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.

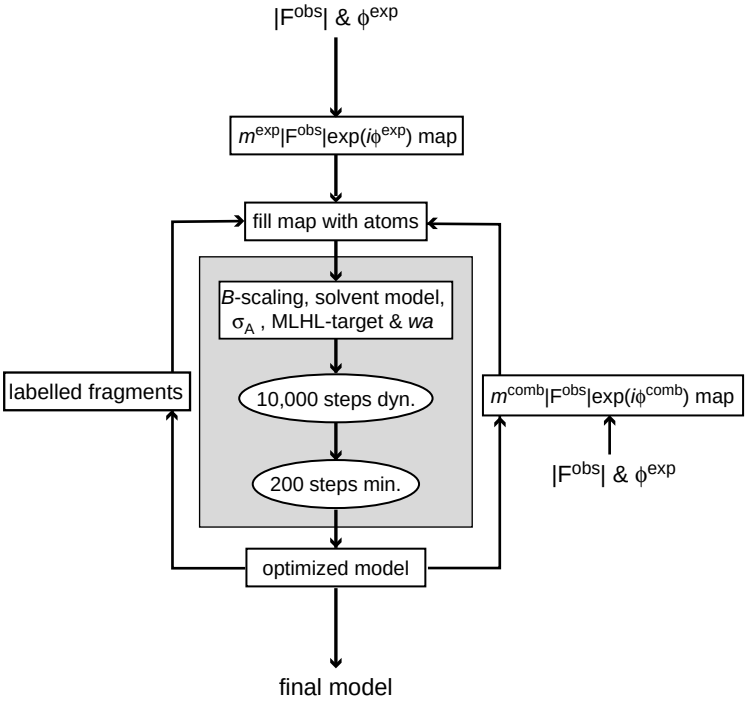

Figure 4.1: *Refinement protocol for automated model building by conditional optimization. Electron density maps were calculated on a gridsize of 0.2 Å. Regions of the experimental map ($m^{\mathrm{exp}}|F^{\mathrm{obs}}|\exp(i\varphi^{\mathrm{exp}})$) with density levels above 1.0σ were filled with 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.*

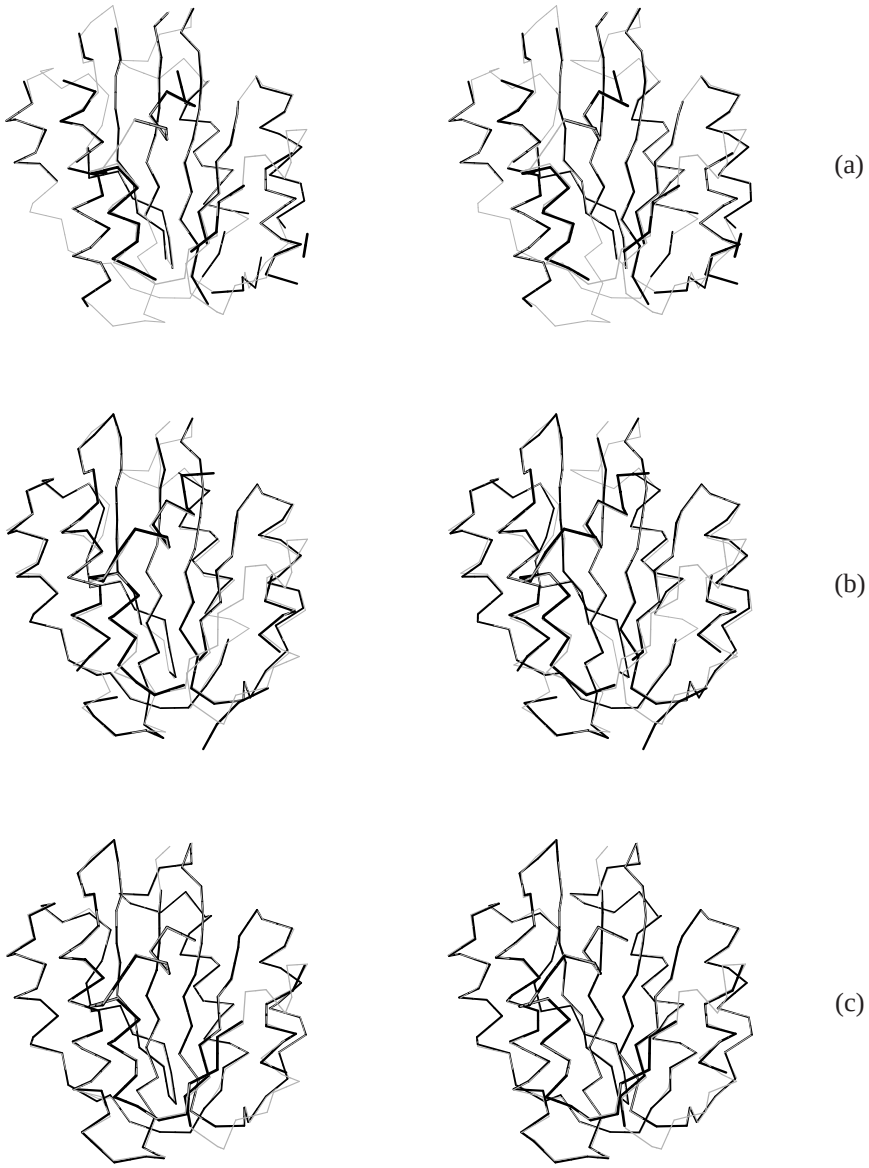

Figure 4.2: *Stereo-views of the automatically built models (black) generated by (a) conditional optimization, (b) ARP/wARP and (c) RESOLVE, superimposed on the target structure of vWF-A3 (gray).*

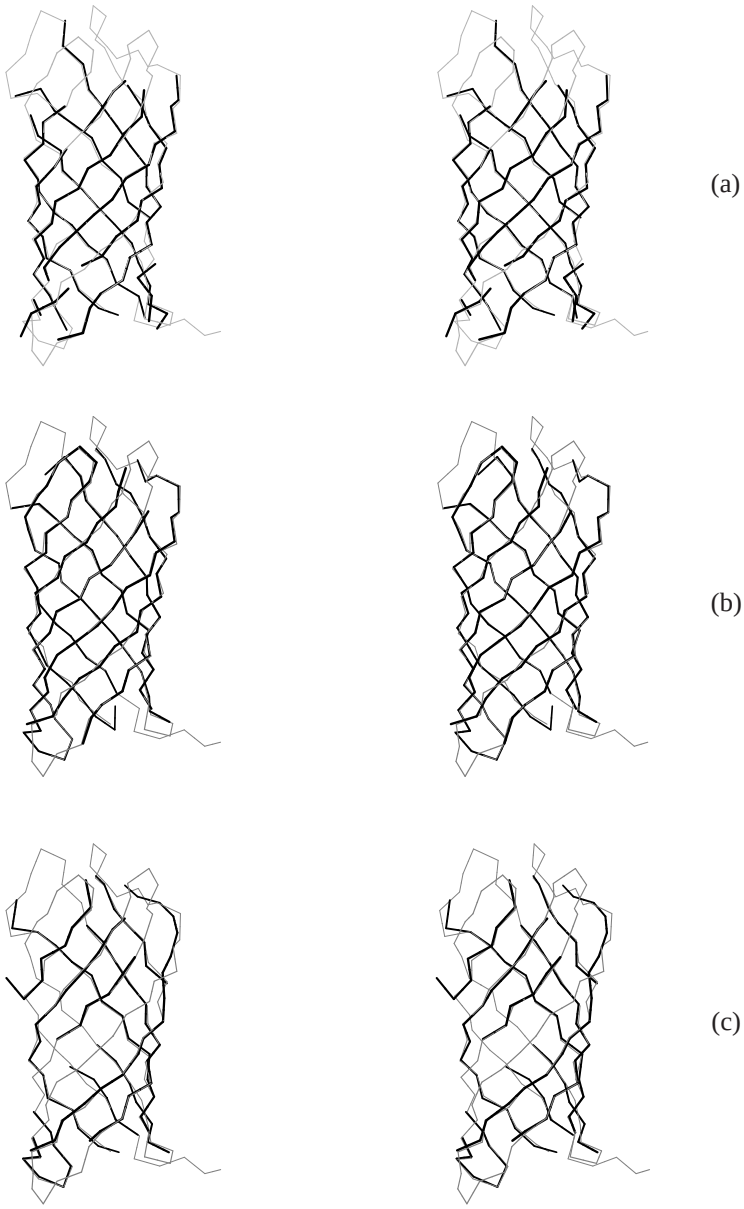

Figure 4.3: *Stereo-views of the automatically built models (black) generated by (a) conditional optimization, (b) ARP/wARP and (c) RESOLVE, superimposed on the target structure of NspA (gray).*

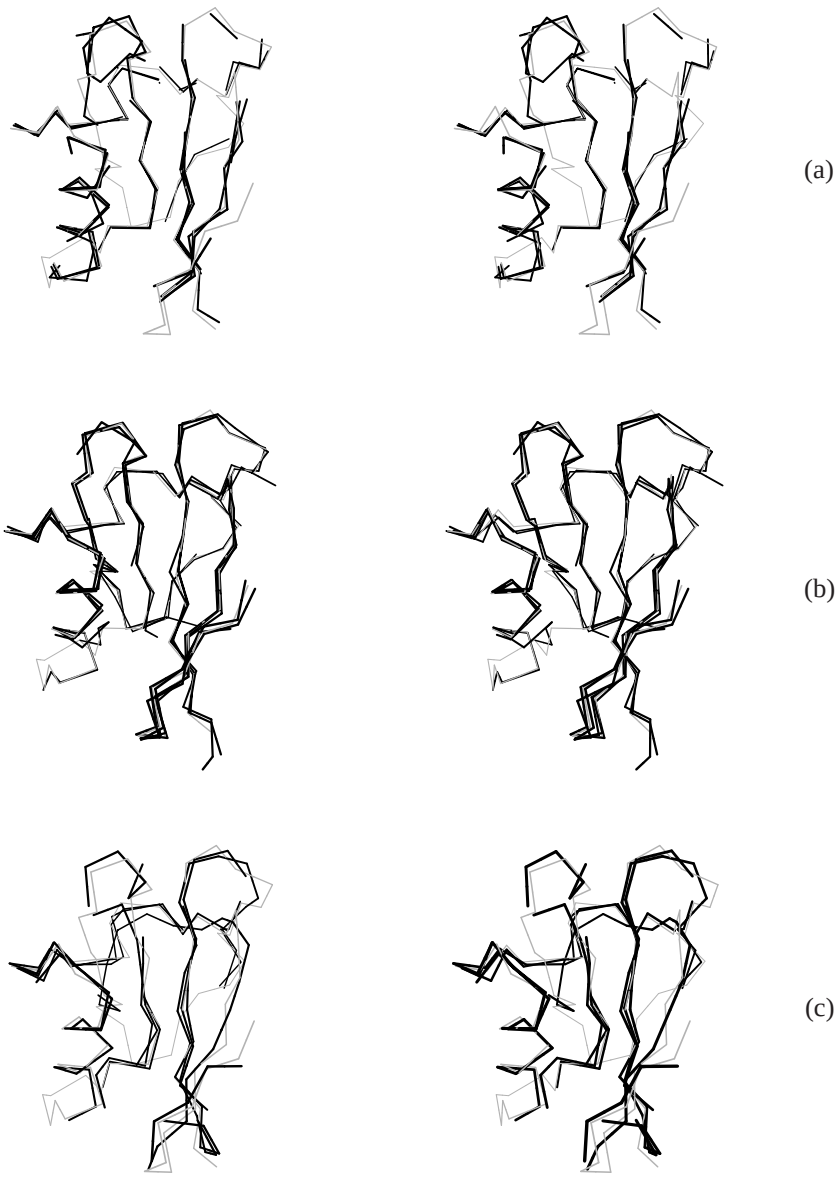

Figure 4.4: *Stereo-views of the automatically built models for LAPP generated by (a) conditional optimization, (b) ARP/wARP and (c) RESOLVE. Auto-built protein fragments for all three molecules in the asymmetric unit (black) are superimposed on one of the target molecules (gray).*

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).