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.

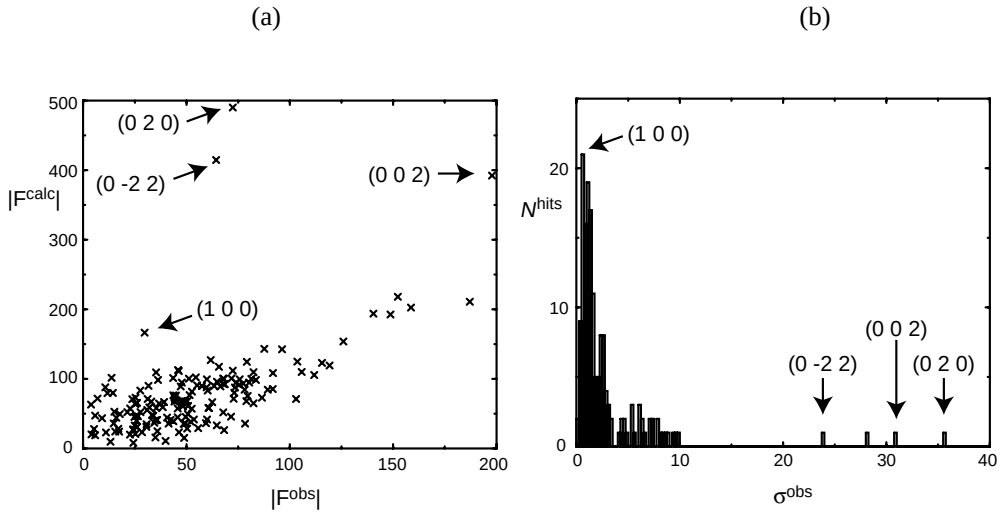

Figure 5.1: *(a) Calculated versus observed structure factor amplitudes for all observed reflections with Bragg spacing $d > 5$ Å. In the model structure factor calculation a bulk solvent contribution was taken into account using the standard mask method as implemented in CNS. Four suspect reflections 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.

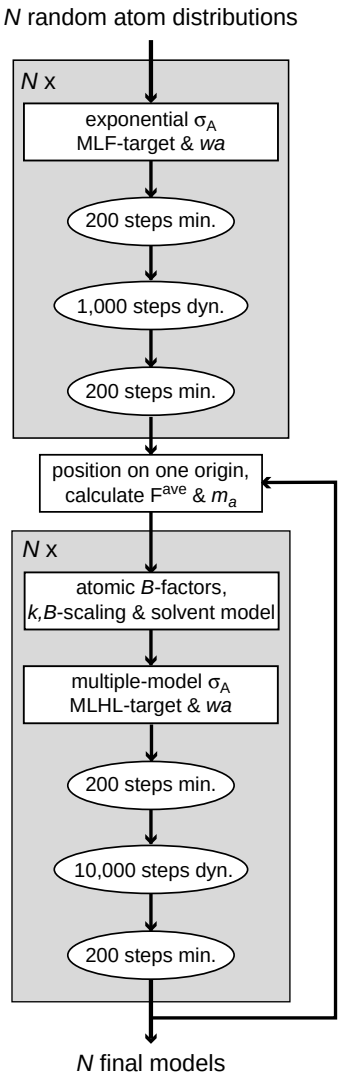

Figure 5.2: *Refinement protocol for ab initio phasing by conditional optimization. In every optimization cycle (gray) conditional dynamics coupled to a temperature bath of 600K (dyn.) was preceded and followed by energy minimization (min.). Positioning of the individual models on a common origin, calculation 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

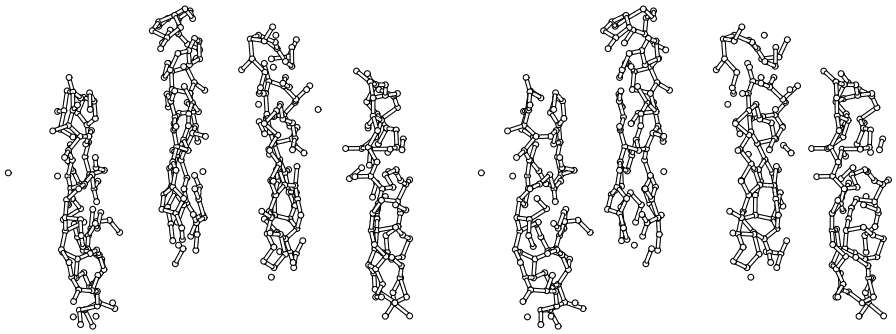

Figure 5.3: *Stereo-view of a ball-and-stick representation of an optimized structure after 1,000 steps of conditional dynamics with a MLF crystallographic target function, showing a typical condensation into four rod-like structures.*

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

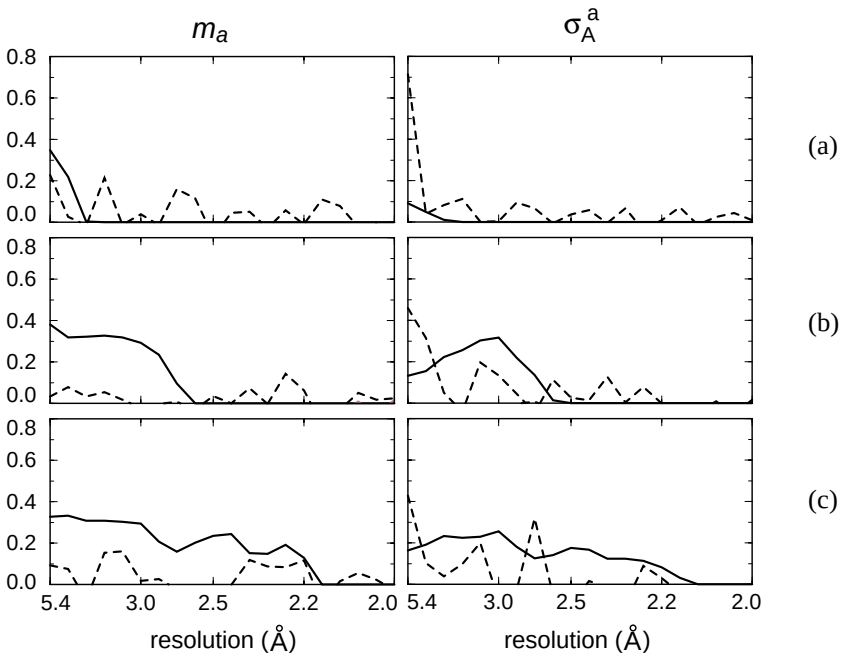

Figure 5.4: *Figures of merit m_a for the phase restraint (left) and σ_A^a for the individual models (right) after condensation (a) and after one (b) and two (c) cycles of MLHL-refinement. Estimated values are shown with solid lines as a function of resolution; their corresponding true values are shown 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.

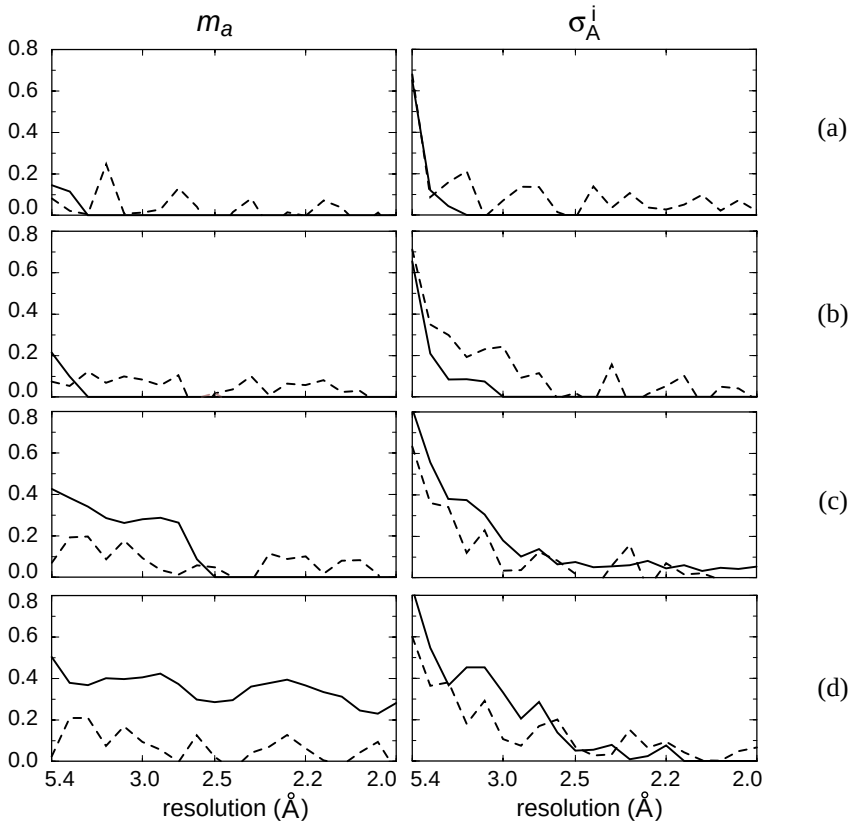

Figure 5.5: *Figures of merit m_a for the phase restraint (left) and σ_A^i for one of the individual models (right) after 2 (a), 6 (b), 11 (c) and 15 (d) cycles of MLHL-refinement. Estimated values are shown with solid lines as a function of resolution; their corresponding true values are shown with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

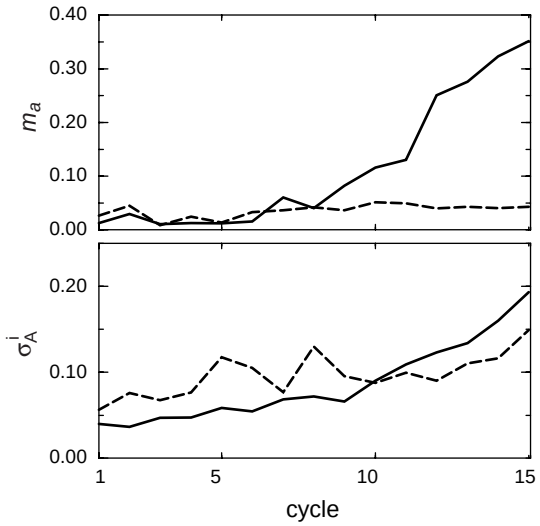

Figure 5.6: *Overall figures of merit m_a for the phase restraint (a) and σ_A^i-values for one of the individual models (b) in the optimization with m_a-estimates that were updated every cycle. Estimated values are shown with solid lines; their corresponding true values with dashed lines.*

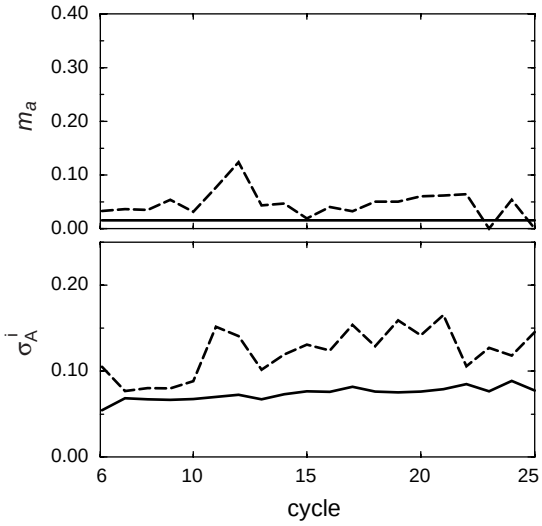

Figure 5.7: *Overall figures of merit m_a for the phase restraint (a) and σ_A^i-values for one of the individual models (b) in the optimization with fixed m_a-estimates after cycle 7. Estimated values are shown with solid lines; their corresponding true values with dashed lines.*

	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.

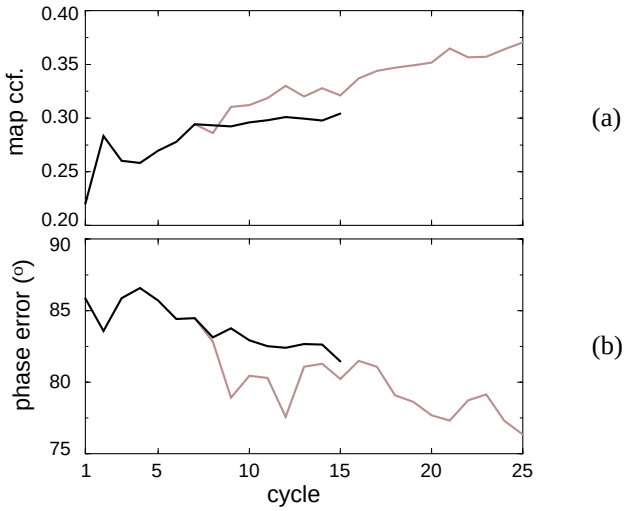

Figure 5.8: *Map correlation coefficients (ccf.) (a) and $|F^{obs}|$-weighted phase errors (b) to 2.0 Å resolution of F^{ave} with respect to phases calculated from the published coordinates for every optimization cycle. In black the results are shown for the optimization with σ_A^i-estimates where m_a-estimates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

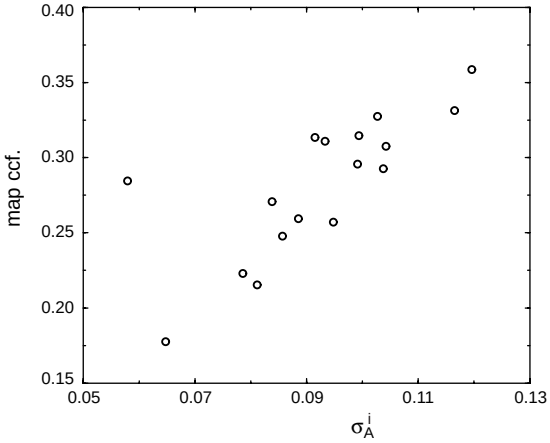

Figure 5.9: *Map correlation coefficients for the 17 individual structures obtained after 25 cycles of conditional optimization with σ_A^i-estimates and fixed values of m_a after cycle 7, plotted against their overall σ_A^i-estimates to 2.0 Å resolution. A correlation coefficient of 0.77 was observed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

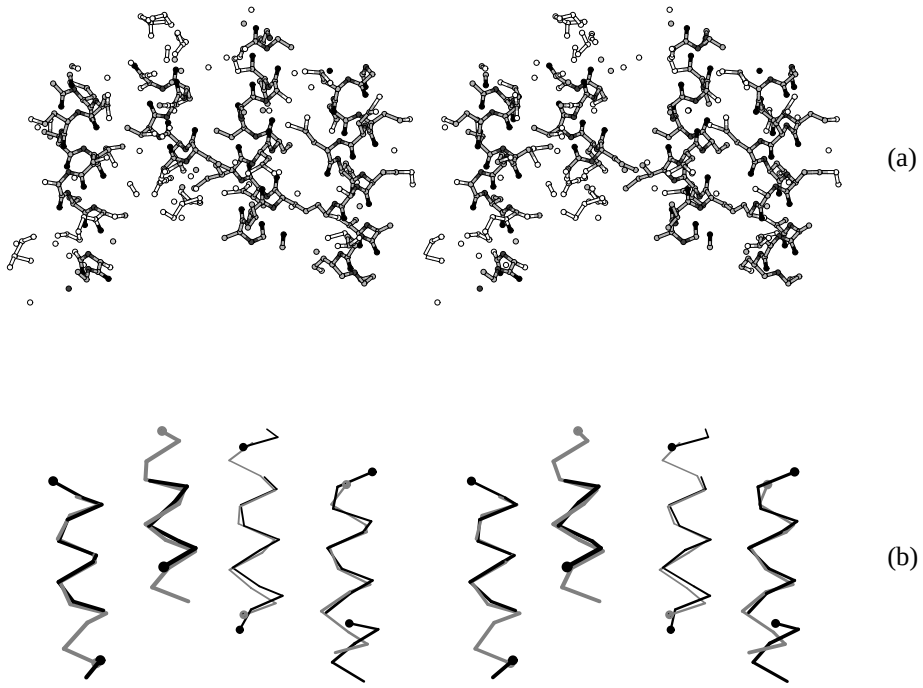

Figure 5.10: *Stereo-views of the best structure based on map correlation coefficients, obtained after 25 cycles of conditional optimization with* σ_A^i*-estimates and fixed values of* m_a *after cycle 7. (a) A ball-and-stick representation with automatic assignment of atom types based on gradient contributions 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

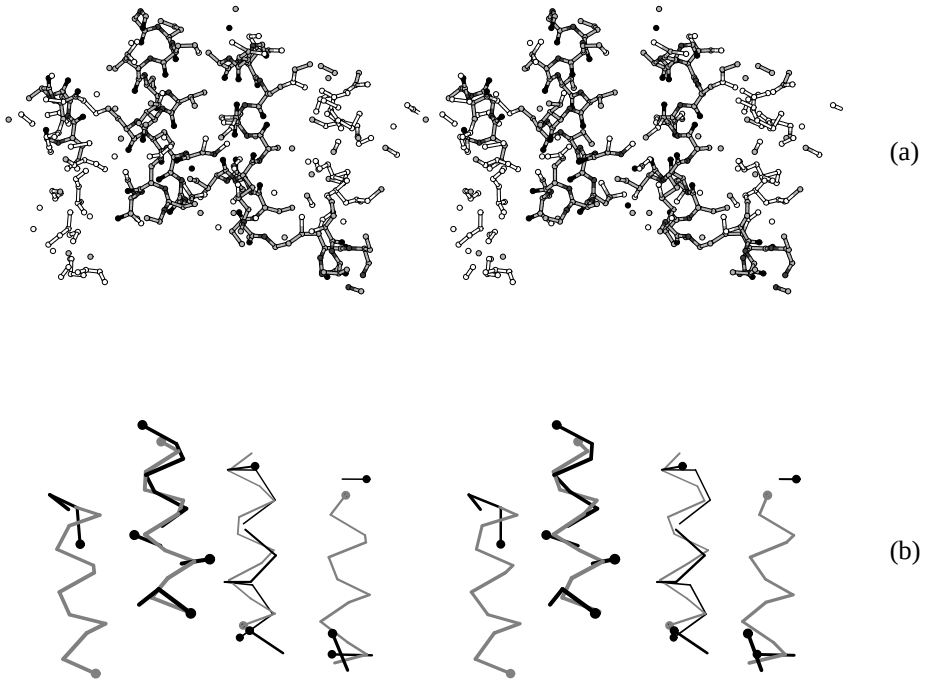

Figure 5.11: *Stereo-views of the worst structure based on map correlation coefficients, obtained after 25 cycles of conditional optimization with σ_A^i-estimates and fixed values of m_a after cycle 7. (a) A ball-and-stick representation with automatic assignment of atom types based on gradient contributions 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

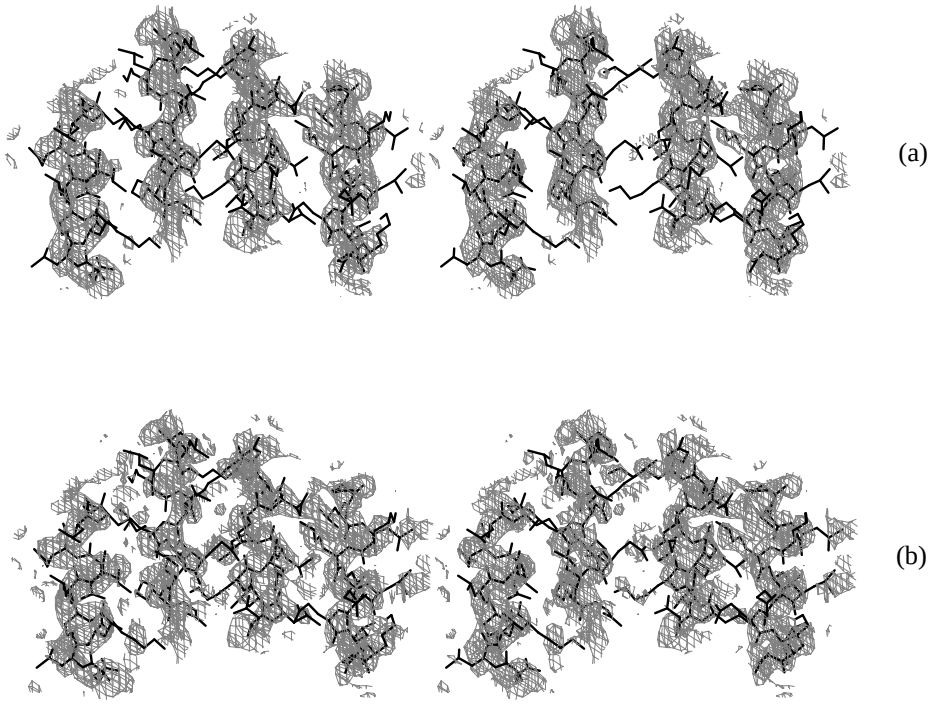

Figure 5.12: *Electron density maps ($m|F^{\mathrm{obs}}|\exp(i\varphi^{\mathrm{ave}})$) after 25 cycles of conditional optimization with σ_A^i-estimates and fixed values of m_a after cycle 7, (a) including calculated intensities for the four suspect low-resolution reflections and (b) excluding these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

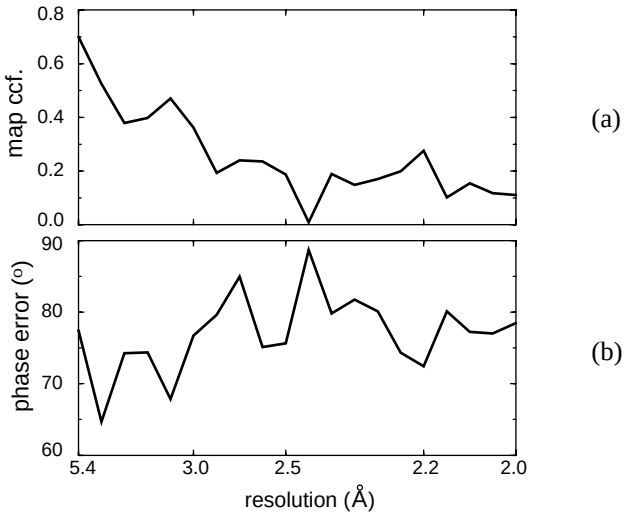

Figure 5.13: *Map correlation coefficients (ccf.) (a) and $|F^{\mathrm{obs}}|$-weighted phase errors (b) of F^{ave} as a function of resolution after 25 cycles of conditional optimization with σ_A^i-estimates and fixed values of m_a after cycle 7.*

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).