent content (%)	35	70	70
resolution limit (Å)	2.4	2.6	3.0
I/σ_I in outer shell	10.7	5.2	2.6
completeness (%)	99.7	98.5	96.6
no. reflections	6404	9768	11402
phasing methods	MAD	SAD + DM	SIRAS + DM
$\Delta\varphi(^\circ)$	35.6	38.3	28.2
$\cos(\Delta\varphi)$	0.59	0.49	0.63
sec. structure content (% α, % β, % loop)	50, 30, 20	0, 75, 25	40, 60, 0
no. atoms in CO	1450	1200	2200

Table 4.1: *Main characteristics of the three test cases presented: the A3-domain of von Willebrand factor (vWF-A3), outer-membrane protein NspA and the C-terminal domain of leech anti-platelet protein (LAPP). Phase errors of the experimental phases ($|F^{\text{obs}}|$-weighted $\Delta\varphi$ and unweighted $\cos(\Delta\varphi)$) were calculated with respect to the refined structures. For vWF-A3, the published coordinates in space group $P2_1$ were transformed to space group $P2_12_12_1$ and subjected to rigid body and energy minimization refinement. For NspA, a structure from the final stages of refinement (Vandeputte-Rutten, personal communication) was used. For LAPP, rigid body and energy minimization refinement was performed with the published coordinates to compensate for the differences in unit cell parameters between the high and low-resolution native data sets. For all three test cases the secondary structure contents that were used to generate the conditional force fields are given in precentages α-helix, β-sheet and loop. The numbers of atoms used in the conditional optimization (CO) runs correspond to approximately 1.05 times the number of atoms in the published models.*

tion according to the protocol shown in figure 4.1. We used a phase-restrained maximum-likelihood crystallographic target function (MLHL) (Pannu *et al.*, 1998) with cross-validated σ_A-values, estimated by Read's procedure (1986). Target values for the phase restraints and corresponding figures of merit were obtained from the MAD-experiment for the vWF-A3 case and from the density modification procedure for the NspA and LAPP test cases. Protein-specific force fields for conditional optimization were generated from the general force field as described before (Scheres & Gros, 2003), using secondary structure contents as given in table 4.1. For LAPP loop conformations were excluded from the force field to limit computer memory requirements. A starting model for the first optimization cycle was generated by filling the experimental map with unlabelled atoms for the regions with density levels above 1.0σ. The number of atoms positioned in the electron density map of each test case are given in table 4.1. After each cycle of conditional optimization, phases from the optimized model were combined with the experimental phases and a new, combined electron density map was calculated. A starting model for the next optimization cycle was generated by maintaining the positions of the atoms in the optimized model that were recognized as part of a protein fragment, and filling the remaining regions of the combined map with new atoms. Recog-

nition of protein fragments was based on the gradient coefficients towards all possible atom assignments as calculated for every atom in the conditional optimization. Atoms were recognized to be part of a protein fragment if the gradient contribution towards one of the atom types N, C^{α}, O, C, C^{β}, C^{γ} or S^{γ} was at least two times as large as the second largest contribution. Only protein fragments consisting of at least two consecutive atoms were taken into account. No attempts to model side chains extending beyond the γ-position were made. A final protein model was constructed from the protein fragments that were recognized in the optimized model after the last optimization cycle. For the vWF-A3 and NspA test cases two cycles of conditional optimization were performed; for the LAPP test case four.

For comparison of the results obtained by conditional optimization, the electron density maps of all three test cases were also subjected to automated model building by *ARP/wARP* (version 6.0) and *RESOLVE* (version 2.03), using *REFMAC* version 5.1.24 for refinement. These calculations were performed using only default values for all parameters. In *RESOLVE* docking of the primary sequence on the constructed fragments by side-chain modelling was included in the model building process. Modelling of the side chains was not performed with *ARP/wARP*, since this option of the program yielded significantly worse results (not shown).

All calculations were performed on a 667 MHz single-processor Compaq XP1000 work station with 2 Gb of computer memory.

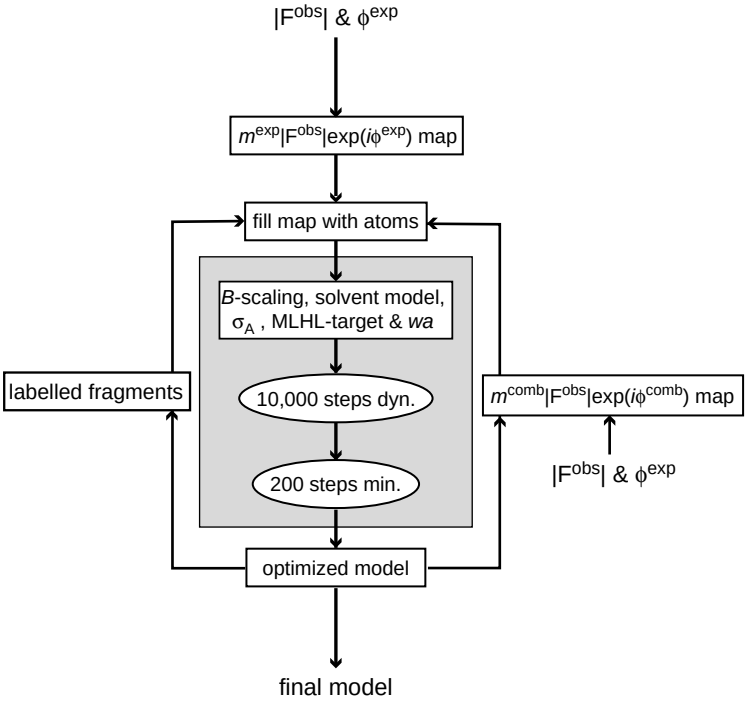

Figure 4.1: *Refinement protocol for automated model building by conditional optimization. Electron density maps were calculated on a gridsize of 0.2 Å. Regions of the experimental map ($m^{\mathrm{exp}}|F^{\mathrm{obs}}|\exp(i\varphi^{\mathrm{exp}})$) with density levels above 1.0σ were filled with atoms labelled "X", regarding criteria of minimum and maximum inter-atomic distances of 1.1 and 1.8 Å respectively, and a maximum of four neighbouring atoms within a distance of 1.8 Å. Every optimization cycle (gray) comprised 10,000 steps of conditional dynamics (dyn.) and 200 steps of energy minimization (min.). A time step of 0.2 fs was used for the dynamics calculations and the velocities were scaled to a constant temperature of 600K. A bulk solvent model was calculated using the mask method as implemented in CNS. Protein masks were calculated around the atoms with the highest number of neighbours within 3.6+0.9 Å (see Scheres & Gros, 2001 for the definition of the number of neighbouring atoms within a distance of $d+\sigma_d$ Å). The cutoff in the number of neighbouring atoms was chosen such that the remaining solvent region was as least as large as the expected solvent content. Atoms with a lower number of neighbours ended up in the solvent region and their occupancy was set to zero. Overall anisotropic B-factor scaling, σ_A-estimation, determination of weight wa on the crystallographic part of the target function (MLHL) and phase combination were performed using standard CNS-routines (Brünger et al., 1998). Recognition of protein fragments in the optimized models was performed as described in the main text. The atomic positions of these fragments were maintained in the filling of the combined electron density map ($m|F^{\mathrm{obs}}|\exp(i\varphi^{\mathrm{comb}})$) after the first optimization cycle.*

4.3 Results

For all three test cases, automated model building by conditional optimization yielded models of comparable phase quality as the models built by *ARP/wARP* and *RESOLVE*. Figures 4.2, 4.3 and 4.4 display the models of respectively vWF-A3, NspA and LAPP, generated by the three methods. Overall quality criteria for these models are shown in table 4.2. In general, conditional optimization generated models of a lower connectivity and with less loops and turns, compared to the models built by *ARP/wARP* or *RESOLVE*. In none of the three test cases conditional optimization yielded the most complete model, but the accuracy of the fragments positioned by conditional optimization was relatively good. For vWF-A3 both conditional optimization and *RESOLVE* obtained a significant phase improvement with respect to the MAD-phases, while the phases resulting from *ARP/wARP* were not better than the experimental ones.

For vWF-A3, conditional optimization yielded a model consisting of almost all α-helical and β-strand segments, except one small α-helix. Only one of the loops was modelled correctly. Another loop was modelled with a reversed chain direction, and also a small strand flanking the central β-sheet was built in the wrong chain direction. Both *ARP/wARP* and *RESOLVE* built models of higher completeness for this case, mainly due to a better modelling of the loop regions. The most complete model was built by *RESOLVE*, including a correct modelling of most of the side chains. In this model only one loop is missing, as well as the same small α-helix that was not built by conditional optimization. In the model built by *ARP/wARP* this α-helix is also missing, as well as one β-strand and two loops. No main chain trace errors were observed for the models built by *ARP/wARP* and *RESOLVE*.

For NspA, conditional optimization built most of the strands in the β-barrel, but none of the turns. The main errors in the generated model are reversed chain directions for one entire β-strand and for two smaller fragments. *ARP/wARP* built a model of higher completeness, including also two of the turns. Two of the β-strands in this model were built with a reversed chain direction. *RESOLVE* built the model with the lowest completeness, and this model contains a tracing error in the form of a crossing from one strand to a neighbouring one, resulting in a reversed chain direction for part of the neighbouring strand.

For LAPP, conditional optimization yielded a model consisting of partially modelled β-sheets and most of the α-helical segments for the three molecules in the asymmetric unit. Reversed chain directions were observed for some of the β-strands and for one α-helix. One of the loops was modelled incorrectly by an α-helical turn. *ARP/wARP* built a more complete model with more β-strands and more loops. One incorrect main-chain trace from an α-helix to a neighbouring β-strand was observed in this model. As for NspA, *RESOLVE* obtained the model with the lowest completeness, and besides a low accuracy of the positioned fragments, more main-chain trace errors were observed for this model than for the models built by conditional optimization and *ARP/wARP*.

test case:	**vWF-A3 (2.4 Å)**			**NspA (2.6 Å)**			**LAPP (3.0 Å)**		
method:	**CO**	**ARP**	**RESOLVE**	**CO**	**ARP**	**RESOLVE**	**CO**	**ARP**	**RESOLVE**
no. residues	148	160	170	121	130	101	169	232	136
frac. built (%)	80	87	93	78	84	65	64	87	51
no. chains	20	8	4	16	6	10	38	11	13
rmsd (Å)	1.5	1.9	0.9	1.3	1.8	0.9	1.2	1.3	1.8
$\Delta\varphi(^\circ)$	27.2	36.0	23.9 (22.7)	35.8	33.6	42.4 (35.6)	25.4	27.9	56.5 (26.0)
$\langle\cos(\Delta\varphi)\rangle$	0.67	0.56	0.72 (0.69)	0.53	0.60	0.46 (0.52)	0.67	0.64	0.32 (0.66)
CPU (h)	36	1	15	38	2.5	18	105	1.5	14

Table 4.2: *Overall quality criteria for the models of vWF-A3, NspA and LAPP built by conditional optimization (CO), ARP/wARP (ARP) and RESOLVE. The number of residues and chains in every generated model are given, as well as the fraction (frac.) of the total number of residues built. Root-mean-square coordinate errors (rmsd) were calculated as the distance between atoms in the modelled protein fragments to the nearest atom with the corresponding label in the refined structure. Phase errors with respect to the refined structures ($|F^{obs}|$-weighted $\Delta\varphi$ and unweighted $\cos(\Delta\varphi)$) were calculated for the all-atom models resulting from the three methods. For RESOLVE phase errors of the resulting electron density map are given between brackets. Also given are the CPU-times required for ARP/wARP, RESOLVE and the performed cycles of conditional optimization.*

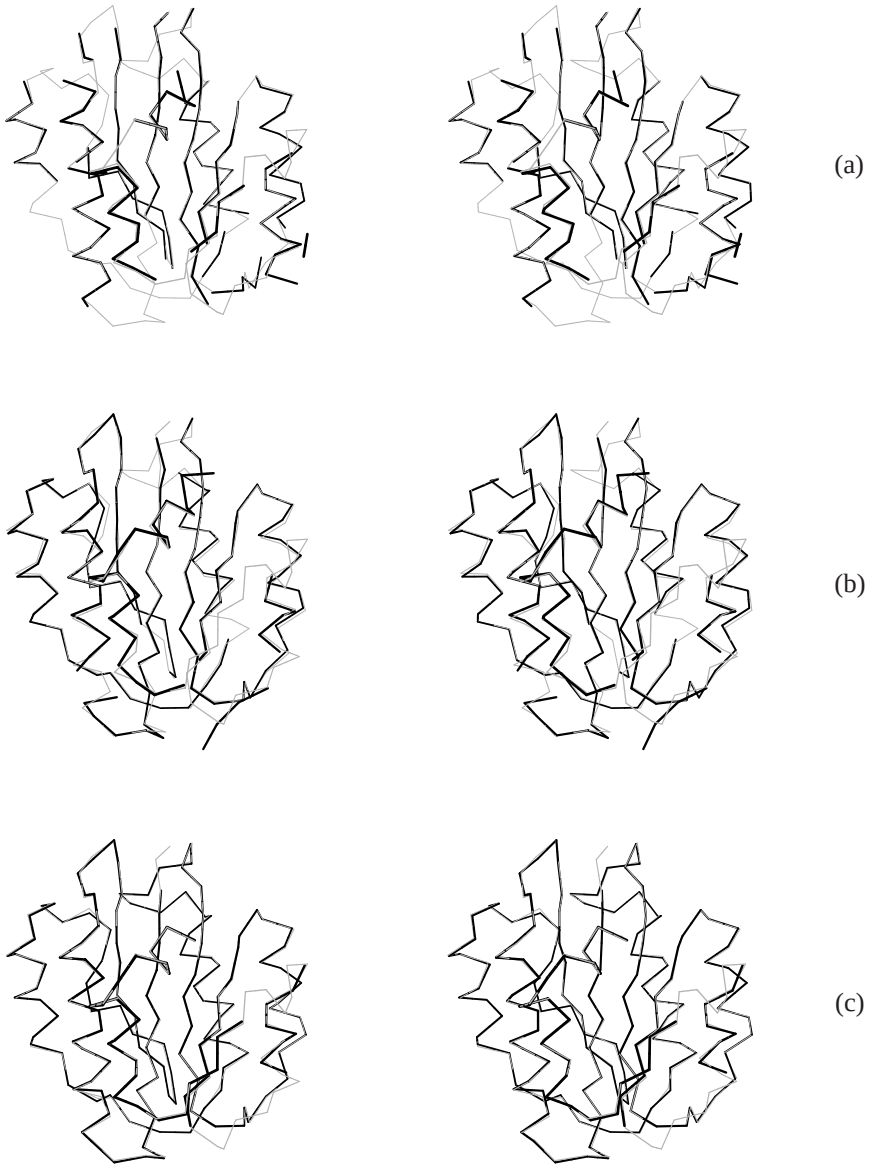

Figure 4.2: *Stereo-views of the automatically built models (black) generated by (a) conditional optimization, (b) ARP/wARP and (c) RESOLVE, superimposed on the target structure of vWF-A3 (gray).*

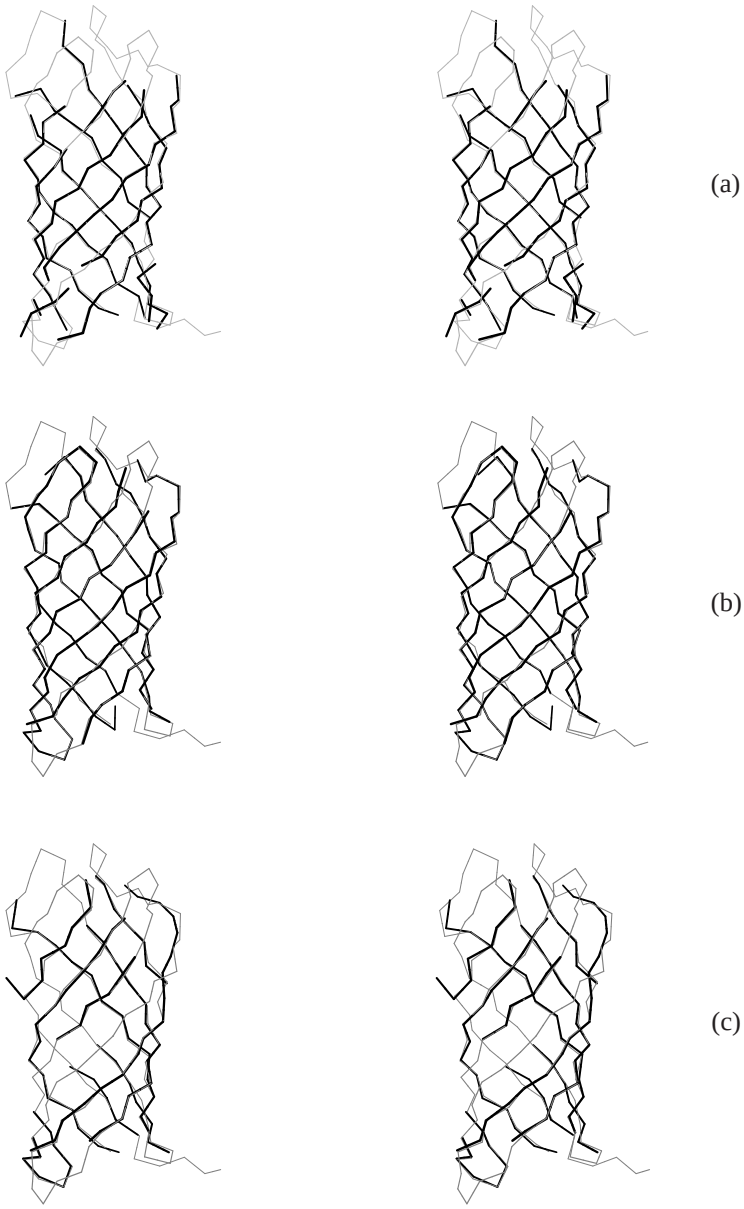

Figure 4.3: *Stereo-views of the automatically built models (black) generated by (a) conditional optimization, (b) ARP/wARP and (c) RESOLVE, superimposed on the target structure of NspA (gray).*

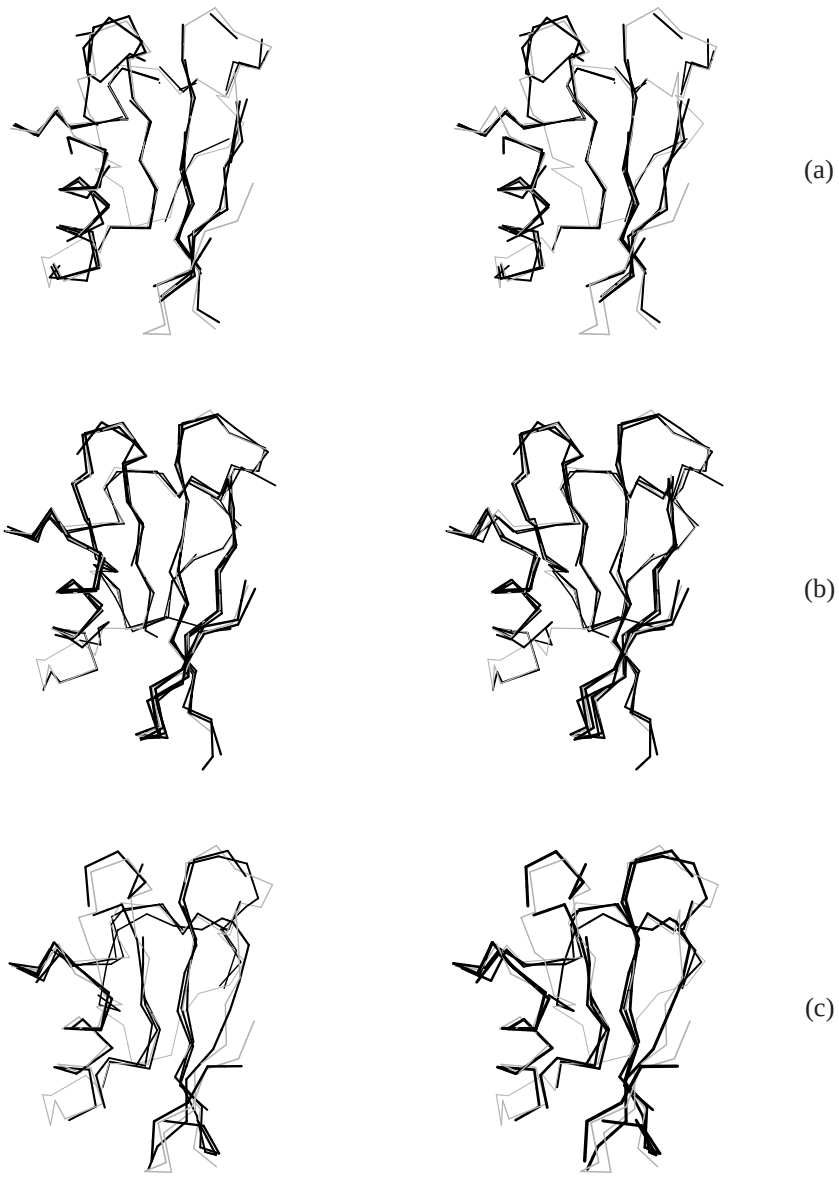

Figure 4.4: *Stereo-views of the automatically built models for LAPP generated by (a) conditional optimization, (b) ARP/wARP and (c) RESOLVE. Auto-built protein fragments for all three molecules in the asymmetric unit (black) are superimposed on one of the target molecules (gray).*

Conditional optimization required large amounts of CPU-time compared to *ARP/wARP* and *RESOLVE* (see table 4.2). With more than an order of magnitude difference with conditional optimization, *ARP/wARP* was the fastest program. Conditional optimization required also much more computer memory than *ARP/wARP* or *RESOLVE*. Up to 1.5 and 1.4 Gb of memory was allocated for conditional optimization of the vWF-A3 and NspA test cases respectively. To allow conditional optimization of LAPP on our work station with 2 Gb of memory, the program was re-compiled using single-precision instead of double-precision numbers for the storage of all condition values in computer memory. With the re-compiled program still up to 2.4 Gb of memory was allocated.

4.4 Discussion & conclusions

Conditional optimization yielded relatively accurate models for all three test cases, which consisted of most of the α-helical and β-strand segments. The models obtained were of comparable phase quality as the models built by *ARP/wARP* or *RESOLVE*, but they consisted of less loops and turns and a higher number of separate chains. The most prominent errors observed in these models were reversed chain directions for β-strands. In the lowest resolution test case, *i.e.* LAPP, also one α-helix was built with a reversed chain direction and one loop was incorrectly modelled by a helical turn. The models generated by *ARP/wARP* and *RESOLVE* showed similar errors, which reflect the difficulties of model building at lower resolutions.

Very few loops and turns were built by conditional optimization, and this forms the main difference with the results from *ARP/wARP* and *RESOLVE*. The poor modelling of loops and turns may be ascribed to two factors. Firstly, loop conformations were poorly defined in the applied conditional force fields. For LAPP, loop conformations were not taken into account at all, and for vWF-A3 and NspA only loop conformations corresponding to the A and B-region of the Ramachandran plot were included (see Scheres & Gros, 2003). Most of the loops and turns in these proteins contain residues with conformations that were not defined by the force field and thus could not be modelled. This concerns residues with conformations in the L-region or residues (mainly glycines) with conformations outside any of the allowed regions of the Ramachandran plot. For automated model building, extension of the force field with these conformations may yield better results. Secondly, conditional optimization of loops may converge less readily than observed for α-helices and β-strands, because the information content of the force field is lower for loop conformations. Similar problems may exist for side chain conformations extending beyond the γ-position. This is a consequence of the higher structural variability of loops and side chains compared to main chain conformations in α-helices and β-strands.

The models built by conditional optimization showed a lower connectivity than the models generated by *ARP/wARP* or *RESOLVE*. This concerns not only the poor modelling of loops and turns, but also the presence of additional breaks in β-strands and α-helices. These breaks result in a larger number of separate chains, which would require more manual rebuilding. The occurrence of the breaks may be attributed to the absence of decision-based model building steps as implemented in *ARP/wARP* and *RESOLVE*. In these programs, complete oligo-peptide fragments are positioned in discrete steps. In the applied conditional op-

timization protocol, protein fragments could be formed solely during the continuous process of refinement. The procedure to recognize protein fragments based on gradient contributions did not alter any coordinates; only fragments with correct geometries were recognized. Therefore, minor topological errors in the optimized distributions of loose atoms resulted in breaks in the recognized protein chains. Iteration of conditional optimization cycles with discrete model building steps, where optimized distributions of loose atoms would be replaced by atoms with the geometric arrangements of ideal protein fragments could lead to models of higher connectivity and completeness. A second option that might enhance the formation of longer protein fragments would be to explicitly assign corresponding atom labels to atoms that are recognized as part of a protein fragment. For these atoms no longer all possible chemical assignments would be taken into account in the conditional optimization, which would reduce the degeneracy of the force field. Consequently, labelled protein fragments may grow faster into longer segments than in the current approach, where constantly all assignments are considered.

For the NspA and LAPP test cases, *ARP/wARP* yielded better results than *RESOLVE* and conditional optimization. For the vWF-A3 case however, *RESOLVE* yielded the most complete model, and in contrast to conditional optimization and *RESOLVE*, *ARP/wARP* did not achieve a phase improvement with respect to the MAD-phases. Despite the higher resolution limit of this test case, the number of observations is relatively low due to a small solvent content. Possibly, the unrestrained ARP-refinement suffered from an unfavourable observation-to-parameter ratio. The large amounts of geometric restraints in conditional optimization and the restrained refinement as implemented in *RESOLVE* may have provided a better defined refinement of the partially built models against the limited number of reflections. The phase improvements obtained with these methods indicate that also with limited amounts of diffraction data significant phase extension by automated model building may be feasible, as was already observed for model building at higher resolutions using *ARP/wARP* (for an extreme example see Tame, 2000).

The main drawback of the conditional optimization approach is formed by the extensive computational cost. Memory and CPU-time requirements are more than an order of magnitude larger than for *ARP/wARP*, and conditional optimization is more than two times as slow as *RESOLVE*. Iteration of conditional optimization cycles with discrete model building steps or explicit assignment of atom labels may provide a more efficient procedure, but the computational cost will remain high. For the vWF-A3 case, *ARP/wARP* may have suffered from a relatively low number of reflections. *RESOLVE* yielded a rather incomplete structure for LAPP, possibly due to the inaccuracy of the positioned fragments in the low-resolution map. Conditional optimization yielded relatively accurate models for all three test cases and does not depend on large numbers of reflections. Besides, a large radius of convergence has been observed for this method before (Scheres & Gros, 2003). Therefore, with the developments mentioned above, conditional optimization might eventually allow automated model building at lower resolutions and in electron density maps of lower quality. Dividing the computational cost over multiple processors by parallelization of the program could then render conditional optimization suitable for common practice.

Acknowledgements

We are very grateful to Lucy Vandeputte-Rutten and Eric Huizinga for providing diffraction data and coordinates and for useful comments and discussions. This work is supported by the Netherlands Organization for Scientific Research (NWO-CW: Jonge Chemici 99-564).

Chapter 5

Testing conditional optimization for application to *ab initio* phasing of protein structures

Abstract

At the resolution limits typically obtained in protein crystallography, the phase problem is underdetermined and requires incorporation of additional prior knowledge. Conditional optimization allows expression of geometric knowledge about protein structures to be combined with refinement of loose, unlabelled atoms. We have tested the application of conditional optimization to *ab initio* structure determination of four-helix bundle *Alpha-1*. The results obtained with observed diffraction data to 2.0 Å resolution and with calculated intensities for four reflections illustrate the importance of low-resolution reflections and reliable phase probability estimates. Although convergence was very slow, a steady improvement in map correlation coefficients and phase errors was observed, illustrating that for this case *ab initio* phasing by conditional optimization is possible. Further development is currently hindered by excessive computational costs, but possibilities for advances are indicated that hopefully may lead to a practical application of this approach in protein crystallography.

5.1 Introduction

After obtaining diffracting crystals, the phase problem is a critical step in protein crystallography. When diffraction data to atomic resolution is available the problem is overdetermined and therefore solvable in principle. Exploiting relationships among structure factors, direct methods nowadays allow routine *ab initio* phase calculation in small molecule crystallography (reviewed by Hauptman, 1997). Data to atomic resolution is usually not observed for protein crystals and *ab initio* phasing by direct methods has not yet been commonly possible in protein crystallography. To supplement the limited information from the observed intensities alone, other sources of information are critical to obtain phase information in protein crystallography. Typically, isomorphous or anomalous intensity differences are used in experimental phasing techniques (see for example Drenth, 1999). In the favourable cases that the structure of a homologous protein is known, molecular replacement (reviewed by Rossmann, 2001) can be used to obtain initial phases.

A wealthy source of prior information is formed by the available knowledge about the geometry of protein structures. Protein structures consist of polypeptide chains arranged in secondary structure elements with well-known geometries. Although the power of this knowledge has been illustrated through the successful application of geometric restraints in protein structure refinement, it has yet been scarcely used in *ab initio* phasing. The main reason for this lies in the difficulty of expressing this knowledge in a way that allows efficient optimization when no or limited crystallographic phase information is available. With conditional optimization we presented a method that allows expression of geometric knowledge, without the requirement of a topological assignment of the individual atoms (Scheres & Gros, 2001). Given an estimate about the secondary structure content of the crystal, this knowledge can be expressed in the absence of any phase information through geometric restraints acting on distributions of unlabelled atoms.

For a simple test case of four poly-alanine helices, we showed that in principle successful refinement of random atom distributions against medium-resolution diffraction data is possible (Scheres & Gros, 2001). Standard routines to estimate phase probabilities fail for models with such large coordinate errors, and a novel procedure to estimate σ_A-values from the distribution of multiple models was necessary for successful optimization of random models. These calculations were performed with model diffraction data and protein structures are more complex than the simplified model of this test case. Therefore, the feasibility of this approach remains to be shown for protein structures using observed diffraction data.

Here, we present conditional optimization of random atom distributions against 2.0 Å observed diffraction data of four-helix bundle *Alpha-1* (Privé *et al.*, 1999). In the first instance, calculations were performed according to the protocols as developed for the *ab initio* phasing of the poly-alanine test structure (Scheres & Gros, 2001), and the optimization of three small protein structures against observed diffraction data (Scheres & Gros, 2003). Since optimization according to these protocols did not result in convergence for this case, an alternative multiple-model procedure to estimate the phase quality of the optimized structures was investigated. Also, the influence of four reflections at low resolution, which were likely measured incorrectly, was examined. Replacing the suspect intensities with calculated values and estimating the phase quality of each individual structure separately appeared critical for convergence towards an interpretable electron density map in terms of helical elements.

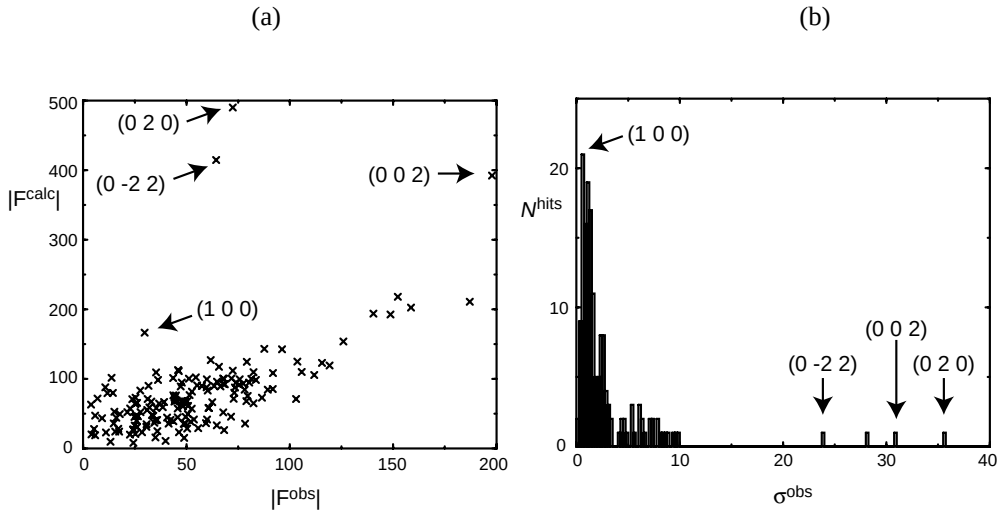

Figure 5.1: *(a) Calculated versus observed structure factor amplitudes for all observed reflections with Bragg spacing $d > 5$ Å. In the model structure factor calculation a bulk solvent contribution was taken into account using the standard mask method as implemented in CNS. Four suspect reflections show a large difference between observed and calculated structure factor amplitudes. For these reflections (corresponding hkls are indicated with arrows), observed structure factor amplitudes were replaced by calculated values. (b) Histogram of the measurement errors (σ^{obs}) of all observed reflections with Bragg spacing $d > 5$ Å. For three of the four suspect reflections a large measurement error was observed. For a fourth reflection with a large measurement error (with hkl = 001) no large discrepancy between observed and calculated structure factor amplitudes was observed.*

5.2 Experimental

5.2.1 Test case

Four-helix bundle *Alpha-1* was selected as a test case. This structure consists of 396 protein atoms in space group *P*1 with unit-cell parameters *a* = 20.846, *b* = 20.909, *c* = 27.057 Å, α = 102.40, β = 95.33, γ = 119.62^o (PDB-code 1byz; Privé *et al.*, 1999). The structure was originally solved by direct metods using all observed diffraction data to 0.9 Å resolution. Here, we truncated deposited structure-factor amplitudes to 2.0 Å resolution. Analysis of this nearly complete data set (1 out of 2549 reflections is missing) showed that up to 5 Å resolution, four reflections were measured with much lower intensity than calculated from the deposited coordinates after scaling and bulk solvent correction (see figure 5.1a). For three of these reflections also a significantly higher value for the measurement error was observed (figure 5.1b). The observed structure-factor amplitudes of these four suspect reflections were replaced by their calculated values.

A force field for conditional optimization of this all-helical test structure was generated using the general parameter set as described by Scheres & Gros (2003). An expected secondary structure content of 100% α-helix was used. The defined force field contained condi-

tions describing linear protein fragments of up to twelve bonds long in an α-helical conformation. Side chain conformations up to the γ-position were described in the two χ_1-rotamers that are commonly observed in α-helices. Limited information was included about side chains extending beyond the γ-position.

5.2.2 Optimization protocol

Figure 5.2 shows the refinement protocol, as implemented in the program *CNS* (Brünger *et al.*, 1998a), for conditional optimization starting from multiple models consisting of randomly positioned atoms in the unit cell. In the absence of any prior phase information, a maximum likelihood crystallographic target function on amplitudes (MLF; Pannu & Read, 1996) was set for the first optimization cycle, and σ_A-values were calculated according to an exponential decrease with the length of scattering vector $\vec{S}$: $\sigma_A = \exp(-150 \times |\vec{S}|^2)$. After 1,000 steps of conditional dynamics, the N individual structures were positioned on a common origin by iteratively shifting each structure using a phased translation function with phases from the average structure factor F^{ave}. With all individual structures sharing a common origin, the phases from F^{ave} served as target values in the phase-restrained maximum likelihood crystallographic target function (MLHL; Pannu *et al.*, 1998) of subsequent conditional optimization cycles. These cycles comprised 10,000 steps of conditional dynamics and phase probabilities were estimated as described in section 5.2.3. After each cycle the individual structures were re-positioned on a common origin and F^{ave} was updated.

Within each cycle of MLHL-refinement, atomic B-factors were assigned based on the numbers of neighbouring atoms as described before (Scheres & Gros, 2001). To avoid negative atomic B-factors after overall isotropic B-factor scaling, inverse scaling was applied to $|F^{\text{obs}}|$ rather than scaling $|F^{\text{calc}}|$. A bulk solvent contribution was calculated using the standard mask routines implemented in *CNS*. Given an expected solvent content of 20%, a mask covering 80% of the unit cell volume was calculated around the atoms with the highest numbers of neighbours. The occupancy of all atoms inside the remaining solvent region was set to zero. Weights *wa* on the crystallographic part of the target function were calculated based on a relationship with the sum of D (as calculated from σ_A, see Read, 1986) over all reflections: $wa \propto 1/\Sigma D$. A randomly selected 10% of all data with Bragg spacing $d < 10$ Å were selected for cross-validation purposes (Brünger, 1993). As described before (Scheres & Gros, 2001), an additional 5% of the data were taken out of refinement and this selection was modified every 1,000 steps to avoid stalled progress owing to local minima in the crystallographic target function.

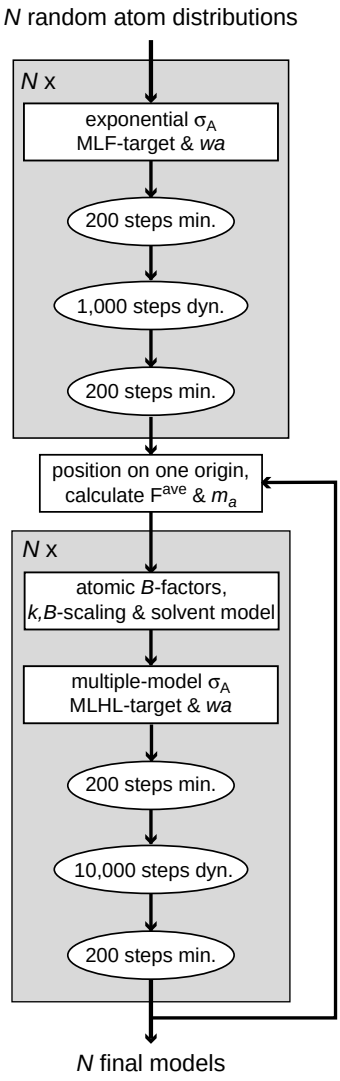

Figure 5.2: *Refinement protocol for ab initio phasing by conditional optimization. In every optimization cycle (gray) conditional dynamics coupled to a temperature bath of 600K (dyn.) was preceded and followed by energy minimization (min.). Positioning of the individual models on a common origin, calculation of atomic B-factors, overall temperature-factor scaling, calculation of a bulk solvent contribution, determination of weight wa on the crystallographic part of the target function and estimation of figures of merit (m_a) and σ_A-values (using an exponential function or multiple-model procedures) were performed as described in section 5.2.*

5.2.3 Phase probability estimation

Two types of phase probabilities need to be estimated for phase-restrained (MLHL) maximum-likelihood refinement: *i.* figures of merit for the average structure factors F^{ave} of the phase restraint and *ii.* σ_A-estimates for the individual models. Both must be estimated as a function of resolution.

Shell-wise estimates m_a for the figures of merit of the phase restraint were calculated by (5.1), using only test-set reflections:

$$m_a = \sqrt{\frac{N(m'_a)^2 - 1}{N - 1}} \tag{5.1}$$

where $m'_a = \sum_{i=1}^{N} F^i / \sum_{i=1}^{N} |F^i|$ and individual structure factor sets F^i are calculated from the corresponding N models (Scheres & Gros, 2003).

Two ways to estimate σ_A-values for the individual models were tested.
i. As described before for conditional optimization of three small protein structures (Scheres & Gros, 2003), cross-validated figures of merit m_a for the average structure factor were converted to $\sigma_A^{a(i)}$-estimates for every model i by (5.2):

$$\sigma_A^{a(i)} = \frac{\langle |E^{\text{obs}}||E^i| m_a \rangle}{\sqrt{\langle |E^{\text{obs}}|^2 \rangle \langle |E^i|^2 \rangle}} \tag{5.2}$$

where $|E^{\text{obs}}|$ and $|E^i|$ are observed and calculated normalized structure factor amplitudes. These estimates will be referred to as σ_A^a because the differences between the different models are small due to the common figure of merit.
ii. Different σ_A-estimates were calculated for each individual model, assuming that the true phase error of a model relates to the observed phase differences of that model with all other models. σ_A^i-Values for every model i were calculated by averaging shell-wise σ_A^{ij}-estimates over all other models j (5.3):

$$\sigma_A^i = \langle \sigma_A^{ij} \rangle_j = \left\langle \frac{\langle |E^{\text{obs}}||E^i| \cos(\varphi^i - \varphi^j) \rangle}{\sqrt{\langle |E^{\text{obs}}|^2 \rangle \langle |E^i|^2 \rangle}} \right\rangle_j \tag{5.3}$$

σ_A^i-Values were calculated using all reflections because this calculation was unstable for the low numbers of reflections in the test set alone.

All calculations were performed on four, 667 MHz single-processor Compaq XP1000 workstations with at least 1.2 Gb of computer memory.

5.3 Results

5.3.1 Condensation and the influence of low-resolution data

Thirty-six random atom distributions were subjected to an initial optimization cycle comprising 1,000 steps of conditional dynamics using a MLF crystallographic target function. Figure 5.3 shows a typical arrangement of the atoms resulting from these optimizations, where

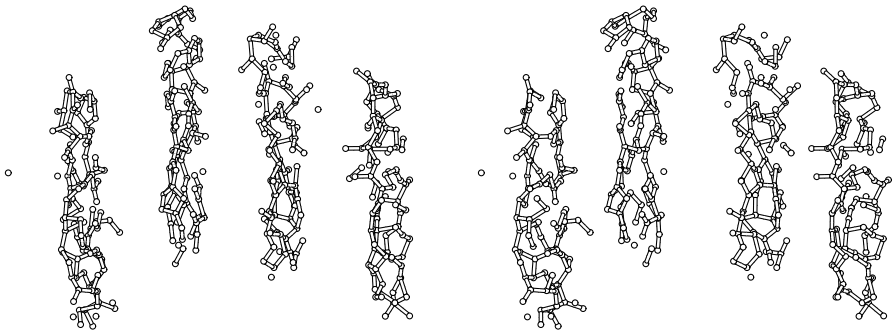

Figure 5.3: *Stereo-view of a ball-and-stick representation of an optimized structure after 1,000 steps of conditional dynamics with a MLF crystallographic target function, showing a typical condensation into four rod-like structures.*

condensation of the random atom distributions into four rod-like structures is observed. In these rods, the lowest resolution feautures of the model have been accounted for, without yet forming α-helical structures. Optimizations against data where the four suspect intensities at low resolution were not replaced by calculated values did not yield this condensation behaviour. Also omitting these reflections from the data set gave optimizations without any observable condensation after the initial optimization cycle (results not shown). Three of the corrected reflections (with hkl = 020, 0 – 22 & 002) account for the strongest reflections in the data set. These three reflections appeared critical for the observed condensation behaviour, since condensation was also observed for optimizations where only the fourth suspect reflection (with hkl = 001) was omitted from the data or where its observed intensity was used (results not shown).

Of the 36 initial optimization runs, three runs did not yield optimized structures due to formation of highly branched structures requiring more computer memory than available. From the remaining models, 17 structures were selected that appeared to have optimized towards a common hand based on a comparison of the highest peak in the phased translation function of the optimized coordinates and of their inverse. These structures were positioned on a common origin and subjected to subsequent cycles of MLHL-refinement.

5.3.2 Quality of the phase probability estimates

In initial calculations, the phase quality of all individual models was estimated by calculating σ_A^a-values derived from figures of merit m_a of the phase restraint. With this procedure, two cycles of MLHL-refinement were performed. Figure 5.4 displays the resulting m_a and σ_A^a-estimates and their true values after condensation and after both cycles of phase-restrained refinement. Severe over-estimation of figures of merit m_a as well as σ_A-values was observed after one cycle of MLHL-refinement. Optimization with these over-estimated values resulted in even larger over-estimation after the second cycle. The resulting models did not show any α-helical structure and no significant phase improvement was observed (results not shown).

Alternatively, σ_A^i-estimates were calculated for each of the 17 models separately, based

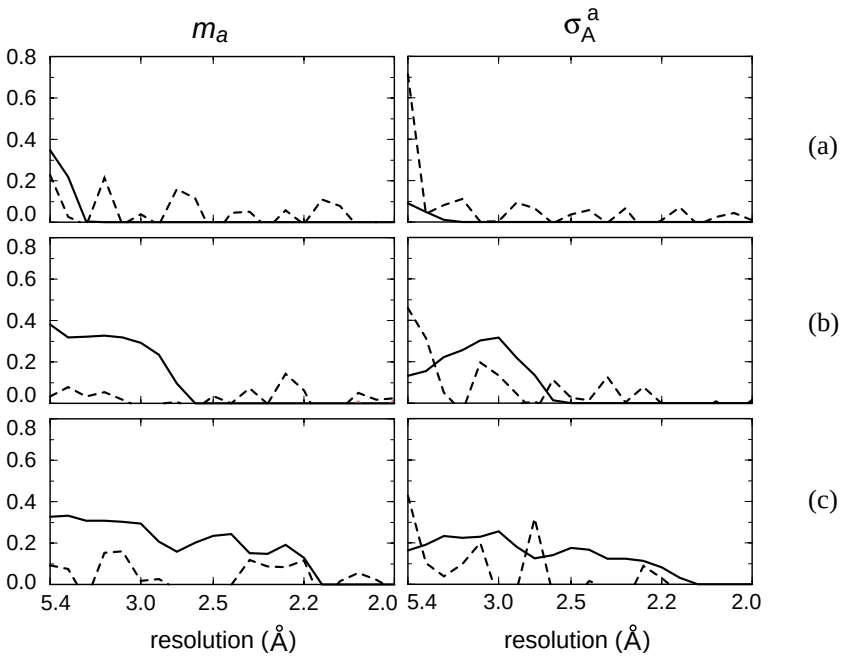

Figure 5.4: *Figures of merit m_a for the phase restraint (left) and σ_A^a for the individual models (right) after condensation (a) and after one (b) and two (c) cycles of MLHL-refinement. Estimated values are shown with solid lines as a function of resolution; their corresponding true values are shown with dashed lines. In this figure and in figures 5.5, 5.6 and 5.7, true values for the figure of merit and σ_A are calculated using the phases as calculated from the published atomic coordinates of Alpha-1.*

on observed phase differences between the individual structures. Fifteen cycles of MLHL-refinement were performed with σ_A^i-estimates. Figure 5.5 and 5.6 show shell-wise and overall estimates for m_a and σ_A^i and their corresponding true values throughout this run. During the first six cycles of refinement, estimates m_a for the figures of merit of the phase restraint corresponded rather well to the true cosine of the average phase error, but from cycle seven on an increasing over-estimation was observed. σ_A^i-Values were under-estimated during the first nine cycles. From cycle ten on, also these values were over-estimated.

An additional run was performed where the optimization with σ_A^i-estimates was resumed at cycle seven. In this calculation m_a-estimates obtained at cycle six were not updated anymore. With fixed estimates for m_a, eighteen additional cycles of MLHL-refinement were performed. Overall estimates for m_a and σ_A^i and their corresponding true values are shown in figure 5.7. As expected, the fixed figures of merit m_a were under-estimated. Also estimation of the phase quality of the individual structures by calculation of σ_A^i yielded under-estimated values throughout this run.

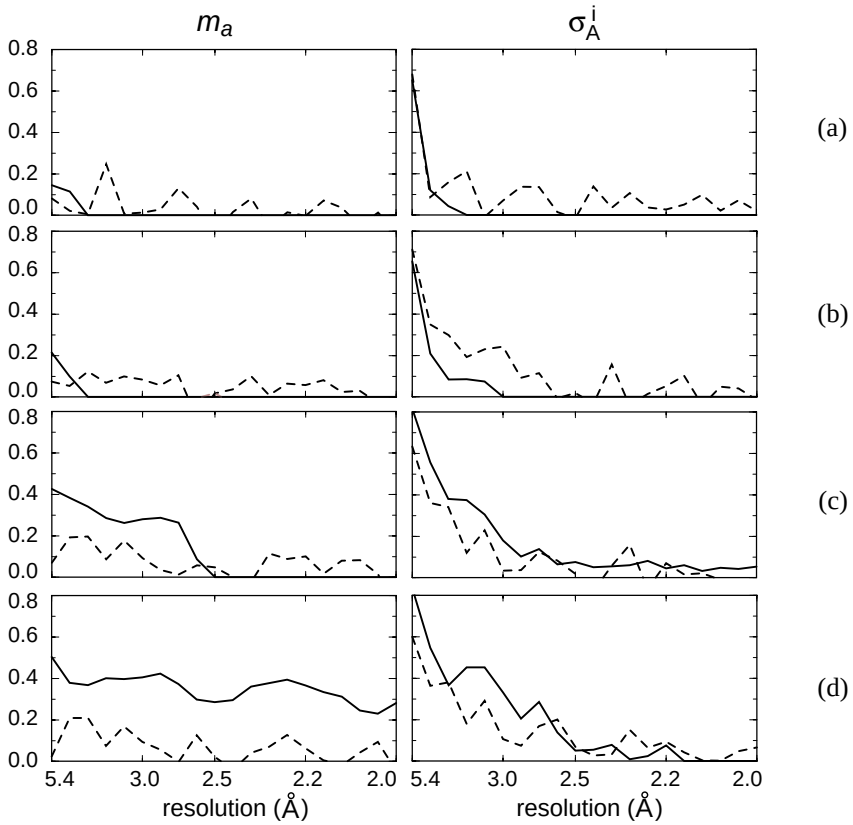

Figure 5.5: *Figures of merit m_a for the phase restraint (left) and σ_A^i for one of the individual models (right) after 2 (a), 6 (b), 11 (c) and 15 (d) cycles of MLHL-refinement. Estimated values are shown with solid lines as a function of resolution; their corresponding true values are shown with dashed lines.*

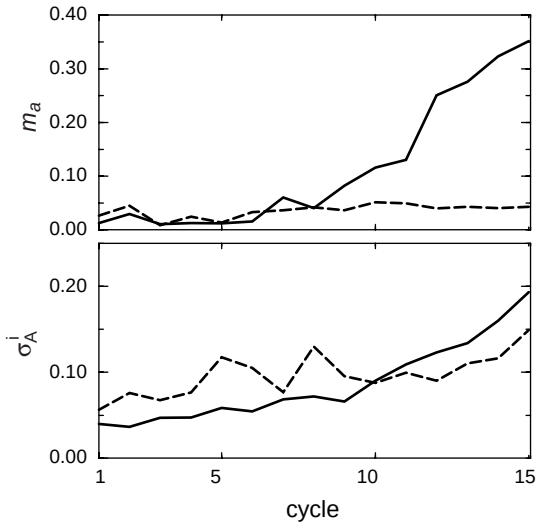

Figure 5.6: *Overall figures of merit m_a for the phase restraint (a) and σ_A^i-values for one of the individual models (b) in the optimization with m_a-estimates that were updated every cycle. Estimated values are shown with solid lines; their corresponding true values with dashed lines.*

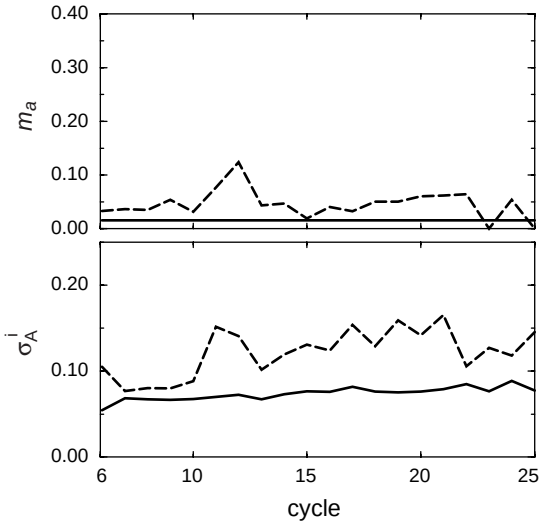

Figure 5.7: *Overall figures of merit m_a for the phase restraint (a) and σ_A^i-values for one of the individual models (b) in the optimization with fixed m_a-estimates after cycle 7. Estimated values are shown with solid lines; their corresponding true values with dashed lines.*

	best structure	worst structure	average
map ccf.	0.36	0.18	0.37
$\Delta\varphi$ ($^\circ$)	74.3	86.1	76.3
$\cos(\Delta\varphi)$	0.16	0.03	0.12
rmsd (Å)	1.54	1.74	-

Table 5.1: *Overall quality criteria for the best and the worst structure and for the average over the 17 individual structure factor sets after 25 optimization cycles. Map correlation coefficients (map ccf.) and phase errors ($|F^{\rm obs}|$-weighted $\Delta\varphi$ and unweighted $\cos(\Delta\varphi)$) were calculated using phases from the published coordinates. Root-mean-square coordinate errors (rmsd) were calculated as the nearest distances from atoms in the optimized structures to any of the atoms in the published structure.*

5.3.3 Convergence behaviour

An improvement in map correlation coefficients and phase errors was observed for both optimization runs with σ_A^i-estimates (see figure 5.8). Fastest convergence was observed for the run with fixed, under-estimated values for figures of merit m_a of the phase restraint. For this run a steady increase in average map correlation coefficient (of $\sim$0.005 per cycle) was observed throughout the optimization, as well as a decrease in the values of the overall phase errors. In the run where m_a-estimates were updated every cycle, over-estimation of the phase probabilities coincided with a significantly slower improvement in map quality.

After cycle 25 of the run with fixed figures of merit m_a, the individual structures with the best and worst map correlation coefficients could be identified by their overall σ_A^i-estimates (see figure 5.9). In the best structure (figure 5.10), three and a half α-helices have been formed, of which two in the correct orientation and one and a half with a reversed chain direction. The worst structure (figure 5.11) shows multiple small α-helical fragments, of which most with incorrect orientations. Overall quality criteria for these two structures and for the average over all 17 individual structure factor sets are shown in table 5.1. Map correlation coefficients for the average map $m|F^{\rm obs}|\exp(i\varphi^{\rm ave})$ tend to be better than for the best of the maps calculated with the phases of individual structure factor sets F^i. The average electron density map at cycle 25 is shown in figure 5.12a. In this map, two right-handed helices are clearly visible and two helical-like rods with a less distinct choice of hand are observed. Map correlation coefficients and phase errors for this map as a function of resolution are shown in figure 5.13. An average map calculated without the four suspect reflections (figure 5.12b) shows more electron density for the side chains.

The calculations presented here took in total approximately 115 CPU-days. Computer memory was allocated up to a maximum of 1.5 Gb.

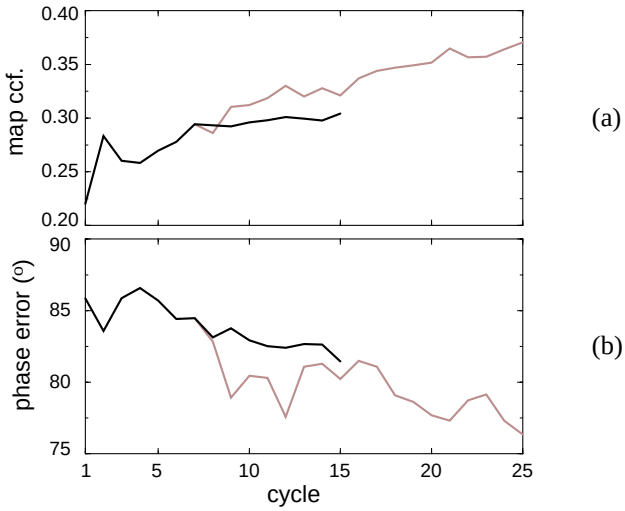

Figure 5.8: *Map correlation coefficients (ccf.) (a) and $|F^{obs}|$-weighted phase errors (b) to 2.0 Å resolution of F^{ave} with respect to phases calculated from the published coordinates for every optimization cycle. In black the results are shown for the optimization with σ_A^i-estimates where m_a-estimates were updated every cycle. In gray the results are shown for the optimization with σ_A^i-estimates and fixed m_a-estimates after cycle 7.*

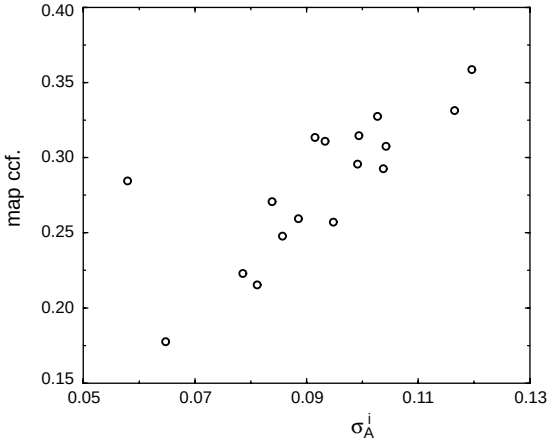

Figure 5.9: *Map correlation coefficients for the 17 individual structures obtained after 25 cycles of conditional optimization with σ_A^i-estimates and fixed values of m_a after cycle 7, plotted against their overall σ_A^i-estimates to 2.0 Å resolution. A correlation coefficient of 0.77 was observed between these values.*

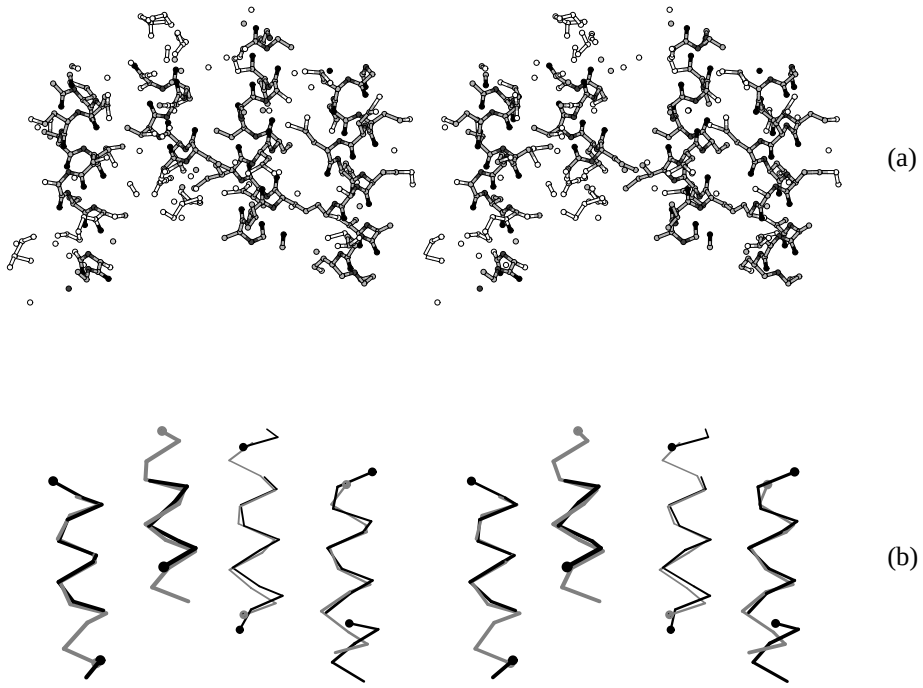

Figure 5.10: *Stereo-views of the best structure based on map correlation coefficients, obtained after 25 cycles of conditional optimization with* σ_A^i*-estimates and fixed values of* m_a *after cycle 7. (a) A ball-and-stick representation with automatic assignment of atom types based on gradient contributions from the conditional force field (white, unassigned; light gray, carbon; dark gray, nitrogen; black, oxygen). Atoms within a distance of 1.8 Å are connected. (b) A backbone trace between the assigned* C^α*-atoms (black), superimposed on the backbone trace of the target structure (gray). A sphere marks the N-terminal* C^α*-atoms of all fragments.*

5.4 Discussion

5.4.1 The Alpha-1 test case

After the first cycle of MLF-refinement, condensation of the random atom distributions into four rod-like structures was observed. The lowest resolution features of the model were accounted for in these models and condensation was considered to be favourable for the subsequent cycles of MLHL-refinement. This condensation behaviour may be attributed to strong reflections at low resolution, which indicate a bias away from uniform random atom distributions (as was already pointed out by Bricogne, 1993). For the three strongest reflections in the applied dataset, model structure factor amplitudes were used instead of observed values. These low-resolution reflections showed a large discrepancy between observed and calculated intensities, as well as a large measurement error. Correction of these reflections appeared critical for condensation, indicating the importance of strong reflections at low resolution. A fourth reflection was corrected (with *hkl* = 001), which had a lower calculated

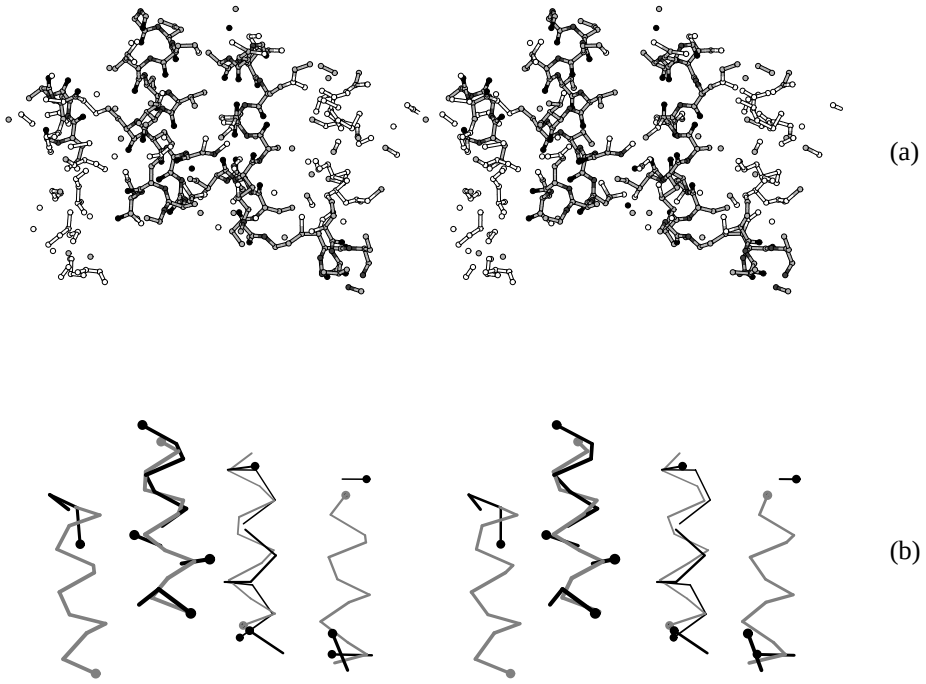

Figure 5.11: *Stereo-views of the worst structure based on map correlation coefficients, obtained after 25 cycles of conditional optimization with σ_A^i-estimates and fixed values of m_a after cycle 7. (a) A ball-and-stick representation with automatic assignment of atom types based on gradient contributions from the conditional force field (white, unassigned; light gray, carbon; dark gray, nitrogen; black, oxygen). Atoms within a distance of 1.8 Å are connected. (b) A backbone trace between the assigned C^α-atoms (black), superimposed on the backbone trace of the target structure (gray). A sphere marks the N-terminal C^α-atoms of all fragments.*

intensity and the discrepancy between the observed and calculated values was smaller. Correction of this reflection appeared not to be critical for condensation. Regarding the low measurement error of this reflection, correction may not have been justified. Final electron density maps calculated without the four suspect reflections showed more side-chain density than maps including the corrected reflections, indicating that the corrected intensities may have been too high.

Estimation of reliable phase probabilities is a critical factor in optimization of random atom distributions. Because standard procedures fail for models of such low phase quality, figures of merit for the phase restraint and σ_A-values for the individual models were estimated from the distribution of multiple models. Iterative estimation of phase probabilities has the risk of introducing bias. Even when using cross-validation in the calculations presented here, iterative estimation of the figures of merit leads to over-estimation of the phase probability of the average structure factor. Over-estimation of the figures of merit coincided with a significantly slower rate of convergence. Fastest convergence was obtained by keeping the figures of

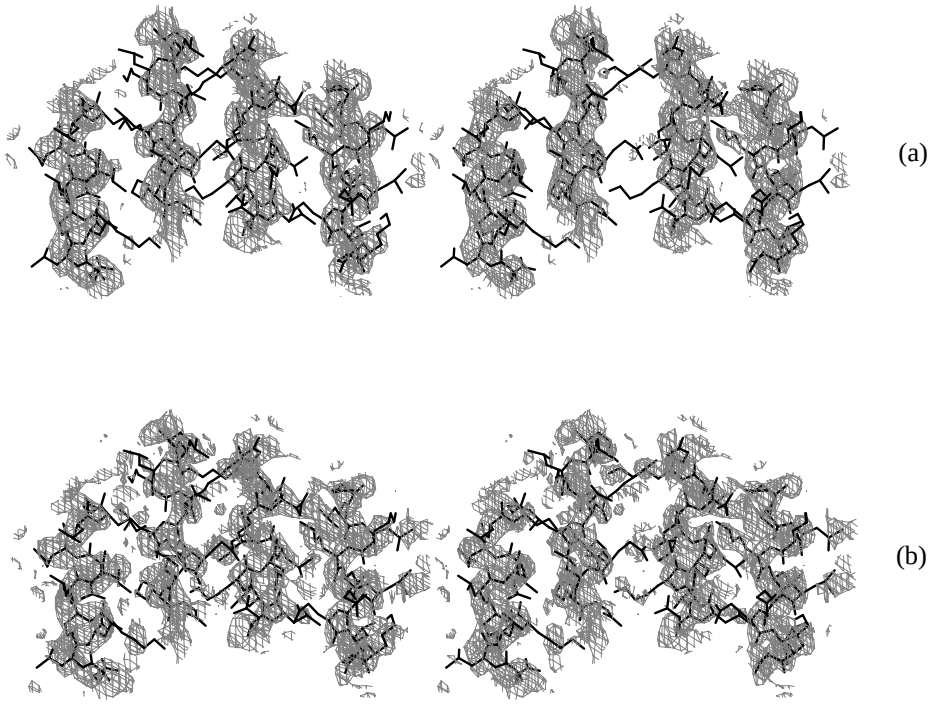

Figure 5.12: *Electron density maps ($m|F^{\mathrm{obs}}|\exp(i\varphi^{\mathrm{ave}})$) after 25 cycles of conditional optimization with σ_A^i-estimates and fixed values of m_a after cycle 7, (a) including calculated intensities for the four suspect low-resolution reflections and (b) excluding these reflections from the map calculation.*

merit fixed at under-estimated values, indicating that the procedure to iteratively estimate figures of merit requires further investigation. Two ways to estimate σ_A-values for the individual structures were tested. From the figures of merit of the average structure factors σ_A^a-estimates were derived for all structures. Although with this procedure significant phase improvements had been obtained before (Scheres & Gros, 2003), here it lead to a fast introduction of bias. Better results were obtained with a second procedure where σ_A^i-estimates were calculated for every structure based on the average cosine of the phase differences between that structure and all other structures. All reflections were used for this calculation. After 25 optimization cycles a correlation coefficient of 0.77 was observed between the σ_A^i-estimates and the map correlation coefficients of all individual structures. Also between the average cosine of the mutual phase differences and the cosine of the true phase errors of the individual structures a strong correlation was observed (with a correlation coefficient of 0.83, results not shown). This illustrates that the σ_A^i-estimates allow a relevant differentiation in phase quality of the individual structures.

After 25 cycles of MLHL-refinement an average electron density map with a correlation coefficient to the target map of 0.37 up to 2.0 Å resolution was obtained. This map may have allowed manual building of the four-helix bundle. Also, the best individual model could be

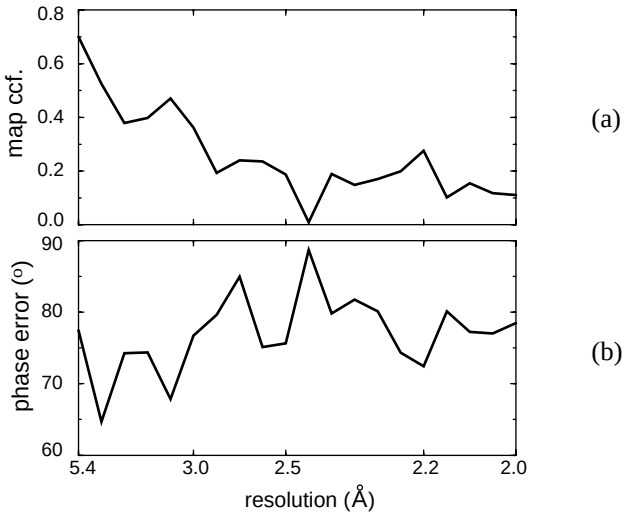

Figure 5.13: *Map correlation coefficients (ccf.) (a) and $|F^{\mathrm{obs}}|$-weighted phase errors (b) of F^{ave} as a function of resolution after 25 cycles of conditional optimization with σ_A^i-estimates and fixed values of m_a after cycle 7.*

identified by its σ_A-estimates and this model consisted of three and a half α-helices. However, these are no arguments that the structure was solved by conditional optimization. Taking the prior knowledge into account that the structure consists of α-helices, it may have been solved based on the lowest resolution reflections alone or after the initial condensation step. The steadily improving map correlation coefficients and phase errors during the subsequent 25 refinement cycles indicate that *ab initio* structure determination by conditional optimization may be possible for this test case. Progress was very slow: the average map correlation coefficient increased with 0.005 per cycle, whereas each cycle took approximately three CPU-days. With an r.m.s. coordinate error of 1.54 Å for the best structure and a phase error of 76.3° for the average structure factor set, the optimization process clearly was not finished yet. Still, without any prior phase information, conditional optimization yielded apparently meaningful gradients resulting in a set of models that was significantly better than the initial random atom distributions.

5.4.2 Implications for further development

In several aspects, the test case presented here may have been favourable for *ab initio* phasing by conditional optimization compared to other cases. The protein consists of four α-helices of near-ideal geometry. These helices are described accurately by the applied force field and the information content of the force field is higher for α-helices than for β-strands or loops. Besides, helices have a large chiral volume compared to β-strands and loops. This chirality is modelled in our approach and this breaks the ambiguity for the choice of hand. Furthermore, in the crystal these helices are arranged side-by-side in sheets spanning the

width of the crystal. This packing results in a few relatively strong low-resolution reflections. As mentioned above, such reflections are favourable for the observed condensation behaviour. Replacing the observed intensities with possibly too large calculated values may have further enhanced the effect of condensation. For those protein crystals where the packing does not result in such strong low-resolution reflections, initial condensation may be more difficult and subsequent optimization more cumbersome. On the other hand, the relatively small solvent region in this test structure leads to an unfavourable, low number of reflections compared to other protein structures. The effect of these contributions will have to be adressed in future calculations with other test cases.

Currently, the main limitations for further development of this method are the excessive CPU-time and the large amount of computer memory required for these calculations. This small test case took in total four months of CPU-time, which severely limits the number of variations that can be tested. Nevertheless, several possibilities for advances exist. As mentioned before, the estimation of reliable phase probabilities is crucial. Iterative estimation of figures of merit for the phase restraint lead to over-estimation, and further development of this procedure is needed. Promising results were obtained with estimation of phase quality for each individual structure. In analogy to procedures developed by Lunin *et al.* (2000), the observed correlation between estimated σ_A-values and the true quality of the individual structures may be exploited in procedures to enrich the average structure factors of the phase restraints. Furthermore, the protocol applied in the calculations presented here consisted of continuous optimization steps alone. Possibly, the introduction of discrete steps in the optimization process may allow a more readily escape from local minima, like for example wrongly oriented α-helices. Possibilities include re-positioning of atoms based on various electron density maps or recognition of protein fragments among the distribution of loose atoms (as described in chapter 4). In this respect a challenge will probably lie in obtaining multiple models that differ in a statistically valid way, yielding reliable phase probability estimates.

Faster convergence may also be obtained by adjusting some of the procedures that were used in the presented calculations and which may not have been optimal. Calculated intensities were used for four suspect reflections at low resolution, and these values may have been too large. Preferably, complete and reliable data is used to test the full potentials of this method. In protein crystallographic data collection it is common practice to ignore the lowest resolution reflections, owing to experimental inconveniences. However, by a few modifications to the standard experiment these problems can be overcome and reliable low-resolution data can be collected on a home source (Evans *et al.*, 2000). Other points of interest are the applied procedures for temperature-factor scaling and determination of the weight on the crystallographic part of the target function. These procedures were transferred from calculations involving models with smaller coordinate errors than random atom distributions. Possibly, for models with such large errors other procedures may yield better results. Also the selection of structures with a common hand after condensation should be reconsidered, since in the early stages of optimization the hand appeared not to be fixed yet.

5.5 Conclusions

The results for the single test case presented here indicate that *ab initio* phasing of observed diffraction data by conditional optimization of random atom distributions may be possible. Although convergence was very slow, a steady improvement in map correlation coefficients and phase errors was obtained. The importance of low-resolution reflections and estimation of reliable phase probabilities were illustrated. Correction of the three strongest, low-resolution reflections appeared crucial for condensation of the random starting models into rod-like structures. Under-estimation of phase probabilities yielded the best results in subsequent phase-restrained optimization cycles. Promising results were obtained with estimating different σ_A-estimates for each individual structure, based on phase differences between these structures. Iterative estimation of figures of merit for the average structure factors of the phase restraints gave over-estimated values, indicating that this procedure requires further examination. Further development of the applied procedures is currently limited by the excessive computational cost, but there are several possibilities for potential improvement. Hopefully, these may lead to a practical application of conditional optimization in *ab initio* protein structure determination.

Acknowledgements

This work is supported by the Netherlands Organization for Scientific Research (NWO-CW: Jonge Chemici 99-564).

Chapter 6

Summary & general discussion

In the standard crystallographic experiment, information about the phases of the reflections is lost. From the observed intensities alone, the electron density of the unit cell cannot be reconstructed and this problem is known as the crystallographic phase problem. At the resolution limits typically obtained in protein crystallography the phase problem is underdetermined and requires incorporation of additional information. Chapter 1 of this thesis describes existing, experimental and computational methods to determine and improve phases. The computational methods exploit prior knowledge about the content of the unit cell to supplement the limited experimental information. The quality of the available phase information typically determines how much prior knowledge can be expressed. In chapter 2 a novel refinement method is presented, called conditional optimization. The conditional formalism offers an N-particle solution for the assignment of topology to loose, unlabelled atoms. Thereby, conditional optimization allows expression of large amounts of geometrical knowledge about protein structures without the requirement of a molecular model, and thus potentially in the absence of phase information. Initial tests with a simplified structure and calculated data show that with this method, in principle, random atom distributions can be successfully optimized and *ab initio* phasing of medium-resolution data can be achieved. Conventional methods for the estimation of phase probabilities fail for models of very low phase quality and a novel multiple-model procedure to estimate σ_A-values was necessary for these optimizations. In chapter 3 a mean-force potential for conditional optimization of protein structures is presented, which expresses knowledge about common protein conformations like α-helices, β-strands and loops. Using this generally applicable parameter set, conditional optimization of three small protein structures against 2.0 Å observed diffraction data shows a large radius of convergence, validating the presented force field and illustrating the feasibility of the approach. The application of conditional optimization to automated model building is explored in chapter 4. For three test cases with data to medium resolution and good experimental phases, conditional optimization yields models of comparable quality as obtained by the commonly used programs *ARP/wARP* and *RESOLVE*. Chapter 5 describes the application of conditional optimization to *ab initio* phasing of observed diffraction data. Low-resolution reflections and reliable phase probability estimates appear to be important for convergence. For the presented test case promising results are obtained, indicating that also with observed

diffraction data successful optimization of random atom distributions may be possible.

Given the success of direct methods in small-molecule crystallography, a commonly applicable method for *ab initio* protein structure determination would be a valuable tool for structural biology. The results described in this thesis indicate that conditional optimization is a promising candidate for *ab initio* phasing of medium-resolution, protein diffraction data. On the other hand, major advances have been obtained with experimental approaches to solve protein structures. The possibility to incorporate seleno-methionine in protein molecules using bacterial expression systems has contributed significantly to the successful application of anomalous dispersion methods, and eukaryotic expression systems for this purpose are being developed. Although these methods still require biochemical modifications to the native protein molecule, techniques exploiting the anomalous signal of sulfur atoms in cysteines and methionines require only a single diffraction experiment on a native protein crystal. Currently, the weak anomalous signal of sulfur is still difficult to assess experimentally, but improvements in data collection and processing techniques may also render these experimental methods a serious alternative for *ab initio* approaches. In addition, also molecular replacement may be expected to become more generally applicable given the rapidly increasing number of different folds in the protein structure database.

All the approaches mentioned here share the ultimate goal of providing a commonly applicable solution to the phase problem in protein crystallography. Conditional optimization may benefit from advances in the other approaches and become suitable for common practice before reaching this ultimate goal. Starting from initial phases as provided by the other methods, the flexible searching behaviour with unlabelled atoms and the incorporation of prior geometric knowledge by the conditional formalism may be exploited to extend phase information in a commonly applicable way. As expressed in chapter 4 and 5, excessive computational costs form a stumble block for the applicability of conditional optimization. Parallelization of the program may be a way of dealing with this problem and such a development is underway. Common application of this method may then induce developments to further improve the applied procedures. Improvements in the rate of convergence may be expected from extension of the force field (especially concerning loops and side chains), optimization of hybrid models with explicit topology assignments for recognizable protein fragments, and development of discrete modelling steps in the optimization protocols. Estimation of reliable phase probabilities for models of low phase quality is also a point of interest. Various multiple-model procedures have been postulated in this thesis, but none of them proved generally applicable and further investigation in this direction is needed. Eventually, also the goal of *ab initio* phasing by conditional optimization would favour from these developments. From an optimistic point of view, and given the promising results described in chapter 2 and 5, this method may then become, perhaps among others, a commonly applicable tool for protein structure determination.

Bibliography

Abrahams, J.P. & de Graaff, R.A.G. (1998) *Curr. Opin. Struct. Biol.* **8**, 601-605.

Adams, P.D., Pannu, N.S., Read, R.J. & Brunger, A.T. (1997) *Proc. Natl. Acad. Sci. USA* **94**, 5018-5023

Ban, N., Nissen, P., Hansen, J., Moore, P.B. & Steitz, T.A. (2000) *Science* **289**, 905-920.

Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J.D. & Zardecki, C. (2002) *Acta Cryst.* D**58**, 899-907.

Bricogne, G. (1993) *Acta Cryst.* D**49**, 37-60.

Bricogne, G. (1997) *Methods Enzymol.* **276**, 361-423.

Brünger, A.T. (1993) *Acta Cryst.* D**49**, 24-36.

Brünger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S., Read, R.J., Rice, L.M., Simonson, T. & Warren, G.L. (1998a) *Acta Cryst.* D**54**, 905-921.

Brünger, A.T., Adams, P.D. & Rice, L.M. (1998b) *Curr. Opin. Struct. Biol.* **8**, 606-611.

Cowtan, K. (1998) *Acta Cryst. D***54**, 750-756.

Cusack, S., Belrhali H., Bram, A., Burghammer, M., Perrakis, A. & Riekel, C. (1998) *Nat. Struct. Biol.* **5**, 634 - 637.

Drenth, J. (1999) *Principles of Protein X-ray Crystallography*, 2 *nd* ed., Heidelberg: Springer.

Evans, G., Roversi, P. & Bricogne, G. (2000) *Acta Cryst.* D**56**, 1304-1311.

Foadi, J., Woolfson, M., Dodson, E. J., Wilson, K. S., Jia-xing, Y. & Chao-de, Z. (2000) *Acta Cryst.* D**56**, 1137-1147.

Gilliland, G.L. & Ladner, J.E. (1996) *Curr. Opin. Struc. Biol* **6**, 595-603.

Glykos, N.M. & Kokkindis, M. (2000) *Acta Cryst.* D**56**, 169-174.

Guo, D.Y., Blessing, R.H. & Langs, D.A. (2000) *Acta Cryst.* D**56**, 1148-1155.

Harris, G.W. (1995) *Acta Cryst.* D**51**, 695-702.

Hauptman, H. (1997) *Curr. Opin. Struc. Biol.* **7**, 672-680.

Hendrickson, W.A. & Ogata, C.M. (1997) *Methods Enzymol.* **276**, 494-523.

Holton, T., Ioerger, T.R., Christopher, J.A. & Sacchettini, J.C. (2000) *Acta Cryst.* D**56**, 722-734.

Huizinga, E.G., Plas van der, R.M., Kroon, J., Sixma, J.J. & Gros, P. (1997) *Structure* **5**, 1147-1156.

Huizinga E.G., Schouten, A., Connolly, T.M., Kroon, J., Sixma, J.J. & Gros, P. (2001) *Acta Cryst.* D**57**, 1071-1078.

Hauptman, H. (1997) *Curr. Opin. Struc. Biol.* **7**, 672-680.

Isaacs, R.C. & Agarwal, N.W. (1977). *Proc. Natl. Acad. Sci. USA* **74**, 2835-2839.

Jones, T.A., Zou, J.-Y., Cowan, S.W. & Kjeldgaard, M. (1991) *Acta Cryst.* A**47**, 110-119.

Ke, K. (1997). *Methods Enzymol.* **276**, 448-461.

Kissinger, C.R., Gehlhaar, D.K. & Fogel, D.B. (1999) *Acta Cryst.* D**55**, 484-491.

Lamzin, V.S. & Wilson, K.S. (1993) *Acta Cryst.* D**49**, 129-147.

Lamzin, V.S., Perrakis, A., Bricogne, G., Jiang, J., Swaminathan, S. & Sussman, J.L. (2000) *Acta Cryst.* D**56**, 1510-1511.

Lunin, V.Y., Lunina, N.L., Petrova, T.E., Vernoslova, E.A., Urzhumtsev, A.G. & Podjarny, A.D. (1995) *Acta Cryst.* D**51**, 896-903.

Lunin, V.Y., Lunina, N.L., Petrova, T.E., Urzhumtsev, A.G., & Podjarny, A.D. (1998) *Acta Cryst.* D**51**, 896-903.

Lunin, V.Y., Lunina, N.L., Petrova, T.E., Skovoroda, T.P., Urzhumtsev, A.G. & Podjarny, A.D. (2000) *Acta Cryst.* D**56**, 1223-1232.

Lunina, N.L., Lunin., V.Y. & Urzhumtsev, A.G. (2000) ECM-19 abstracts, *XIXth European Crystallographic Meeting*, 25-31 August 2000, Nancy, France

Morris, R.J., Perrakis, A. & Lamzin, V.S. (2002) *Acta Cryst.* D**58**, 968-975.

Mowbray, S.L., Helgstrand, C., Sigrell, J.A., Cameron, A.D. & Jones, T.A. (1999) *Acta Cryst.* D**55**, 1309-1319.

Murshudov, G.N., Vagin, A.A. & Dodson, E.J. (1997) *Acta Cryst.* D**53**, 240-255.

Navaza, J. & Saludjian P. (1997) *Methods Enzym.* **276**, 581-595.

Ogata, C.M. (1998) *Nat. Struc. Biol.* **5**, 638-640.

Oldfield, T. (2000) *Crystallographic Computing 7*: Macromolecular Crystallographic Data. Edited by Bourne, P.E., Watenpaugh, K., Oxford: Oxford University Press.

Pannu, N.S. & Read, R.J. (1996). *Acta Cryst.* A**52**, 659-668.

Pannu, N.S., Murshudov, G.N., Dodson, E.J. & Read, R.J. (1998) *Acta Cryst.* D**54**, 1285-1294.

Perrakis, A., Morris, R. & Lamzin, V.S. (1999) *Nat. Struc. Biol.* **6**, 458-463.

Privé, G.G., Anderson, D.H., Wesson, L., Cascio, D. & Eisenberg, D. (1999). *Protein Sci.* **8**, 1400-1409.

Read, R.J. (1986) *Acta Cryst.* A**42**, 140-149.

Read, R.J. (2001) *Acta Cryst.* D**57**, 1373-1382.

Rice, L.M., Shamoo, Y. & Brunger, A.T. (1998) *J. Appl. Cryst.* **31**, 798-805.

Rossmann, M.G. (1990) *Acta Cryst.* A**46**, 73-82.

Rossmann, M.G. (1995) *Curr. Opin. Struc. Biol.* **5**, 650-655.

Rossmann, M.G. (2001) *Acta Cryst.* D**57**, 1360-1366.

Schomaker, V. & Trueblood, K.N. (1968) *Acta Cryst.* B**24**, 63-76.

Scheres, S.H.W. & Gros, P. (2001), Chapter 2 of this thesis, *Acta Cryst.* D**57**, 1820-1828.

Scheres, S.H.W. & Gros, P. (2003), Chapter 3 of this thesis, *Acta Cryst.* D**59**, 438-446.

Sheldrick, G.M. & Gould, R.O. (1995) *Acta Cryst.* B**51**, 423-431.

Sheriff S., Klei H.E. & Davis, M.E. (1999) *J. Appl. Cryst.* **32**, 98-101.

Sippl, M.J. (1995) *Curr. Opin. Struct. Biol.* **5**, 229-235

Srinivasan, R. & Parthasaraty, S. (1976). *Some Statistical Applications in X-ray Crystallography*. Oxford: Pergamon Press.

Stevens, R.C. (2000) *Curr. Opin. Struc. Biol.* **10**, 558-563.

Subbiah, S. (1991) *Science* **252**, 128-133.

Subbiah, S. (1993) *Acta Cryst.* D**49**, 108-119.

Tame, J.R.H. (2000) *Acta Cryst.* D**56**, 1554-1559.

Terwilliger, T.C. & Berendzen, J. (1999) *Acta Cryst.* D**55**, 849-861.

Terwilliger, T.C. (2000) *Acta Cryst.* D**56**, 965-972.

Terwilliger, T.C. (2001) *Acta Cryst.* D**57**, 1763-1775.

Terwilliger, T.C. (2002) *Acta Cryst.* D**58**, 1937-1940.

Urzhumtsev, A.G., Lunina, N.L., Skovoroda, T.P., Podjarny, A.D. & Lunin, V.Y. (2000) *Acta Cryst.* D**56**, 1233-1244.

Vriend, G., Sander, C. & Stouten, P. F. (1994) *Protein Eng.* **7**, 23-29.

Wang, B.C. (1985) *Methods Enzymol.* **115**, 90-112.

Weeks, C.M., DeTitta, G.T., Miller, R. & Hauptman, H.A. (1993) *Acta Cryst.* D**49**, 179-181.

Winn, M.D., Isupov, M.N. & Murshudov, G.N. (2001) *Acta Cryst.* D**57**, 122-133.

Samenvatting & algemene discussie

In het standaard kristallografisch experiment gaat alle informatie over de fasen van de reflecties verloren. De electronendichtheid in de eenheidscel kan met enkel de gemeten intensiteiten van de reflecties niet gereconstrueerd worden. Dit staat in de kristallografie bekend als het fasenprobleem. Bij de resolutielimieten die kenmerkend zijn voor de eiwitkristallografie is het fasenprobleem onderbepaald. Het oplossen ervan is daarom afhankelijk van additionele informatie. Hoofdstuk 1 van dit proefschrift beschrijft bestaande, experimentele en rekenkundige methoden voor het bepalen en verbeteren van de fasen. De rekenkundige methoden maken gebruik van vooraf beschikbare kennis over de inhoud van de eenheidscel om de gebrekkige experimentele informatie aan te vullen. Kenmerkend hierbij is dat de kwaliteit van de faseninformatie bepaalt hoeveel van die kennis gebruikt kan worden. In hoofdstuk 2 wordt een nieuwe verfijningsmethode geïntroduceerd: conditionele optimalisatie. Het conditionele formalisme biedt een N-deeltjes oplossing voor het toekennen van een topologie aan losse, ongelabelde atomen. Daarmee kunnen grote hoeveelheden kennis over de geometrie van eiwitstructuren tot uitdrukking worden gebracht zonder gebruik te maken van een moleculair model en dus mogelijk ook zonder enige faseninformatie. Aanvankelijke tests met een vereenvoudigde structuur en berekende data toonden aan dat het in principe mogelijk is om met deze methode random verdelingen van atomen succesvol te verfijnen en daarmee data tot medium resolutie *ab initio* te faseren. Conventionele methoden om fasenwaarschijnlijkheden af te schatten werken niet met modellen die zo weinig faseninformatie bevatten. Daarom was het noodzakelijk een nieuwe procedure te ontwikkelen die gebruik maakt van de spreiding in meerdere modellen. In hoofdstuk 3 wordt een krachtenveld voor conditionele optimalisatie van eiwitmoleculen gepresenteerd. Dit krachtenveld bevat kennis over veel voorkomende eiwitconformaties zoals α-helices, β-strands en loops. Met dit algemeen toepasbare krachtenveld werd een grote convergentiestraal bereikt voor conditionele optimalisatie van drie kleine eiwitmoleculen met experimentele diffractiedata. Hiermee werd het krachtenveld gevalideerd en de haalbaarheid van de methode aangetoond. In hoofdstuk 4 wordt de toepassing van deze methode op het geautomatiseerd bouwen van eiwitmoleculen onderzocht. Voor drie testgevallen met data tot medium resolutie en goede experimentele fasen leverde conditionele optimalisatie een model op van vergelijkbare kwaliteit als de veelgebruikte programma's *ARP/wARP* en *RESOLVE*. Hoofdstuk 5 beschrijft de toepassing van conditionele optimalisatie op het *ab initio* faseren van experimentele diffractiedata. Lage

resolutie reflecties en betrouwbare fasenwaarschijnlijkheden blijken hiervoor van belang te zijn. Met het gepresenteerde testgeval werden veelbelovende resultaten behaald, wat aantoont dat succesvolle optimalisatie van random atoomverdelingen met experimentele data mogelijk is.

Gegeven het succes van directe methoden in de kleine-stofjes kristallografie zou een algemeen toepasbare methode voor *ab initio* structuurbepaling van eiwitten een waardevol instrument zijn voor de structuurbiologie. De resultaten die in dit proefschrift staan beschreven tonen aan dat conditionele optimalisatie een veelbelovende kandidaat is voor het *ab initio* faseren van medium-resolutie, eiwitdiffractiedata. Van de andere kant is er ook grote vooruitgang geboekt met experimentele methoden voor eiwitstructuurbepaling. De mogelijkheid om met bacteriële expressiesystemen seleniummethionine in eiwitmoleculen in te bouwen heeft significant bijgedragen aan de succesvolle toepassing van methoden die gebruik maken van anomale dispersie. Momenteel worden er ook eukaryote expressiesystemen voor deze doeleinden ontwikkeld. Deze methoden vereisen echter biochemische modificatie van het natieve eiwitmolecuul. Technieken die gebruik maken van het anomale signaal van zwavelatomen in cysteine en methione vereisen daarentegen slechts één diffractieëxperiment aan een natief eiwitkristal. Momenteel is het nog moeizaam het kleine anomale signaal van zwavel experimenteel te bepalen, maar door verbeteringen in data opname- en verwerkingstechnieken kunnen dit soort experimentele methoden een serieus alternatief gaan vormen voor *ab initio* technieken. Daarnaast kan ook van de *molecular replacement* techniek worden verwacht dat ze algemener toepasbaar wordt, gezien het snel toenemend aantal verschillende eiwitvouwingen in de databank van eiwitstructuren.

De bovenstaande technieken delen het ultieme doel om een algemeen toepasbare oplossing voor het fasenprobleem in de eiwitkristallografie te leveren. Conditionele optimalisatie kan haar voordeel doen met vorderingen die worden gemaakt met de overige benaderingen. Initiële faseninformatie die is verkregen met een van de andere technieken kan met conditionele optimalisatie uitgebreid worden, gebruik makende van het flexibele zoekgedrag met ongelabelde atomen en de benutte kennis over de eiwitgeometrie. Daarmee kan deze methode al algemeen toepasbaar worden voordat haar ultieme doel is bereikt. Zoals reeds vermeld in hoofdstuk 4 en 5 vormen excessieve rekentijden en geheugengebruik een struikelblok voor algemene toepassing van conditionele optimalisatie. Parallellisatie van het programma zou hierbij kunnen helpen en momenteel wordt daaraan gewerkt. Versnelde convergentie zou bereikt kunnen worden door uitbreiding van het krachtenveld (met name wat betreft loops en zijketens), optimalisatie van hybride modellen met expliciete toekenning van topologie aan herkenbare eiwitfragmenten en de ontwikkeling van discrete modelbouwstappen in de optimalisatieprotocollen. Het afschatten van betrouwbare fasenwaarschijnlijkheden voor modellen met weinig faseninformatie is hierbij ook van belang. In dit proefschrift zijn hiervoor verschillende procedures voorgesteld, maar geen enkele is algemeen toepasbaar gebleken. Daarom is verder onderzoek in deze richting noodzakelijk. Uiteindelijk zou ook het ultieme doel van *ab initio* fasering kunnen profiteren van deze ontwikkelingen. Vanuit een optimistisch standpunt en gegeven de veelbelovende resultaten beschreven in hoofdstuk 2 en 5, zou deze methode dan, misschien naast andere methoden, een algemeen toepasbaar instrument voor eiwitstructuurbepaling kunnen worden.

Curriculum Vitae

De auteur van dit proefschrift werd op 5 augustus 1975 geboren in Roermond. In 1993 voltooide hij zijn opleiding aan het Gymnasium van het Bisschoppelijk College in Weert en begon aan de opleiding Scheikunde van de Universiteit Utrecht. Aldaar werd binnen een jaar het propedeuse-diploma *cum laude* behaald. Het afstudeeronderzoek werd verricht onder begeleiding van dr. Graafsma bij de Europese faciliteit voor synchrotronstraling (ESRF) in Grenoble, Frankrijk. Op 31 maart 1998 werd in Utrecht het doctoraal-examen *cum laude* afgelegd. Een dag later trad hij als assistent-in-opleiding in dienst bij de vakgroep Kristal- en Structuurchemie van de Universiteit Utrecht onder begeleiding van prof. dr. Jan Kroon en dr. Piet Gros. Aanvankelijk werd getracht de kristalstructuur van het N-terminale domein van transcriptiecofactor PC4 op te helderen. Later werd het onderzoek verricht dat beschreven staat in dit proefschrift. Vanaf juni zal hij waarschijnlijk werkzaam zijn als post-doc in het laboratorium van dr. José María Carazo in Madrid, alwaar hij beeldverwerkingsmethoden voor de electronenmicroscopie gaat ontwikkelen.

Publicaties

- **S.H.W. Scheres** & P. Gros (2003) "Development of a force field for conditional optimization of protein structures", *Acta Cryst.* D**59**, 438-446.
- **S.H.W. Scheres** & P. Gros (2001) "Conditional Optimization: a new formalism for protein structure refinement", *Acta Cryst.* D**57**, 1820-1828.
- R.B.G. Ravelli, M.L. Raves, **S.H.W. Scheres**, A. Schouten & J. Kroon (1999) "*Ab initio* structure determination of low-molecular-weight compounds using synchrotron radiation Laue diffraction", *J. Synchrotron Rad.* **6**, 19-28.
- G.W.J.C. Heunen, A. Puig-Molina, **S.H.W. Scheres**, C. Schultze, D. Bourgeois & H. Graafsma (1998) "Broad energy band techniques for perturbation crystallography", Proceedings of SPIE, **3448**, 166-175.

Dankwoord

Beste Jan, jij was het die mij strikte voor een hoofdvak in Grenoble en daarmee voor de kristallografie. Dank voor al je vertrouwen en steun en dat je mijn promoter wilde zijn. Helaas maak jij mijn promotie niet meer mee. Je was (ook als onze professor) een fantastisch mens en aan jou draag ik dit boekje op. Piet, jij nam de stok van Jan over en ik bewonder de manier waarop je dat doet. Reeds daarvoor maakte je mij enthousiast voor jouw ideeën over ”conditional dynamics”. Bedankt voor al je begeleiding daarbij, al jouw enthousiasme en je eindeloze stroom van ideeën, van ’s ochtends voor negenen tot laat in de middag. Ik heb veel van je geleerd en ik denk dat we een goed team vormden. Ik hoop dat dit boekje zal bijdragen aan het verwezenlijken van jouw wetenschappelijke droom. Loes, mijn innig medeleven vanwege Jan gaat uit naar jou. Ook jou wil ik hartelijk danken, voor het samen tennissen, het altijd open staan voor vragen en je exacte antwoorden daarop. Marjan, jij bent de spil van de vakGROEP, bedankt daarvoor en voor al je hulp van de afgelopen jaren.

Ik heb mezelf al die tijd zeer gelukkig geprezen met een vakgroep als Kristal & Structuurchemie! Lucy en Clasien, met jullie ben ik daar samen aan een promotie-avontuur begonnen. Ik wil jullie hartelijk danken voor alle gezelligheid, steun en de vriendschap die wij al vijf jaar hebben mogen delen. Ik wens jullie veel succes met jullie laatste loodjes. Toen wij kwamen gingen Raimond en Jeroen net weg. Jeroen, ik nam aanvankelijk jouw PC4-project over en hoewel ons gezamenlijk verblijf op de vakgroep kort was wil ik je danken voor je hulp bij mijn opstarten. Jouw ziekte en overlijden hebben ons allemaal erg geschokt. Raimond, jij wijdde me in in de geheimen van het dataverwerken. Bedankt daarvoor, ik heb er altijd veel aan gehad. Ook de rest van de ’oude stempel’ wil ik hartelijk bedanken voor alle gezelligheid en collegialiteit in de vakgroep. Martin W. en Roeland (o.a. voor een geweldige West-Highland ervaring; jammer dat jullie allebei weer roken...), Stephane (Martin’s spätzles waren toch echt lekkerder), Wijnand (ook bedankt voor je gastvrijheid in Cambridge), Barend (spelletjes), Carien (blijven lachen), Ton L. (een biertje met een gulle lach), Diane (carnaval in Limburg), Anne (The Netherlands are so cold, but great fun), Jean (dank voor veel gezelligheid en je hulp met PC4), Ap (koffie-discussies) en Bouke (tennissen). Ik heb jullie allemaal best gemist toen jullie vertrokken waren. Gelukkig is er ook weer een nieuwe lichting enthousiaste mensen gekomen, waar ik steeds graag mee heb samengewerkt: Annika (dank voor gezelligheid op onze kamer), Nicole & Jenny (brug naar de BOC), Mitja (nu eindelijk op de fiets naar Primus), Allison (koffietijd!), Hans, Fin & Bert (een aangename invasie uit Wageningen) en Wiegert (van spinnen niet bang). En dan nog dank aan die mensen die er altijd al waren en zonder wie de vakgroep de vakgroep niet zou zijn: Arie-laat-mij-hier-maar-achter (voor veel lol in Sheffield, Nancy en bij talloze andere gelegenheden), Martin L. (voor

je grote behulpzaamheid bij vele computerproblemen), Eric (voor al je nuttige commentaren), Ton Spek (voor hulp bij kristallografische problemen), Huub (voor de lekkere koffie), Toine (voor al je steun met de workstations) en alle studenten. Jan, Ingrid en Aloys van de AV-dienst wil ik bedanken voor alle hulp bij de posters en de kaft van dit boekje. Janneke, bedankt voor je hulp bij mijn beursaanvragen. Van de NMR-groep wil ik verder Gert, Albert en Henri bedanken voor onmisbare hulp met de bacteriën, helaas kunnen jullie er in dit boekje niks over lezen. Ook andere leden van de BOC & NMR wil ik bedanken voor hun bijdragen aan gezellige borrels en avondjes in Primus en De Vooghel. Speciale dank ook aan BOC *Int.* voor veel volleybalplezier.

Piet, gelukkig heb je ongelijk gekregen en kan ik ook na vijf jaar promoveren nog veel mensen van buiten het lab bedanken voor hun vriendschap en voor mooie tijden: al mijn scheikundevriendjes voor veel gezelligheid en alle huisgenootjes voor het (t)huisgevoel in Huize Rembrandt. Verder dank aan alle vrienden en familie in Limburg: ook daar kom ik graag weer thuis. Bart & Cindy, dank voor jullie vriendschap en voor het aan Rosa laten zien hoe het in Limburg is. Alle minse van Scouting Tungelroy: hertstikke bedanktj veur hieël vùl plezeer. Auch alle minse van de bouw in Utrecht wil ich hertstikke bedanke veur vùl lol en veur os sjoeëne huuske: wae woeëne dur hieël gaer. Ook dank aan het Weert-clubje voor leuke VWO-etentjes (ik hoop binnenkort weer eens te kunnen!), Guido & Marie-Hélène voor hun vriendschap-op-afstand, Heinz o.a. voor hulp bij aanbevelingen en de mensen van de KCV, speciaal het jongerenteam + helpers, voor alle verdieping en steun.

Paco, Rosa y Fran, gracias por toda vuestra ayuda y hospitalidad en España. Despues de trabajar mucho siempre me he podido relajar con vosotros en San Lorenzo. Tambien gracias a todos nuestros amigos allí por esas noches de cachondeo. Mario, hartelijk dank voor onze mooie vriendschap en dat jij deze keer míjn paranimf wilt zijn. Waarvandaan je ook mag komen, onze deur blijft altijd voor je open. Marc, op zo'n broertje kun je rekenen. Bedankt dat je altijd klaar staat om te helpen, ook nu weer als paranimf. Succes met de voorbereidingen van jullie feest! Pa en ma, ook al vonden jullie het zelf moeilijk iets aan mijn proefschrift bij te dragen, jullie steun is altijd onmisbaar geweest! Bedankt voor alle (in)directe hulp bij mijn promotie en dat jullie altijd voor mij en Rosa klaar staan.

Rosita, tu me haces muy feliz. TQM. Juntos continuaremos por nuestro camino, donde sea. Ik ben blij dat we daarbij samen God mogen danken, voor elkaar en voor de dingen van iedere dag, ook een promotiedag.