Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Conditional optimization: a new formalism for protein structure refinement

Sjors H. W. Scheres and Piet Gros*

Department of Crystal and Structural Chemistry, Bijvoet Center for Biomolecular Research, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands

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

Received 17 July 2001
Accepted 19 September 2001

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

1. Introduction

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

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

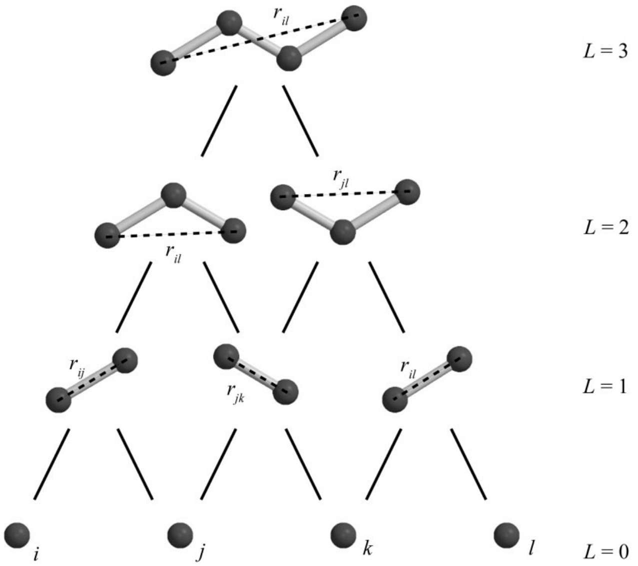

Figure 1
Formation of a peptide plane by binary combinations of four loose atoms, three bonds and two bond angles. For each binary combination of two sub-elements of length $L - 1$ into one element of length L, a condition is assigned. These conditions represent geometrical criteria, depending for example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

Here, we present a new formalism that allows conditional formulation of target functions in structure optimization. Using this formalism, we can express the geometrical information of protein structures in terms of loose atoms. Our approach overcomes the problem that in general a chemical topology cannot be assigned unambiguously to loose atoms. We consider all possible interpretations, based on the structural similarity between the distribution of loose atoms and that of given protein fragments. Weighted geometrical restraints are applied in the optimization according to the extent by which the individual interpretations could be made. In effect, the formalism presented here yields an N-particle solution to the problem of assigning a topology to a given atomic coordinate set. Thereby, the method of conditional optimization combines the search efficiency of loose atoms with the possibility of including large amounts of geometrical information. The information expressed using the conditional formalism includes structural fragments of protein structures from single bonds up to secondary-structural elements. We show that for a simple test case this method yields reliable phases when starting from random atom distributions.

2. Conditional formalism

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

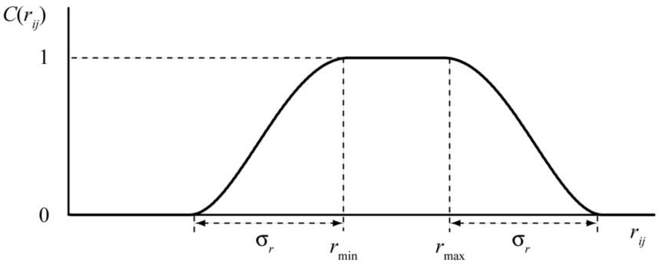

Figure 2
Conditions $C(r_{ij})$ are defined by an optimal range of distances from $r_{\min}$ to $r_{\max}$ and a fourth-order polynomial slope with a width of σ_r: $C(r_{ij}) = 0$ for $r_{ij} \le r_{\min} - \sigma_r$; $C(r_{ij}) = \{1 - [(r_{\max} - r_{ij})/\sigma_r]^2\}^2$ for $r_{\min} - \sigma_r 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

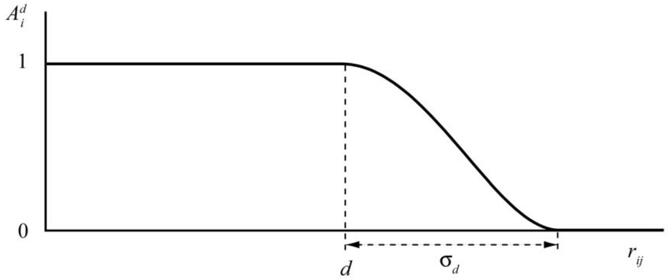

Figure 3
Neighbouring atoms j around atom i are counted using a continuous function n_i^d: $n_i^d(r_{ij}) = 1$ for $r_{ij} \leq d$; $n_i^d(r_{ij}) = \{1 - [(d - r_{ij})/\sigma_d]^2\}^2$ for $d < r_{ij} < d + \sigma_d$; $n_i^d(r_{ij}) = 0$ for $r_{ij} \geq \; d + \sigma_d$. The total number of neighbours, $N_i^d = \sum_j n_i^d(r_{ij})$, is used to calculate a neighbour condition $C_i^0(N_i^d)$. Given an optimal range for the number of neighbouring atoms $N_{\min}$ to $N_{\max}$ and a width σ_N for the fourth-order polynomial slope, this condition can be calculated using the functional form as described in Fig. 2.

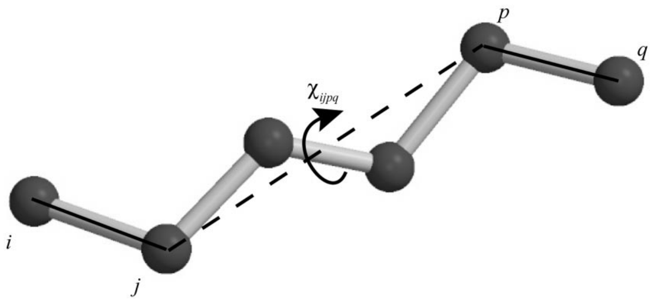

Figure 4
A dihedral angle χ_{ijpq} is defined for the four outermost atoms i, j, p and q of any linear element $ij\ldots pq$ of length $L = \geq 3$. Given an optimal value $\chi_{\rm opt}$ for this dihedral angle, a condition $C_\chi^{\rm type}(\chi_{ijpq})$ can be defined as $C_\chi^{\rm type}(\chi_{ijpq}) = \{1 - [(\chi_{\rm opt} - \chi_{ijpq})/\pi]^2\}^2$.

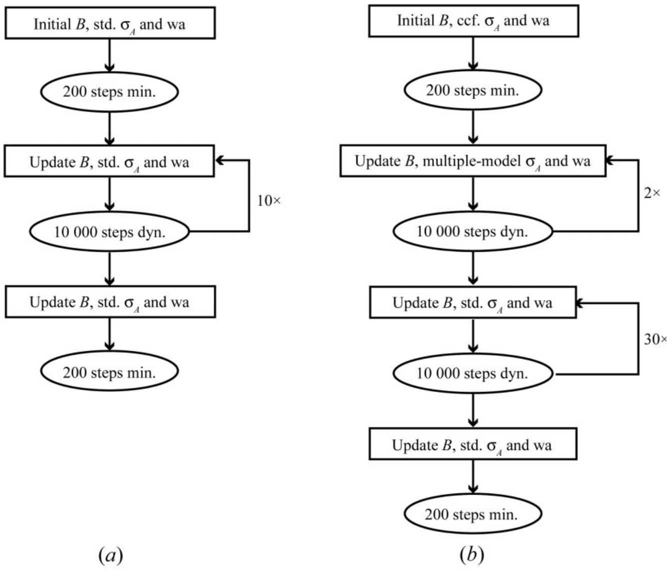
Initial B, std. σ_A and wa
200 steps min.
Update B, std. σ_A and wa
10 000 steps dyn.
10×
Update B, std. σ_A and wa
200 steps min.
(a)
Initial B, ccf. σ_A and wa
200 steps min.
Update B, multiple-model σ_A and wa
10 000 steps dyn.
2×
Update B, std. σ_A and wa
10 000 steps dyn.
30×
Update B, std. σ_A and wa
200 steps min.
(b)

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

the expected number of structural elements present as given by the protein sequence and expected secondary structure.

We define a linear structural element as a non-branched sequence of atoms $ij\ldots pq$ of L bonds long, containing $L + 1$ atoms. A linear structural element of atoms $ij\ldots pq$ of length L is composed of two linear sub-elements $ij\ldots p$ and $j\ldots pq$, both of length $L - 1$ (see Fig. 1). We define conditions C, which are continuous functions with $C = [0, 1]$, assigned to each of these elements. Conditions C reflect the degree to which a geometrical criterion is fulfilled associated with forming a specific type of element from its two sub-elements. When considering only distance criteria, the conditions C become pairwise atomic interaction functions (see Fig. 2). A linear element of length L is then described by a joint condition JC, which is a product of conditions C according to the binary decomposition of the linear element into its sub-elements. Thus, the $(L + 1)$-particle function $JC_{i\ldots q}$ for a linear structure consisting of atoms $i\ldots q$ forming L bonds is expressed in a (binomial) product of $L(L + 1)/2$ pairwise functions.

Fig. 1 shows an example of a binary combination of four atoms i, j, k and l resembling the atoms CA—C—N—CA in a peptide plane. A peptide plane, CA–C–N–CA, is composed of six types of linear elements: bonds CA—C, C—N and N—CA, bond angles CA—C—N and C—N—CA and peptide plane CA—C—N—CA. For each type of element, a pairwise interaction function $C^{\rm type}$ is assigned. The resemblance of the four atoms to a peptide plane can then be expressed by the following multiplication of functions $C^{\rm type}$, yielding joint condition $JC_{ijkl}^{\rm CA-C-N-CA}$ which depends on all six interatomic distances r_{ij}, r_{jk}, r_{kl}, r_{ik}, r_{jl} and r_{il},

$$\begin{aligned} JC_{ijkl}^{\rm CA-C-N-CA} =& C^{\rm CA-C}(r_{ij})C^{\rm C-N}(r_{jk})C^{\rm CA-C-N}(r_{ik})C^{\rm C-N}(r_{jk}) \\ & \times C^{\rm N-CA}(r_{kl})C^{\rm C-N-CA}(r_{jl})C^{\rm CA-C-N-CA}(r_{il}). \quad (1) \end{aligned}$$

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

$$JC_{ijk}^{\rm type} = C^{\rm subtype-A}(r_{ij})C^{\rm subtype-B}(r_{jk})C^{\rm type}(r_{ik}), \quad (2a)$$

$$JC_{ij\ldots pq}^{\rm type} = JC_{ij\ldots p}^{\rm subtype-A}JC_{j\ldots pq}^{\rm subtype-B}C^{\rm type}(r_{iq}), \quad (2b)$$

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

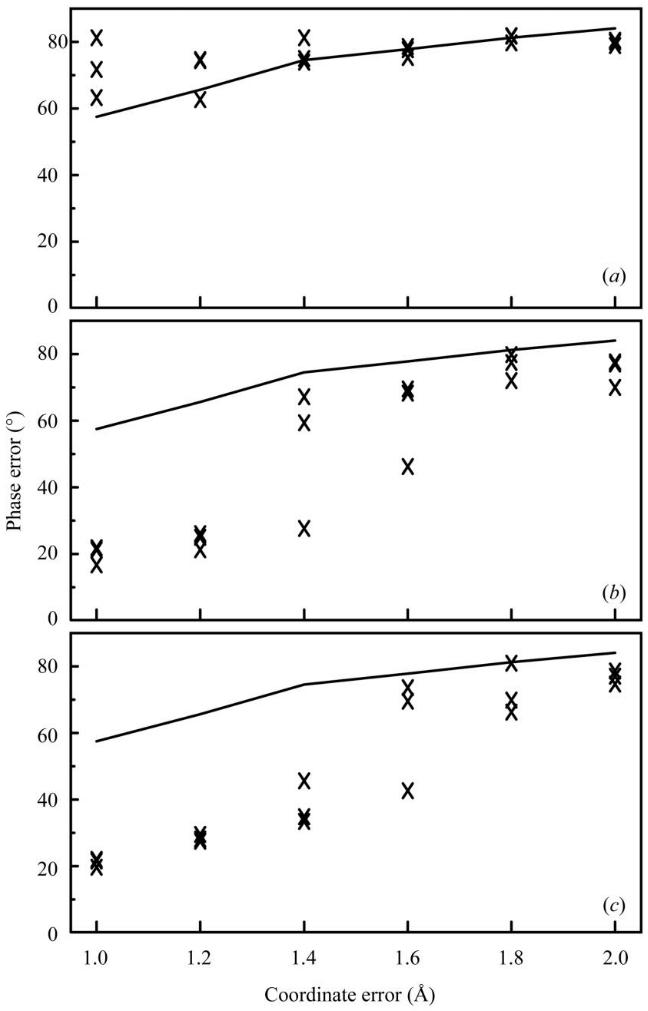

Figure 6
Optimizations of scrambled models with different initial coordinate errors against 2.0 Å resolution diffraction data. Overall amplitude-weighted phase errors are shown for the starting models (solid lines) and the refined structures (crosses) using (*a*) three, (*b*) six and (*c*) nine layers of conditions, where each run was performed three times starting from different random velocities.

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

$$E^{\text{type}} = w^{\text{type}}\left(TC^{\text{type}} - \sum_{ij\ldots pq} JC^{\text{type}}_{ij\ldots pq}\right)^2, \qquad (3)$$

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

$$E = \sum_{\text{type}} E^{\text{type}} = \sum_{\text{type}} w^{\text{type}}\left(TC^{\text{type}} - \sum_{ij\ldots pq} JC^{\text{type}}_{ij\ldots pq}\right)^2. \qquad (4)$$

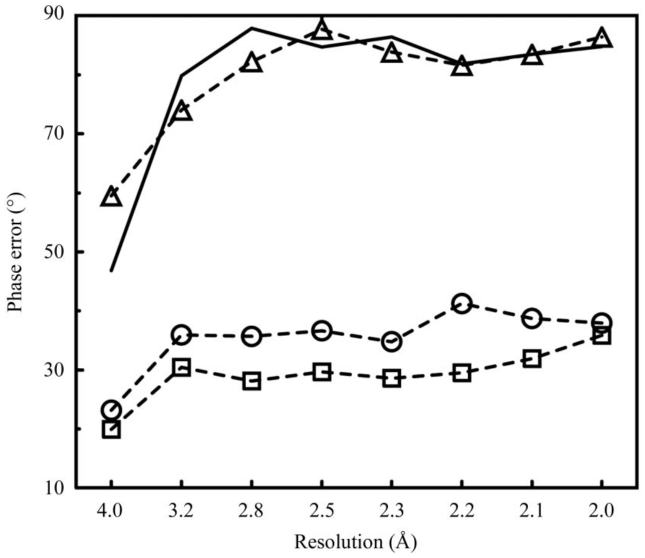

Figure 7
Optimizations of a scrambled model with an initial coordinate error of 1.4 Å r.m.s.d. against 2.0 Å resolution diffraction data. Amplitude-weighted phase errors per resolution shell are shown for the initial model (solid line) and the refined models (dashed lines) using three (triangles), six (squares) and nine (circles) layers of conditions, corresponding to the runs with the lowest overall amplitude-weighted phase error in Fig. 6.

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

$$\frac{\partial}{\partial r_{kl}} \sum_{ij\ldots pq} JC^{\text{type}}_{ij\ldots pq} = \sum_{\ldots k\ldots l\ldots} \frac{nJC^{\text{type}}_{\ldots k\ldots l\ldots}}{C^{\text{subtype}}(r_{kl})}\frac{\partial C^{\text{subtype}}(r_{kl})}{\partial r_{kl}}, \qquad (5)$$

where the summation on the right-hand side runs over linear elements $\ldots k\ldots l\ldots$, which form a subset of linear elements $ij\ldots pq$ that contain both atoms k and l; C^{subtype} is a condition contributing to $JC^{\text{type}}_{\ldots k\ldots l\ldots}$ depending on the interatomic vector r_{kl} and n is the power of C^{subtype} in the binomial distribution of $JC^{\text{type}}_{\ldots k\ldots l\ldots}$. The derivative of the target function given in (3) is

$$\begin{aligned}\frac{\partial E^{\text{type}}}{\partial r_{kl}} &= -2\sum_{\ldots k\ldots l\ldots} w^{\text{type}}\left(TC^{\text{type}} - \sum_{ij\ldots pq} JC^{\text{type}}_{ij\ldots pq}\right)\\ &\quad\times \frac{nJC^{\text{type}}_{\ldots k\ldots l\ldots}}{C^{\text{subtype}}(r_{kl})}\frac{\partial C^{\text{subtype}}(r_{kl})}{\partial r_{kl}}\\ &= G^{\text{type}}_{kl}\frac{1}{C^{\text{subtype}}(r_{kl})}\frac{\partial C^{\text{subtype}}(r_{kl})}{\partial r_{kl}},\end{aligned} \qquad (6)$$

where G^{type}_{kl} is the sum of gradient coefficients from all linear elements depending on $C^{\text{subtype}}(r_{kl})$. (6) shows that the effective weight on a gradient for a particular subtype depends on the extent to which this particular subtype-element is incorporated into larger structural elements. Total gradients can be calculated efficiently because in the summation over all types of linear elements (see equation 4) gradient coefficients G^{type}_{kl} can be pre-calculated for all subtypes, so that for each interacting pair of atoms kl only a summation over the subtypes needs to be performed.

The formulation given above is not restricted to pairwise distance functions. We have extended the description of protein structures with conditions for packing densities and chirality. For all atoms i, atomic conditions $C_i^{\rm atomtype}$ ($L = 0$) are defined, depending on the expected number of neighbouring atoms around an atom of a specific atomtype (see Fig. 3). Thereby, linear elements of a single bond ($L = 1$) are then described by a joint condition

$$JC_{ij}^{\rm type} = C_i^{\rm atomtype\text{-}A} C_j^{\rm atomtype\text{-}B} C^{\rm type}(r_{ij}). \quad (7)$$

Conditions $C_\chi^{\rm type}$ are defined that describe the chirality of linear structures $ij\ldots pq$ with $L \geq 3$ (see Fig. 4). Thereby, (2) becomes

$$JC_{ij\ldots pq}^{\rm type} = JC_{ij\ldots p}^{\rm subtype-A} JC_{j\ldots pq}^{\rm subtype-B} C^{\rm type}(r_{iq}) C_\chi^{\rm type}(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_p, \mathbf{r}_q), \quad (8)$$

where chirality condition $C_\chi^{\rm type}$ depends on the positional vectors $\mathbf{r}_i$, $\mathbf{r}_j$, $\mathbf{r}_p$ and $\mathbf{r}_q$.

3. Experimental

3.1. Implementation

The formalism as described in the previous section has been implemented as a non-bonded routine in the *CNS* program (Brunger, Adams, Clore *et al.*, 1998). A slight modification of (2) is used for the target functions,

$$E^{\rm type} = \frac{\left(TC^{\rm type} - \sum\limits_{ij\ldots pq} JC_{ij\ldots pq}^{\rm type}\right)^2}{TC^{\rm type}} - TC^{\rm type}. \quad (9)$$

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

To compute all non-zero joint conditions, a binary tree is generated starting from the atom-pair list. Joint conditions, see (7) and (8), are computed for all defined types moving from the bottom layer, *i.e.* atoms ($L = 0$), 'upwards' to higher levels of bonded conditions ($L \geq 1$). Energies are computed, see (4) and (9), when all joint conditions are known. Gradients are computed moving 'downwards' from the defined top level to the bottom layer, see (6). The gradient coefficients $G^{\rm type}$ are computed by summation while moving downwards through the binary tree. For each node in the tree the gradient is computed once.

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

3.2. Test case

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

Two types of starting models were generated for testing purposes. Firstly, scrambled starting models with increasing coordinate errors were made by applying random coordinate shifts of increasing magnitude to all atoms in the unit cell. For

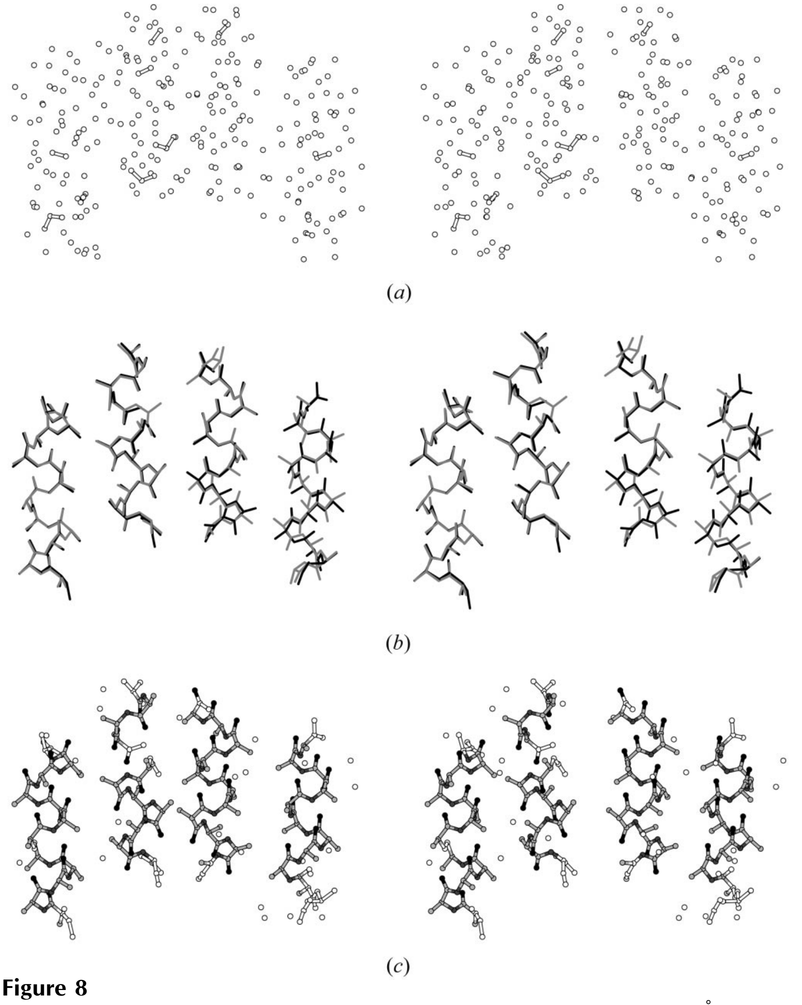

Figure 8
Stereoviews of (*a*) the initial scrambled model with a coordinate error of 1.4 Å r.m.s.d., (*b*) its refined structure (in black) superimposed on the target structure (in grey) and (*c*) the same structure in ball-and-stick representation with automatic assignment of atom types based on the scores of joint conditions (white, unassigned; light grey, carbon; dark grey, nitrogen; black, oxygen). Atoms within 1.8 Å interatomic distance are connected.

Table 1
Conditional force field for alanines in a helical conformation.

(a) Parameters $N_{\min}$, $N_{\max}$ and σ_N for atom types N, CA, C, O and CB, defining the atomic conditions for five neighbour shells with different $d + \sigma_d$ (see Fig. 3).

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

(b) Parameters $r_{\min}$, $r_{\max}$, σ_r and $\chi_{\rm opt}$ (see Figs. 2 and 4), describing the bonded conditions for all types of linear elements with L = 1–9.

Layer	Type (L)	Subtype-A (L − 1)	Subtype-B (L − 1)	$r_{\min}$ (Å)	$r_{\max}$ (Å)	σ_r (Å)	$\chi_{\rm opt}$ (°)
L = 1	N—CA	N	CA	1.43	1.51	0.05	
	CA—C	CA	C	1.51	1.55	0.05	
	C—O	C	O	1.21	1.27	0.05	
	C—N	C	N	1.31	1.35	0.05	
	CA—CB	CA	CB	1.51	1.57	0.05	
L = 2	N—C	N—CA	CA—C	2.41	2.53	0.08	
	CA—O	CA—C	C—O	2.35	2.45	0.08	
	CA—N	CA—C	C—N	2.39	2.49	0.08	
	C—CA	C—N	N—CA	2.39	2.49	0.08	
	O—N	O—C†	C—N	2.21	2.31	0.08	
	O—O	O—C†	C—O	2.10	2.30	0.08	
	N—CB	N—CA	CA—CB	2.39	2.55	0.08	
	CB—C	CB—CA†	CA—C	2.43	2.61	0.08	
L = 3	N—O	N—C	CA—O	3.43	3.61	0.15	138
	N—N	N—C	CA—N	2.71	2.93	0.15	−42
	CA—CA	CA—N	C—CA	3.75	3.87	0.15	178
	C—C	C—CA	N—C	2.91	3.15	0.15	−62
	O—CA	O—N	C—CA	2.69	2.85	0.15	−2
	CB—O	CB—C	CA—O	3.15	3.47	0.15	−98
	CB—N	CB—C	CA—N	3.01	3.37	0.15	82
	C—CB	C—CA	N—CB	3.63	3.79	0.15	174
L = 4	N—CA	N—N	CA—CA	4.11	4.33	0.20	138
	CA—C	CA—CA	C—C	4.29	4.53	0.20	122
	C—O	C—C	N—O	3.69	4.05	0.20	62
	O—C	O—CA	C—C	2.81	3.15	0.20	−58
	C—N	C—C	N—N	3.13	3.47	0.20	−90
	CA—CB	CA—CA	C—CB	4.77	4.99	0.20	−10
	CB—CA	CB—N	CA—CA	4.31	4.71	0.20	−110
	O—CB	O—CA	C—CB	4.17	4.35	0.20	174
L = 5	N—C	N—CA	CA—C	4.59	4.85	0.25	82
	CA—O	CA—C	C—O	5.17	5.51	0.25	142
	O—O	O—C	C—O	3.17	3.71	0.25	14
	CA—N	CA—C	C—N	4.19	4.59	0.25	14
	O—N	O—C	C—N	3.21	3.71	0.25	−114
	C—CA	C—N	N—CA	4.31	4.67	0.25	−6
	N—CB	N—CA	CA—CB	4.81	5.19	0.25	−46
	CB—C	CB—CA	CA—C	5.31	5.61	0.25	−166
	CB—CB	CB—CA	CA—CB	5.13	5.67	0.25	66
L = 6	N—O	N—C	CA—O	5.63	5.97	0.30	158
	CA—CA	CA—N	C—CA	5.27	5.69	0.30	78
	N—N	N—C	CA—N	4.13	4.53	0.30	30
	C—C	C—CA	N—C	4.37	4.75	0.30	22
	O—CA	O—N	C—CA	4.11	4.69	0.30	−50
	CB—O	CB—C	CA—O	6.11	6.51	0.30	−86
	CB—N	CB—C	CA—N	5.47	5.81	0.30	146
	C—CB	C—CA	N—CB	4.99	5.53	0.30	−94
L = 7	CA—C	CA—CA	C—C	5.25	5.77	0.35	90
	C—N	C—C	N—N	3.63	4.07	0.35	−14
	N—CA	N—N	CA—CA	5.13	5.61	0.35	86
	O—C	O—CA	C—C	3.83	4.39	0.35	−30
	C—O	C—C	N—O	5.43	5.85	0.35	98
	CB—CA	CB—N	CA—CA	6.65	7.05	0.35	−166
	CA—CB	CA—CA	C—CB	5.57	6.27	0.35	−18
	O—CB	O—CA	C—CB	5.05	5.81	0.35	−138
L = 8	N—C	N—CA	CA—C	5.43	5.85	0.40	130
	O—O	O—C	C—O	4.73	5.26	0.40	22
	CA—O	CA—C	C—O	6.37	6.91	0.40	130
	CA—N	CA—C	C—N	4.27	4.85	0.40	42
	C—CA	C—N	N—CA	4.33	4.87	0.40	22
	O—N	O—C	C—N	2.99	3.65	0.40	−70
	CB—CB	CB—CA	CA—CB	7.05	7.66	0.40	130
	N—CB	N—CA	CA—CB	5.05	5.77	0.40	22
	CB—C	CB—CA	CA—C	6.71	7.15	0.40	−122
L = 9	N—O	N—C	CA—O	6.65	7.09	0.45	166
	CA—CA	CA—N	C—CA	4.85	5.55	0.45	74
	C—C	C—CA	N—C	4.67	5.17	0.45	90
	N—N	N—C	CA—N	4.61	5.07	0.45	110
	O—CA	O—N	C—CA	3.47	4.09	0.45	−38
	CB—O	CB—C	CA—O	7.79	8.27	0.45	−58
	CB—N	CB—C	CA—N	5.71	6.31	0.45	−114
	C—CB	C—CA	N—CB	3.97	4.81	0.45	−22

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

these starting structures a minimum interatomic distance of 1.4 Å was enforced. Secondly, random atom distributions were made by randomly placing 264 atoms in the unit cell, while enforcing a minimum interatomic distance of 1.8 Å. All atoms in the starting structures were given equal labels and carbon scattering factors were assigned to all of them.

3.3. Refinement protocols

The refinement protocols for optimization starting from the scrambled models and random models are given in Figs. 5(a) and 5(b). These optimization protocols include standard procedures: overall B-factor optimization and weight determination for the X-ray restraint followed by maximum-likelihood optimization by either energy minimization or dynamics simulation. Table 1 contains the set of parameters defining the conditional force field; target values for packing densities and interatomic distances were determined from their distributions in several high-resolution structures in the PDB. Up to nine layers of bonded conditions have been defined, corresponding to linear elements up to, for example, $C^{\alpha}(i)$ to $C^{\alpha}(i + 3)$. During the optimization, the width σ_r of the

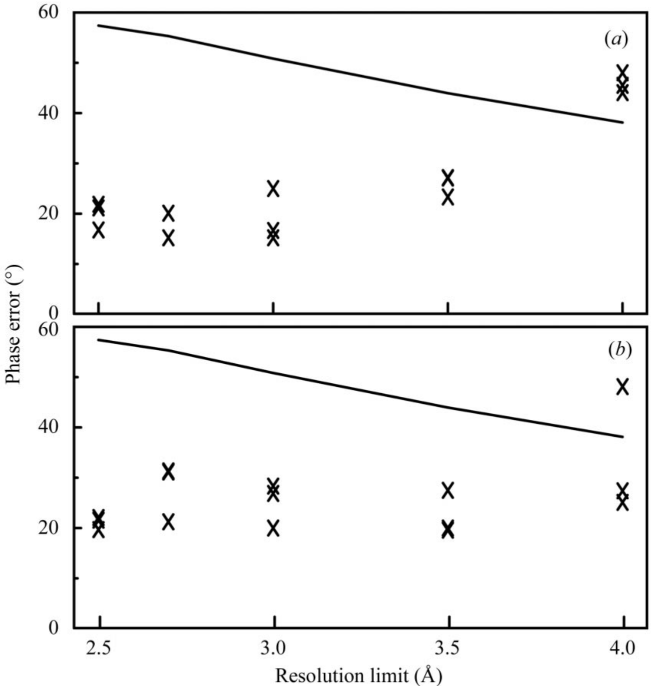

Figure 9
Optimizations of a scrambled model with an initial coordinate error of 1.0 Å r.m.s.d. against diffraction data with different high-resolution limits. The overall amplitude-weighted phase errors are shown for the initial model (solid lines) and the refined structures (crosses) using (*a*) six and (*b*) nine layers of conditions, where each run was performed three times starting from different random velocities.

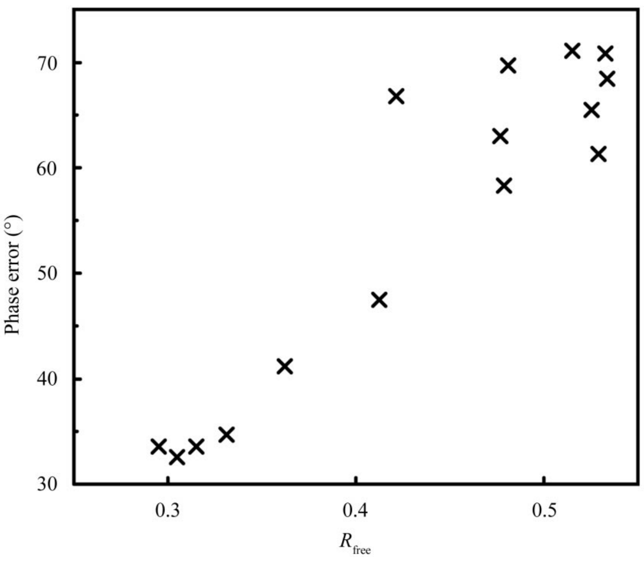

Figure 10
Scatter plot of the amplitude-weighted phase error *vs.* the free *R* factor for the 15 final models that were obtained starting from random atom distributions.

conditional functions was adjusted according to the estimated coordinate error (ε_r) derived from the estimated σ_A values: $\sigma_r' = \sigma_r + \varepsilon_r L^{1/2}$. Atomic *B* factors were assigned using an exponentially decreasing function depending on the number of neighbours N_i^d within a shell d ($+\sigma_d$) of 4.3 (+0.7) Å: $B_i = 150\exp(-0.1N_i^d)$, with a minimum value of 15 Å^2. The time step in these calculations was 0.2 fs and during the dynamics calculations the temperature was coupled to a temperature bath (T_{bath} = 300 K).

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

For optimization starting from randomly placed atoms, all X-ray data to 2.0 Å resolution were included. Compared with the optimization of scrambled models, three modifications were made: alternative protocols were defined for estimating σ_A values and for handling the 'test-set' reflections and to allow faster sampling T_{bath} was set to 600 K. Standard σ_A estimates are based on the correlation coefficient between observed and calculated normalized structure factors, E_{obs} and E_{calc} (Read, 1986). For random atom distributions and structures very far away from the correct answer, the binwise correlation coefficients on normalized structure factors yield spuriously high values. We used a multiple-model approach to obtain estimates of the phase error $\varphi_{\text{obs}} - \varphi_{\text{calc}}$ in the theoretical values for σ_A: $\sigma_A = \langle|E_{\text{obs}}||E_{\text{calc}}|\cos(\varphi_{\text{obs}} - \varphi_{\text{calc}})\rangle$ (Srinivasan & Parthasaraty, 1976). Starting from the coordinate set corresponding to F_{calc}, four dynamics runs of 1000 steps each were performed at an elevated temperature of 900 K using different random starting velocities (yielding structure factors sets F_i). From the resulting four models, we compute the average structure factor F_{ave} and figure of merit m_{ave} ($m_{\text{ave}} = |F_{\text{ave}}|/\langle|F_i|\rangle$). By rewriting ($\varphi_{\text{obs}} - \varphi_{\text{calc}}$) = ($\varphi_{\text{obs}} - \varphi_{\text{ave}}$) + ($\varphi_{\text{ave}} - \varphi_{\text{calc}}$) and assuming $\cos(\varphi_{\text{obs}} - \varphi_{\text{ave}}) \simeq m_{\text{ave}}$, we can estimate σ_A. For a range of test structures far away from the known answer, these estimates had a reasonable correlation to the theoretical values as calculated using known phases φ_{obs} of the test cases. The second feature deviating from normal crystallographic refinement protocols was the handling of the test-set reflections. A conventional test set comprising 7% of all reflections was used to calculate R_{free} and to estimate cross-validated σ_A values according to Pannu & Read (1996) in the later stages of refinement. Additionally, another 7% of the reflections were taken out of the refinement. After every 1000 steps, the selection of these 7% was modified. As a result, the reflections used in the crystallographic target function changed every 1000 steps, resulting in a 'tacking' behaviour during refinement and minimizing the chance of stalled progress owing to local minima in the crystallographic target function.

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

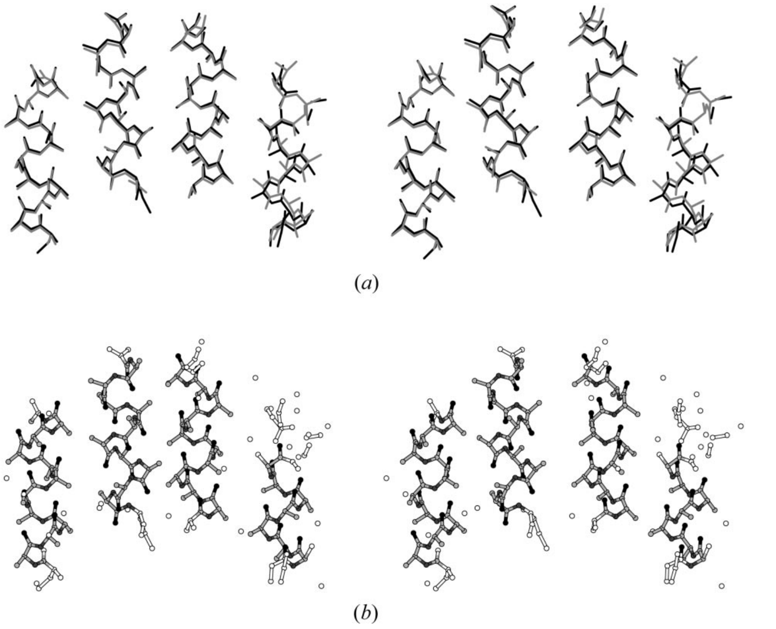

Figure 11
Stereoviews of (*a*) a successfully refined structure starting from a random atom distribution (in black) superimposed on the target structure (in grey) and (*b*) the same structure in ball-and-stick representation with automatic assignment of atom types based on the scores of joint conditions (white, unassigned; light grey, carbon; dark grey, nitrogen; black, oxygen). Atoms within 1.8 Å interatomic distance are connected.

4. Results

4.1. Refinement of scrambled models

Six scrambled models with coordinate errors of 1.0, 1.2, 1.4, 1.6, 1.8 and 2.0 Å root-mean-square deviation (r.m.s.d.) were generated. The dependence of the method on the number of conditional layers was tested performing a series of refinements using three, six or nine layers. The resulting amplitude-weighted phase errors are shown in Fig. 6. Three layers of conditions are not enough to give significant phase improvement. Using six layers, scrambled models with r.m.s.d.s up to 1.4 Å could be improved significantly. Adding another three layers of conditions led to a small increase in the success rate. Fig. 7 shows the phase improvement for the refined 1.4 Å r.m.s.d. structure with the lowest free *R* factor using three, six or nine layers of conditions. Fig. 8 shows an initial model with a coordinate error of 1.4 Å r.m.s.d and the refined structure with the lowest free *R* factor using nine layers of conditions. This structure is representative for all successful runs: the four helices are clearly visible, although some are not completed, contain breaks in the main chain or the N—C direction is reversed. For structures with a coordinate error larger than 1.4 Å r.m.s.d., refinement did not yield improvement of the phases. This coincides with the observation that for models with large errors the *SIGMAA* procedure (Pannu & Read, 1996) gave spurious estimates for the σ_A values (results not shown).

The dependence on the high-resolution limit of the diffraction data was tested by refining the 1.0 Å r.m.s.d. model using data truncated at various resolution limits. Calculations were performed using six or nine layers of conditions. The resulting phase improvements are shown in Fig. 9. All runs using data to a resolution of 3.0 Å were successful. When using only 3.5 Å data all three runs using six layers of conditions failed, while using nine layers of conditions resulted in a success rate of two out of three.

4.2. Refinement of random atom distributions

Sixteen different random atom distributions were refined according to the protocol in Fig. 5(*b*). One run was abandoned because standard σ_A estimates could not be obtained by the *SIGMAA* procedure after the initial 20 000 steps. Of the remaining 15 models, six yielded a final amplitude-weighted phase error of smaller than 50° for data to 2.0 Å resolution. This corresponds to a success rate of one out of three. For these successful runs a condensation into four rod-like structures was observed during the initial stages of the refinement process, thereby establishing a choice of origin for the triclinic cell. Subsequent dynamics optimization lead to the formation of helical fragments that were expanded into near-complete α-helices. Fig. 10 shows a clear correlation between the phase errors and the overall free *R* factor obtained for the final models. The structure with the lowest free *R* factor is shown in Fig. 11[1]. This structure clearly shows the four α-helices and resembles the results obtained from the refinement of the scrambled models. The errors in the model include chain breaks, incomplete helices and chain reversals.

5. Discussion

We introduced a new method for optimization of protein structures that overcomes the necessity of a fixed topology for defining geometrical restraints. This *N*-particle approach offers a 'restrained topology', where weighted gradients over all possible assignments are applied to loose atoms. We tested this method using calculated data and a very simple test case consisting of four polyalanine helices with 244 non-H atoms in total. Optimizations starting from scrambled models show that the method works successfully with diffraction data of at least 3.0 to 3.5 Å resolution and with six or nine layers of conditions, corresponding to linear structural elements of the length of two and three peptide planes, respectively. Moreover, we have shown that our test structure can be optimized successfully starting from randomly distributed atoms when using 2.0 Å resolution diffraction data. Important for successful optimization of random starting models was estimation of reasonable σ_A values for very bad models using a multiple-

[1] A movie showing the formation of the four helices starting from a random atom distribution is available from the IUCr electronic archive (Reference: jn0096) and can be viewed in the online version of this paper. Details on how to access these data are available at the back of the journal.

model procedure. For trials with different random starts a success rate of one out of three was observed. The free *R* factor readily distinguished correct solutions from false ones. To our knowledge, we have presented the first method that in principle allows an *ab initio* optimization of atomic models under conditions relevant for protein crystallography (*i.e.* at medium resolution).

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

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

References

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

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

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

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

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

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

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

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

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.

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

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

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

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

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