electronic reprint

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449
Editors: **E. N. Baker and Z. Dauter**

The potentials of conditional optimization in phasing and model building of protein crystal structures

Sjors H. W. Scheres and Piet Gros

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

The potentials of conditional optimization in phasing and model building of protein crystal structures

Sjors H. W. Scheres and Piet Gros*

Department of Crystal and Structural Chemistry, Bijvoet Center for Biomolecular Research, Utrecht University, The Netherlands

Correspondence e-mail: p.gros@chem.uu.nl

Received 23 January 2004
Accepted 15 April 2004

Model building is a pivotal step in protein-structure determination, because with an atomic model available the vast amount of geometrical prior knowledge may be applied to the structure-determination process. Here, conditional optimization, a method that does not require interpretation of the electron-density map, is described. Instead, this method refines loose atoms for which all chemical interpretations are considered simultaneously using an *N*-particle formalism. This method bears the potential of introducing the geometrical data much earlier in the structure-determination process, *i.e.* well before an interpretable electron-density map has been obtained. Here, results from two tests are presented: automated model building of three proteins with diffraction data extending to 2.4–3.0 Å resolution and *ab initio* phasing of a small four-helical bundle with diffraction data to 2.0 Å resolution. Models built automatically by the widely used programs *ARP/wARP* and *RESOLVE* and those from conditional optimization *per se*, without discrete modelling steps, had comparable phase quality and completeness, except in loop regions, which are poorly modelled by the current force field in conditional optimization. Optimization of multiple random starting models by conditional optimization yielded models revealing the four helices of the four-helical bundle.

1. Introduction

In protein crystallography, the generation of an atomic model of the molecules is a crucial step in the structure-determination process. With an atomic model available, the vast amount of geometrical data of protein structures can be applied in structure refinement in order to generate better phases and a better atomic model. In practice, an atomic model can only be generated when sufficient phase information has been obtained either by experimental means or through the use of known homologous structures to produce an interpretable electron-density map. The model-building task may be far from straightforward, because the phase information may be poor and the resolution of the diffraction data may be limited. For maps at high resolution ($d \leq 2.2$ Å) and with good starting phases, automation of the model-building process has been highly successful in recent years (Perrakis *et al.*, 1999). Automation has reduced enormously the amount of time involved in manual model building using computer graphics programs. Currently, various approaches are being developed to improve the pattern recognition of protein structural features in the electron-density maps (Terwilliger, 2003; Holton *et al.*, 2000; Levitt, 2001) so that automated model building can deal with lower resolution data and poorer phase information. Nonetheless, at increasingly lower resolution and with poorer phase information the interpretation will become increasingly

unreliable. In such cases, the quality of the electron-density map may not allow unique identification of molecular fragments and hence may not allow the assignment of chemical identities to the atoms present in the model. Similar to the treatment of X-ray diffraction data in maximum-likelihood refinement, introduction of the geometrical data would then require a statistical procedure in which all possible assignments are taken into account simultaneously instead of testing individual assignment hypotheses. In recent years, we have developed such an approach that we call 'Conditional Optimization' (Scheres & Gros, 2001, 2003). Such a treatment of an ensemble of structural hypotheses instead of single hypotheses creates the possibility of using the geometrical prior information at an earlier stage in the structure-determination process, thereby merging the steps of phasing, model building and refinement further than current practice.

The most widely applied automated modelling procedure is *ARP*/*wARP*, developed by Lamzin, Perrakis and coworkers (Lamzin & Wilson, 1993; Perrakis *et al.*, 1999; Morris *et al.*, 2002). *ARP*/*wARP* presents a powerful iterative combination of loose-atom positioning, recognition of protein fragments from the distribution of atoms and refinement of the loose atoms and the pieces of assigned protein fragments by *REFMAC* (Murshudov *et al.*, 1997). Though originally limited to diffraction data with resolution limits extending beyond $d \simeq 2.3$ Å, recent (Morris *et al.*, 2002) and current developments in pattern recognition (discussed by Cohen *et al.*, 2004) appear to make this approach feasible at lower resolution limits ($d \leq 2.6$ Å). Because of the coupling of refinement with the process of model building, the resulting models are typically accurate and significant phase improvements are observed. Discrete modelling steps ensure that the model is completed, generating as much as possible of a continuous main chain and assigning side chains using the amino-acid sequence. Other approaches, *e.g. RESOLVE* (Terwilliger, 2001, 2003, 2004) and *MAID* (Levitt, 2001), have been developed that start by positioning fragments into the electron-density map instead of generating fragments from loose-atom positions. These methods have the potential to be applied to data with lower resolution limits. Based on a set of diverse proteins, Badger (2003) reports significant success in building ∼75% of the main chain with these two methods using maps at 2.3–2.7 Å resolution. To some extent similar to *ARP*/*wARP*, the approach in *RESOLVE* (Terwilliger, 2003) consists of an iterative procedure of placing fragments, extending the model into loop regions, docking of the primary sequence alternated by maximum-likelihood density modification and restrained protein-structure refinement using *REFMAC* (Murshudov *et al.*, 1997).

In two recent publications (Scheres & Gros, 2001, 2003) we presented the method of 'Conditional Optimization'. At the heart of this approach is an *N*-particle method in which we assign all possible chemical identities to atoms based on their spatial arrangement. Through this method, we can express prior geometrical knowledge without the requirement for a topological model. For the assignments of chemical identities and possible topologies we use sets of conditions, which are continuous scoring functions [$C = (0, 1)$] that express target values of observed interatomic distances and dihedral angles in known protein structures (see, for example, Figs. 2 and 4 in Scheres & Gros, 2001). In addition to these 'through-bond' conditions, the formulation includes local atomic density conditions describing the observed packing density for atom types (Fig. 3 of Scheres & Gros, 2001). These various types of conditions lend themselves to a logical organization in what we have called layers: layer 0 (atoms defined by local atomic density conditions), layer 1 (bond conditions), layer 2 (angle conditions) *etc*. Assignment of a particular fragment to a set of atoms is then determined by multiplication of conditions into joint conditions. The combinations of conditions for atom types, bond types, angle types *etc.* is made according to the topology of protein molecules (strictly speaking, we consider not only topology but also distinct conformations since the scoring functions are based on the spatial arrangement of the atoms). In this way, any arbitrary set of atoms is evaluated. However, since zero-scoring conditions yield zero-scoring joint conditions (and zero derivatives), we need to consider only non-zero interactions in our computations. Effectively, each joint condition represents a single assignment hypothesis. An example of the conditions making up one hypothesis is depicted in Fig. 1 of Scheres & Gros (2003). The number of hypotheses that can thus be made for a protein structure is of the order of the number of atoms times the number of layers. Considering only a maximum number of layers implies that the number of hypotheses depends linearly on the number of atoms. Therefore, an algorithm based on this formulation is of order *N*, which makes the computational problem of making assignments tractable. As target functions in the optimization process, we choose harmonic potentials that restrain the number of occurrences of structural fragments associated with joint conditions to the expected number based on sequence and secondary-structure prediction. Since we choose to use only continuous functions in the calculation of the (joint) conditions, we can compute derivatives that can be used in an optimization process.

In the first paper (Scheres & Gros, 2001), we showed that Conditional Optimization successfully built the helices of a four-helical bundle starting from randomly distributed atoms in a simple artificial test case with 2 Å resolution diffraction data. In the second paper (Scheres & Gros, 2003), the set of conditions was extended to treat commonly observed conformations in proteins: α-helices, β-sheets, a limited number of loop conformations and side chains up to the γ position. Refinement of three protein structures with large randomly generated coordinate errors against their 2 Å resolution diffraction data showed that Conditional Optimization has a large radius of convergence.

Here, we report two types of tests of Conditional Optimization: automated model building and *ab initio* modelling. The most powerful approach in automated model building would be an iterated process of refinement cycles and discrete model-building steps, as is performed in *ARP*/*wARP* and *RESOLVE*. However, rather than providing a ready-to-use solution, we chose to first test the potential of Conditional Optimization

per se in the model-building process. Therefore, in the calculations presented here no pattern-recognition or model-building techniques were applied other than the conditional formulation itself. Using three test cases from our own laboratory with resolution limits of 2.4, 2.6 and 3.0 Å and good experimental phases, we compared Conditional Optimization to the commonly used programs *ARP*/*wARP* and *RESOLVE*. To test the potential of the method for phasing, we applied Conditional Optimization starting from multiple models of randomly distributed atoms. This time (*cf.* Scheres & Gros, 2001), we used the experimental structure-factor amplitudes of the four-helical bundle Alpha-1 (Privé *et al.*, 1999).

2. Experimental

2.1. Test in automated model building

Three protein structures were selected for testing Conditional Optimization for automated model building: (i) the A3-domain from human von Willebrand Factor, vWF-A3 (Huizinga *et al.*, 1997), (ii) outer-membrane protein NspA from *Neisseria meningitidis* (Vandeputte-Rutten *et al.*, 2003) and (iii) the C-terminal domain of leech anti-platelet protein, LAPP (Huizinga *et al.*, 2001). All three structures were solved in our laboratory at medium to low resolution (2.4–3.0 Å resolution) and with good experimental phases. For all cases, the original models were built manually using the graphics program *O* (Jones *et al.*, 1991). The main characteristics of these test cases are given in Table 1.

Fig. 1(*a*) shows the protocol used for automated model building by Conditional Optimization. The process was started by positioning loose and unlabelled atoms in the $(m|\mathbf{F}_{\rm obs}|\exp\{i\varphi_{\rm best}\})$ electron-density map at sites with $\rho > 1.0\sigma$, with minimum and maximum interatomic distances of 1.1 and 1.8 Å, respectively, and a maximum of four neighbouring atoms within 1.8 Å. Optimization was performed using the program *CNS* (Brünger *et al.*, 1998). For the crystallographic target function, we used the phase-restrained maximum-likelihood function (MLHL; Pannu *et al.*, 1998) with phases and figures of merit from the experimental phasing process (Table 1). σ_A values (Read, 1986) were estimated using test-set reflections. For the geometric target functions we used the conditional formulation (Scheres & Gros, 2001, 2003). For each protein, a specific force field was generated using the general parameter list of conditions for α-helices, β-strands, loops and side chains up to γ-positions (Scheres & Gros, 2003) and an approximate estimate of the secondary-structure content (see Table 1). In the case of LAPP we set the loop content to 0, thereby excluding loop conformations from the force field, to limit computer memory requirements. At the end of each cycle of Conditional Optimization (Fig. 1*a*), we assigned atom labels based on the implicit assignments used in the Conditional Optimization

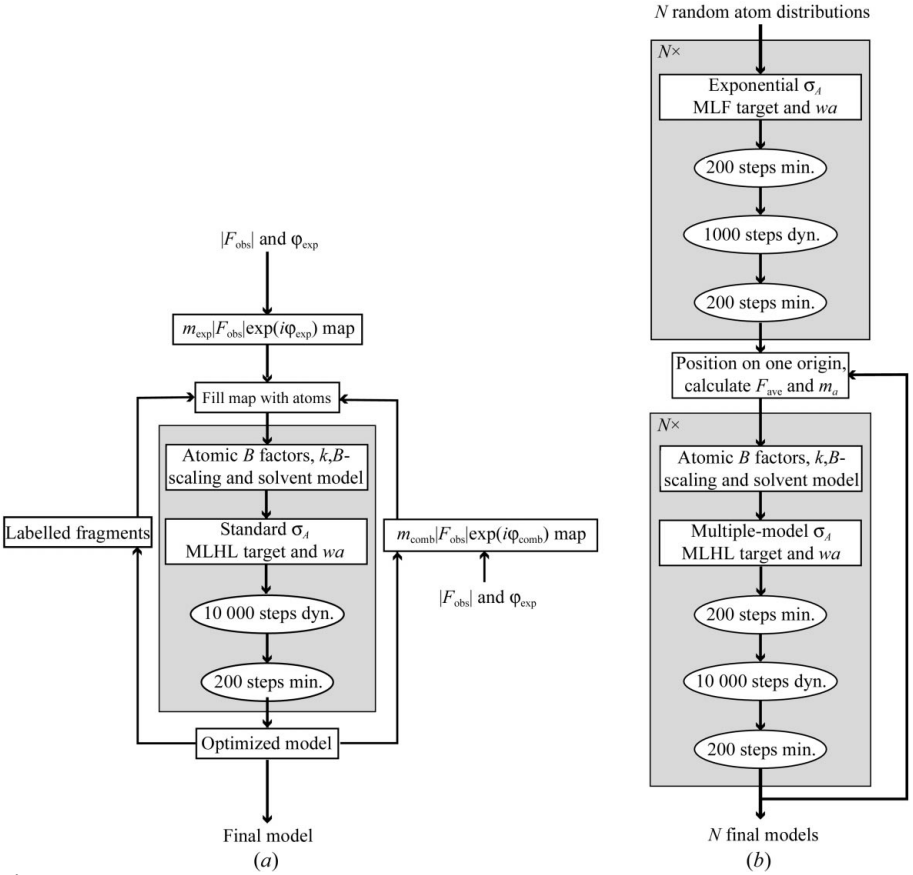

Figure 1
Protocols for automated model building and *ab initio* modelling. (*a*) Automated model building by Conditional Optimization using the program *CNS* (Brünger *et al.*, 1998). The standard geometrical target functions were replaced by our conditional formulation (see main text). Electron-density maps were filled with loose atoms (as described in the main text). Refinement cycles used the MLHL crystallographic target function and consisted of 10 000 steps of dynamics (denoted 'dyn.') and 200 steps energy minimization ('min.'). Velocities in the dynamics calculations were scaled to a constant temperature of 600 K. Overall anisotropic *B*-factor scaling, bulk-solvent correction, σ_A estimation, determination of *wa* weight and phase combination were performed using standard *CNS* routines. After each cycle, the positions of atoms that could be assigned to protein fragments (see main text) were maintained in the model. Placing the remaining number of atoms in the phase-combined electron-density map completed the starting model for a next cycle of optimization. (*b*) Multiple-model optimization protocol starting from *N* models of randomly distributed atoms. The *CNS* program was used with conditional formulation for the geometrical target functions. First, a short condensation step was performed for each model with 200 steps of minimization and 1000 steps of dynamics followed by 200 steps minimization using the MLF crystallographic target function. Subsequent optimization cycles consisted of 200 steps of minimization, 10 000 steps of dynamics and 200 steps of minimization using the MLHL crystallographic target function. Phases of the averaged structure factors and figures of merit (m_a) were used in the phase restraint (see main text). The phased-translation function was used to put the *N* models on a common origin prior to averaging. The bulk-solvent model was generated for 20%(*v*/*v*); atoms present in solvent regions were given zero occupancy.

process. Atoms were labelled with a chemical identity (N, C^{α}, O, C, C^{β}, C^{γ} or S^{γ}) if the gradient contribution towards that particular atom type was at least twice as large as the second largest contribution. Subsequent cycles of optimization were started from phase-combined maps, combining model and experimental phases. Subsequent starting models consisted of the atoms (but not their labels) selected in our labelling procedure with additional atoms placed in the electron-density map as described above. For vWF-A3 and NspA two cycles and for LAPP four cycles of model building by Conditional Optimization were performed.

All three test cases were also subjected to automated model building by *ARP/wARP* (version 6.0; Perrakis *et al.*, 1999; Morris *et al.*, 2002) and *RESOLVE* (version 2.03; Terwilliger, 2003), both using *REFMAC* version 5.1.24 for refinement (Murshudov *et al.*, 1997). These calculations were performed using default values for all input 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).

Table 1
Statistics of the three test cases for automated model building.

	vWF-A3	NspA	LAPP
Data statistics			
Space group	$P2_12_12_1$	$R32$	$P4_322$
Resolution limit (Å)	2.4	2.6	3.0
$I/\sigma(I)$ 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
$\langle\Delta\varphi\rangle$† (°)	35.6	38.3	28.2
$\langle\cos(\Delta\varphi)\rangle$‡	0.59	0.49	0.63
Model statistics			
Z	1	1	3
No. residues	183	155	264
Solvent content (%)	35	70	70
Input to Conditional Dynamics			
No. atoms	1450	1200	2200
Secondary-structure content (% α, % β, % loop)	50, 30, 20	0, 75, 25	40, 60, 0

† Amplitude-weighted mean phase error calculated with respect to the refined structures. ‡ Unweighted mean cosine phase error.

2.2. Testing *ab initio* modelling

We selected the four-helical bundle Alpha-1 (Privé *et al.*, 1999; PDB code 1byz) to explore the potentials of Conditional Optimization in *ab initio* phasing. This structure consists of 396 protein atoms in space group $P1$ and was originally solved by direct methods using all observed diffraction data to 0.9 Å resolution. Here, we truncated deposited structure-factor amplitudes to 2.0 Å resolution.

The protocol used to refine N multiple models using Conditional Optimization is given in Fig. 1(*b*). The number of atoms per model was 400. Initial models consisted of randomly distributed atoms. Calculations were performed using the program *CNS* (Brünger *et al.*, 1998). The target functions from Conditional Optimization replaced the standard geometric target functions. The random models were first subjected to 1000 steps of maximum-likelihood refinement using structure-factor amplitudes (MLF; Pannu & Read, 1996). The σ_A values for this initial cycle were calculated according to $\sigma_A = \exp(-150s^2)$. In subsequent cycles, each containing 10 000 steps of dynamics, we used the phase-restrained maximum-likelihood crystallographic restraint (MLHL; Pannu *et al.*, 1998) with target phases and figures of merit derived from averaging the structure-factor sets of the individual models. To this end, all individual structures were first placed on a common origin using the phased-translation function. Figures of merit (m_a) were computed per resolution shell using only test-set reflections: $m'_a = \sum_{i=1}^{N} \mathbf{F}^i / \sum_{i=1}^{N} |\mathbf{F}^i|$, where $\mathbf{F}^i$ were calculated structure factors from an individual model, and were extrapolated to infinite models by $m_a = \{[N(m'_a)^2 - 1]/(N-1)\}^{1/2}$. For each individual model, we estimated σ_A values per resolution shell. We assumed that the true phase error of a model would be related to phase differences of that model to any other model,

$$\sigma_A^i = \langle \sigma_A^{ij} \rangle_j = \left\langle \langle |\mathbf{E}^{\rm obs}||\mathbf{E}^i| \cos(\varphi^i - \varphi^j) \rangle / (\langle |\mathbf{E}^{\rm obs}|^2 \rangle \langle |\mathbf{E}^i|^2 \rangle)^{1/2} \right\rangle_j .$$

These estimates were calculated using all reflections because the low numbers of reflections in the test set alone yielded unstable results.

Calculations were performed on 667 MHz single-processor Compaq XP1000 workstations with 1–2 Gb of memory. The CPU times for automated model building are given in Table 2.

Table 2
Statistics of automated model building of three crystal structures at medium to low resolution.

	vWF-A3 (2.4 Å)			NspA (2.6 Å)			LAPP (3.0 Å)		
	CO	*ARP*	*RESOLVE*	CO	*ARP*	*RESOLVE*	CO	*ARP*	*RESOLVE*
No. residues built	148	160	170	121	130	101	169	232	136
Fraction built (%)	80	87	93	78	84	65	64	87	51
No. chains	20	8	4	16	6	10	38	11	13
R.m.s.d.† (Å)	1.5	1.9	0.9	1.3	1.8	0.9	1.2	1.3	1.8
$\langle\Delta\varphi\rangle$‡ (°)	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

† Root-mean-square coordinate deviations, or coordinate errors, were calculated based on the distance between atoms in modelled protein fragments to the nearest atom with a corresponding atom label in the refined structure. ‡ Amplitude-weighted mean phase error calculated with respect to the refined structures. § For *RESOLVE*, the phase errors of the resulting electron-density maps are given in parentheses. ¶ Unweighted mean cosine of the phase error with respect to the refined structures. For calculation of both amplitude-weighted and unweighted phase errors all atoms of the resulting models were taken into account, *i.e.* for models generated by conditional optimization (CO) or *ARP/wARP* (*ARP*) atoms that were not recognized as part of a protein fragment were also included.

Ab initio modelling calculations took more than 100 d of CPU time.

3. Results and discussion

3.1. Model-building tests

We compared automated model building by Conditional Optimization, which did not include any discrete decision-making steps, with the commonly used automated building programs *ARP*/*wARP* and *RESOLVE*. Models were built for three test cases, vWF-A3, NspA and LAPP, with data to $d = 2.4$, 2.6 and 3.0 Å resolution, respectively. Criteria for comparison were model completeness, correctness of the trace, accuracy of the positioned fragments assessed by r.m.s. coordinate errors and quality of the phases computed from the models. Statistics of the automatically built models are given in Table 2. C^{α} traces of the generated models are given in Fig. 2. Coordinate

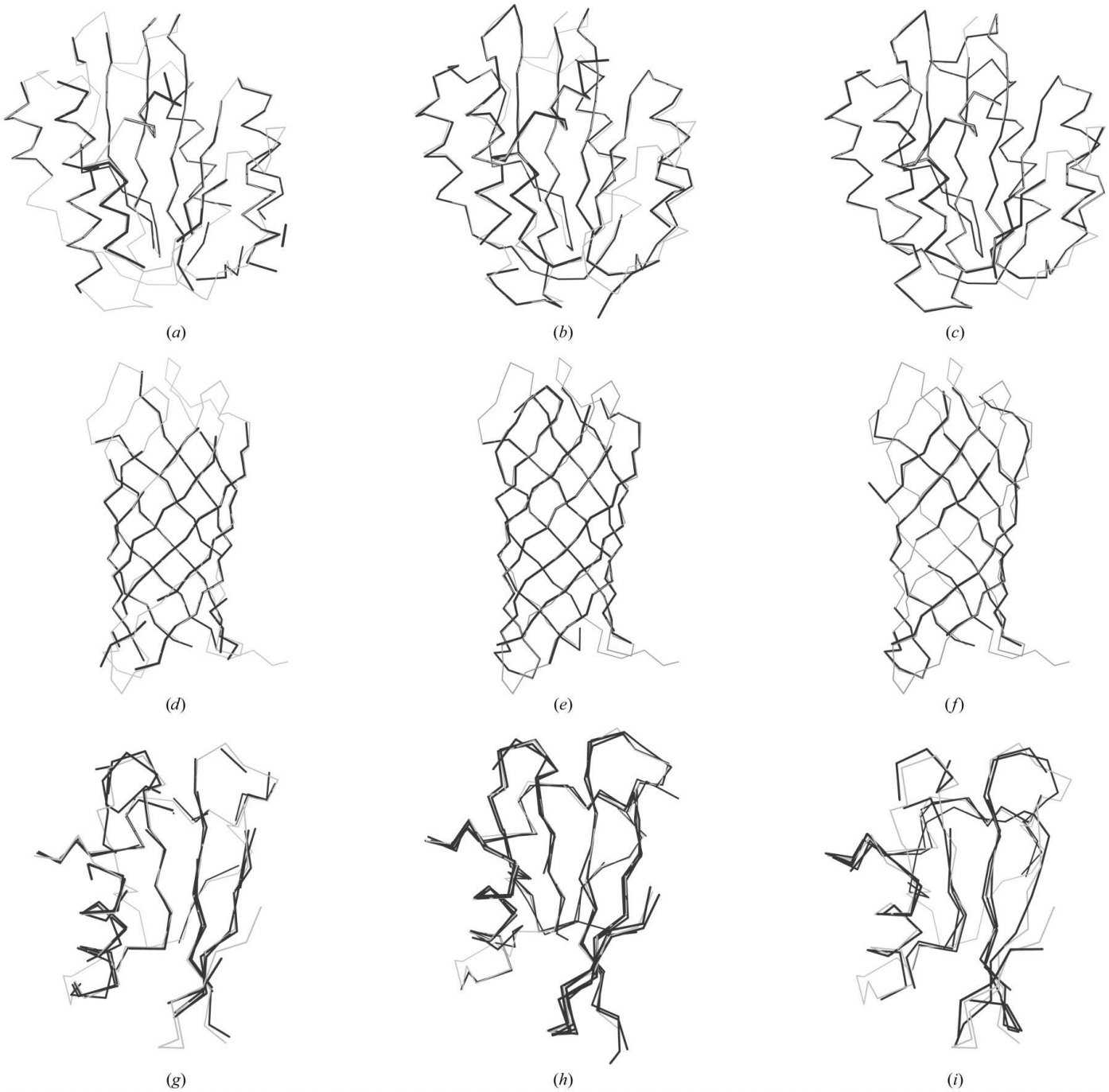

Figure 2
C^{α} traces of automatically built models (black lines) overlaid with traces of the refined structures (grey lines): models of the vWF-A3 domain (using diffraction data to 2.4 Å resolution) obtained by Conditional Optimization (*a*), *ARP*/*wARP* (*b*) and *RESOLVE* (*c*); models obtained for NspA (with data to 2.6 Å resolution) by Conditional Optimization (*d*), *ARP*/*wARP* (*e*) and *RESOLVE* (*f*); superposition of the three models of three independent molecules of LAPP determined at 3.0 Å resolution by Conditional Optimization (*g*), *ARP*/*wARP* (*h*) and *RESOLVE* (*i*).

files for the nine generated models and for the three refined models used in the analysis have been submitted as supplementary material.[1]

For vWF-A3, automated model building was tested using data to 2.4 Å resolution and experimental phases with an (amplitude-weighted) mean phase error of 35.6°. *ARP*/*wARP* and *RESOLVE* built more complete and less fragmented models than did Conditional Optimization (Figs. 2*a*–2*c*). *RESOLVE* produced the best model that was most complete, missing only one loop and a small α-helix, had the smallest r.m.s. coordinate error and the lowest mean-phase error. The model from *ARP*/*wARP* missed one α-helix, one β-strand and two loops. It had a relatively large coordinate error and the phases computed from the model were not better than the experimental phases. Conditional Optimization yielded a fragmented model consisting of almost all α-helical and β-strand segments, except the one small α-helix that was also missed by the other two programs. Only one of the loops was modelled correctly. Another loop was modelled with a reversed chain direction, as was a small β-strand which flanks the central β-sheet. Notwithstanding these errors, the model from Conditional Optimization was more accurate and yielded better phases than the model from *ARP*/*wARP*.

For NspA (Figs. 2*d*–2*f*), using data to 2.6 Å resolution and experimental phases with an (amplitude-weighted) mean phase error of 38.3°, Conditional Optimization built most of the strands in the β-barrel, but none of the turns. The largest errors in this model were reverse chain directions for an entire β-strand and two smaller β-strand fragments. In this case, *ARP*/*wARP* produced the most complete model and the lowest phase error, though the model included two β-strands with reversed chain directions. *RESOLVE* built a smaller portion of the molecule with a strand that crossed over into a neighbouring strand. The mean-phase error using calculated phases from this model was relatively large, possibly reflecting the incompleteness of the model produced by *RESOLVE*.

The data from LAPP represent a situation in which automated model building generally does not work owing to the limited resolution (3 Å) of the diffraction data. However, solvent flattening and threefold non-crystallographic symmetry averaging yielded high-quality phases with an (amplitude-weighted) mean-phase error of 28.2°. Conditional Optimization (Fig. 2*g*) yielded a model with partially built β-sheets and most α-helical segments of 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 also modelled incorrectly by an α-helical turn (note: in this particular case loops were omitted from the force field). This model had a fairly low r.m.s. coordinate error and good phase quality. *ARP*/*wARP* built a more complete model with more β-strands and more loops (Fig. 2*h*). One incorrect main-chain trace from an α-helix to a neighbouring β-strand was observed in this model. This model yielded a phase error comparable to that obtained with conditional optimization. As for NspA, *RESOLVE* (Fig. 2*i*) produced the model with the lowest completeness. In addition, this model had a low accuracy of the positioned fragments and a high mean-phase error. It contained more main-chain trace

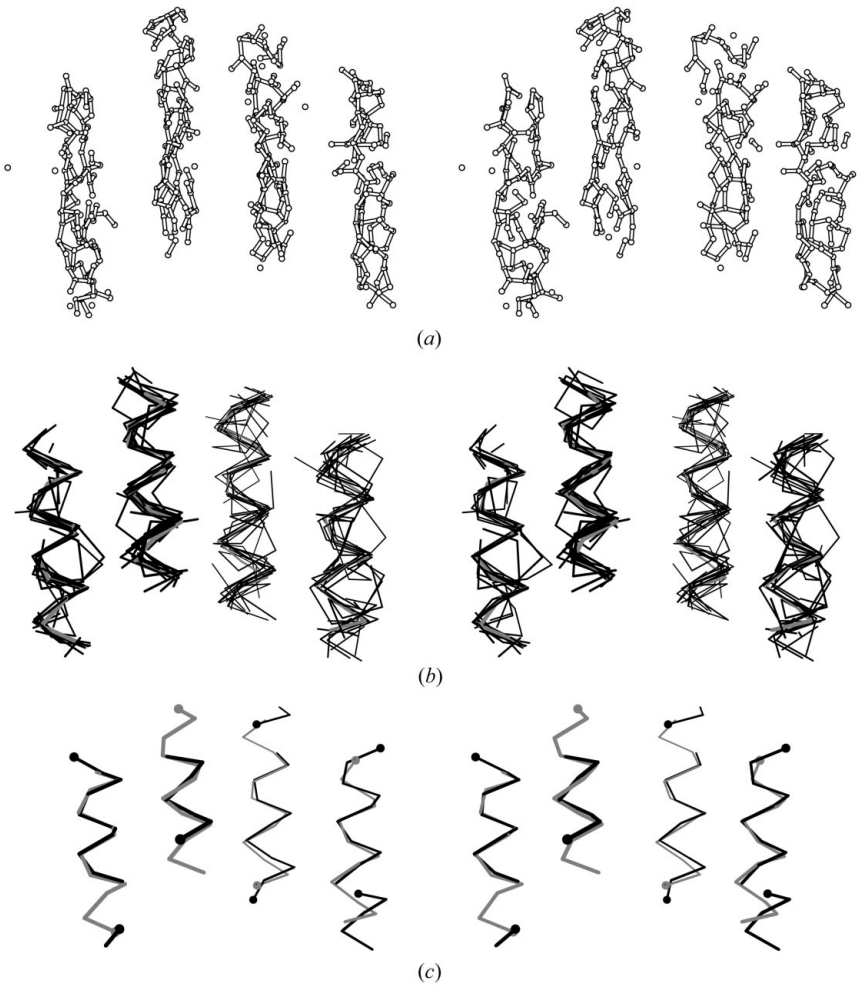

Figure 3
Structures of Alpha-1 obtained by *ab initio* modelling against 2 Å resolution data. (*a*) Stereoview of a ball-and-stick representation of an optimized model after the initial condensation using 1000 steps of conditional dynamics using the MLF target function; (*b*) C^{α} trace of 17 models obtained after 25 cycles of optimization (black lines) with the C^{α} trace of the refined structure overlaid (grey lines); (*c*) stereoview of the C^{α} trace of the model with the highest σ_A value. From left to right the helices are oriented down, up, down and up in the refined structure (the position of the N-terminus is indicated by a ball). The chain directionality in the depicted model is therefore from left to right: incorrect, incorrect, correct and correct. Assignment of atom labels (for *b* and *c*) was based on the gradient contributions in Conditional Optimization (as described in the main text for automated model building).

[1] Supplementary data have been deposited in the IUCr electronic archive (Reference BA5060). Methods for accessing these data are described at the back of the journal.

errors than the models built by either Conditional Optimization or *ARP*/*wARP*.

These three cases clearly demonstrate that Conditional Optimization can produce models of comparable completeness and comparable phase quality as *ARP*/*wARP* and *RESOLVE*. The models from Conditional Optimization have significantly lower connectivity and contain fewer loops and turns. Similar to *ARP*/*wARP* and *RESOLVE*, an increase in connectivity would be most efficiently achieved by introducing discrete model-completion steps. Moreover, unsuccessful modelling of turns and loops reflects the currently limited conditions defined for turn and loop conformations in the conditional force field. The worst aspect of Conditional Optimization is the excessive amount of CPU time and computer memory required.

3.2. *Ab initio* modelling

The possibility of phasing protein structures by *ab initio* modelling was explored using experimental 'real' diffraction data of the four-helical bundle Alpha-1 (Privé *et al.*, 1999), *cf.* the tests of *ab initio* phasing using artificial data in Scheres & Gros (2001). We chose to perform parallel optimization of multiple random models and used the set of models in two ways: (i) the variation among the multiple models provided an indication of the statistical relevance of the individual models and (ii) averaging the individual structure-factor sets provided phases that were used as phase restraints in the MLHL crystallographic target.

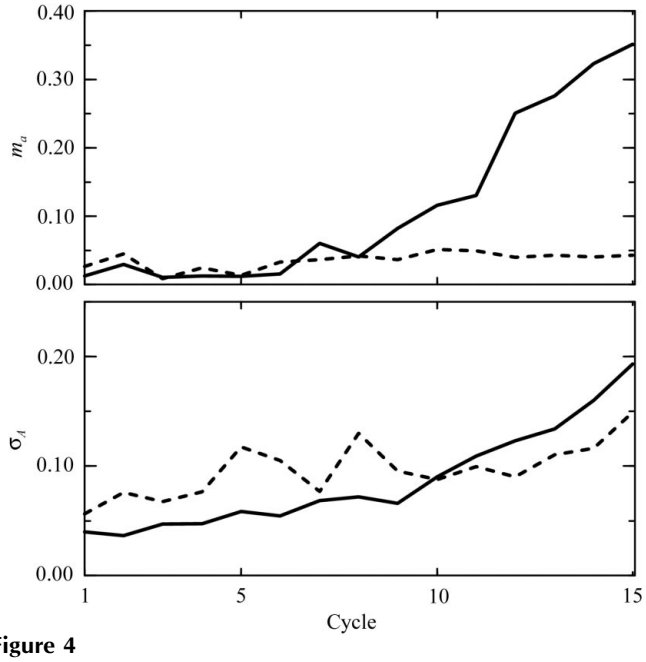

Figure 4
Figures of merit and σ_A values derived from multiple models. Figures of merit m_a (*a*) and σ_A values for one of the individual models (*b*) computed for data up to 2 Å resolution over 15 cycles of optimization. Solid lines represent the estimates used in our calculation; dashed lines indicate the corresponding values, $\langle\cos(\varphi_{\rm ave} - \varphi_{\rm calc})\rangle$ and $\sigma_A = \left\langle \langle |\mathbf{E}^{\rm obs}||\mathbf{E}^{i}|\cos(\varphi^{i} - \varphi_{\rm calc})\rangle / \langle |\mathbf{E}^{\rm obs}|^2\rangle \langle |\mathbf{E}^{i}|^2\rangle \right\rangle$, where $\varphi_{\rm ave}$ is the phase of the average structure factor, $\varphi_{\rm calc}$ is the phase of the structure factors computed from the published structure and φ^{i} is the phase of the individual model.

36 models with randomly distributed atoms were first subjected to 1000 steps of Conditional Optimization with MLF refinement (see Fig. 1*b*). In this cycle, the random atoms of most models condensed into four rod-like structures corresponding to the lowest resolution features of the diffraction data (see Fig. 3*a*). Of these 36 initial optimization runs, three runs did not finish owing to the formation of extensively branched structures requiring more computer memory than was available. 17 models were selected which appeared to have a common hand in an analysis based on the phased translation function. However, *a posteriori* analysis showed that these models did not yet have a significant handedness to allow a useful selection to be made. Subsequently, the selected 17 models were subjected to 25 optimization cycles of 10 000 steps each using a MLHL crystallographic target function (Fig. 1*b*). Initially, the estimated values of the figures of merit (m_a) for the averaged structure factors and σ_A values for the individual models behaved well when analyzed using the phases of the refined model (see Fig. 4). However, after seven cycles the figures of merit m_a were increasingly overestimated. Similarly, the σ_A values were overestimated from cycle ten onwards. Since the overestimation of m_a appeared to coincide with a drop in convergence (as judged by the map-correlation coefficients depicted in Fig. 5*a*), we decided to continue from cycle 7 with fixed figures of merit m_a. Fixed and low values for m_a also avoided subsequent overestimation of the σ_A values. Under these conditions, we observed a slow but steady convergence, as indicated by a ∼0.005 increase in map-correlation coefficient per cycle and an overall decrease in

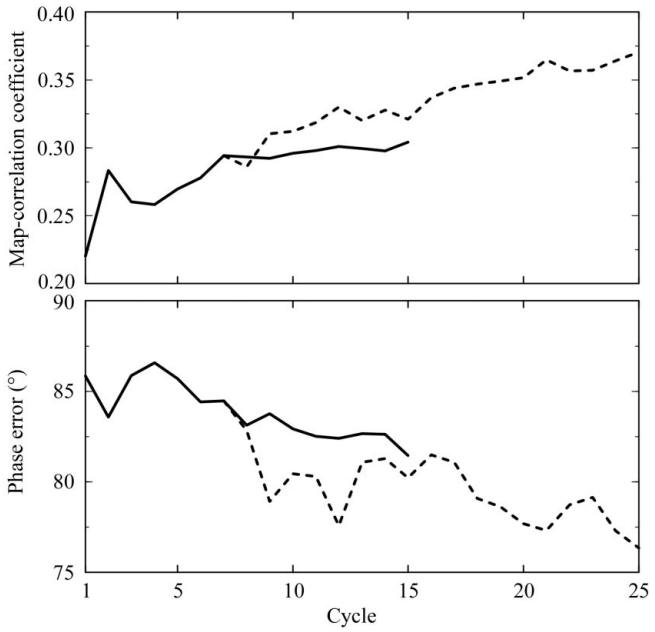

Figure 5
Convergence in *ab initio* modelling of multiple models starting from randomly distributed atoms. Map-correlation coefficients (*a*) and overall $F^{\rm obs}$-weighted phase errors (*b*) to 2.0 Å resolution of the average structure factors with respect to structure factors calculated from the published structure. Solid lines show the results for the optimization cycles with updated figures of merit (cycles 1–15). Dashed lines show the results for the optimization cycles with fixed figures of merit (cycles 7–25).

(amplitude-weighted) mean-phase error of the average structure factors by $\sim 10^{\circ}$ over 25 cycles (Figs. 5*a* and 5*b*). The mean phase error after 25 cycles was 76.3° and the map-correlation coefficient was 0.37 for all data to 2 Å resolution. Inspection of the models indicated the significance of this gain in phasing quality. An overlay of all 17 models showed that right-handed helices have developed to a reasonable extent (Fig. 3*b*). The best model, as identified by the highest σ_A value, had three and a half helices formed (see Fig. 3*c*).

This test of *ab initio* modelling was computationally demanding, which seriously limited the testing of relevant parameters. Nonetheless, the preliminary results presented here clearly indicated the most prominent shortcomings of our current approach. Throughout these calculations, the average structure factors scored among the top three of the individual sets of structure factors (*i.e.* top $\sim 20\%$) with respect to phase quality. This justifies the idea of using these phases as phase restraints. Obviously, applying this information through the MLHL target function introduces serious bias into the calculation. Though our σ_A estimates were without thorough theoretical foundation, we observed a significant correlation (0.77 after cycle 25) between the estimated σ_A values and the map-correlation coefficients of the individual models. This indicates that the use of multiple models for estimating the statistical relevance holds promise.

4. Concluding remarks

From our ongoing effort to test the procedure, we have presented two types of applications. The results from automated model building by Conditional Optimization compared well with the commonly used programs *ARP/wARP* and *RESOLVE*. Clearly, the resulting models from Conditional Optimization could be improved with the appropriate model-completion steps and better modelling of loop conformations. Nonetheless, without discrete building steps our approach already performed well. Our analysis also showed that going from medium- to low-resolution data (from 2.4 to 3.0 Å), the three methods make increasingly more tracing errors, as expected. Still, the models may be useful in a structure determination when not taken fully at face value. Furthermore, it is conceivable that improved decision-making steps may catch a number of the observed errors at low resolution, possibly strand crossings and incorrect strand directionality, which could be evaluated by testing both directions explicitly. Nonetheless, at lower resolution limits and with poorer phase information the process of model building will inherently become more difficult. Results from the preliminary tests in *ab initio* modelling by Conditional Optimization, *i.e.* without any experimental phase information present, indicated the obvious need for proper estimation of the quality of calculated phases. The results obtained imply that a multiple-model approach may benefit significantly from a multi-variate treatment (see Read, 2001) that minimizes introducing bias into the optimization process. Furthermore, an *ab initio* modelling approach may benefit from discrete model-building steps, such as atom relocation in electron-density maps and fixing atom assignments, which may speed up the modelling process.

In conclusion, we have illustrated the potentials of Conditional Optimization in both automated model building and *ab initio* phasing of protein structures. Although currently at excessive computational costs, Conditional Optimization holds great promise for protein structure determination by incorporating extensive geometrical prior information without the necessity of an explicit interpretation of the electron-density map.

We thank Dr Hans Raaijmakers for critically reading the manuscript. This work was supported by the Council for Chemical Sciences of the Netherlands Organization for Scientific Research (NWO-CW: Jonge Chemici 99–564).

References

Badger, J. (2003). *Acta Cryst.* D**59**, 823–827.
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. (1998) *Acta Cryst.* D**54**, 905–921.
Cohen, S. X., Morris, R. J., Fernandez, F. J., Ben Jelloul, M., Kakaris, M., Parthasarathy, V., Lamzin, V. S., Kleywegt, G. J. & Perrakis, A. (2004). *Acta Cryst.* D**60**, 2222–2229.
Holton, T., Loerger, T. R., Christopher, J. A. & Sacchettini, J. C. (2000). *Acta Cryst.* D**56**, 722–734.
Huizinga, E. G., Schouten, A., Connolly, T. M., Kroon, J., Sixma, J. J. & Gros, P. (2001). *Acta Cryst.* D**57**, 1071–1078.
Huizinga, E. G., van der Plas, R. M., Kroon, J., Sixma, J. J. & Gros, P. (1997). *Structure*, **5**, 1147–1156.
Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.
Lamzin, V. S. & Wilson, K. S. (1993). *Acta Cryst.* D**49**, 129–147.
Levitt, D. G. (2001). *Acta Cryst.* D**57**, 1013–1019.
Morris, R. J., Perrakis, A. & Lamzin, V. S. (2002). *Acta Cryst.* D**58**, 968–975.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Pannu, N. S., Murshudov, G. N., Dodson, E. J. & Read, R. J. (1998). *Acta Cryst.* D**54**, 1285–1294.
Pannu, N. S. & Read, R. J. (1996). *Acta Cryst.* A**52**, 659–668.
Perrakis, A., Morris, R. & Lamzin, V. S. (1999). *Nature Struct. 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.
Scheres, S. H. W. & Gros, P. (2001). *Acta Cryst.* D**57**, 1820–1828.
Scheres, S. H. W. & Gros, P. (2003). *Acta Cryst.* D**59**, 438–446.
Terwilliger, T. C. (2001). *Acta Cryst.* D**57**, 1755–1762.
Terwilliger, T. C. (2003). *Acta Cryst.* D**59**, 1174–1182.
Terwilliger, T. C. (2004). *Acta Cryst.* D**60**, 2144–2149.
Vandeputte-Rutten, L., Bos, M. P., Tommassen, J. & Gros, P. (2003). *J. Biol. Chem.* **278**, 24825–24830.