Chapter 3

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

Abstract

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

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

3.1 Introduction

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

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

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

3.2 Mean-force potential for protein structures

3.2.1 Brief review of the conditional formalism

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

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

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

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

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

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

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

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

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

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

3.2.2 The general parameter set

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

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

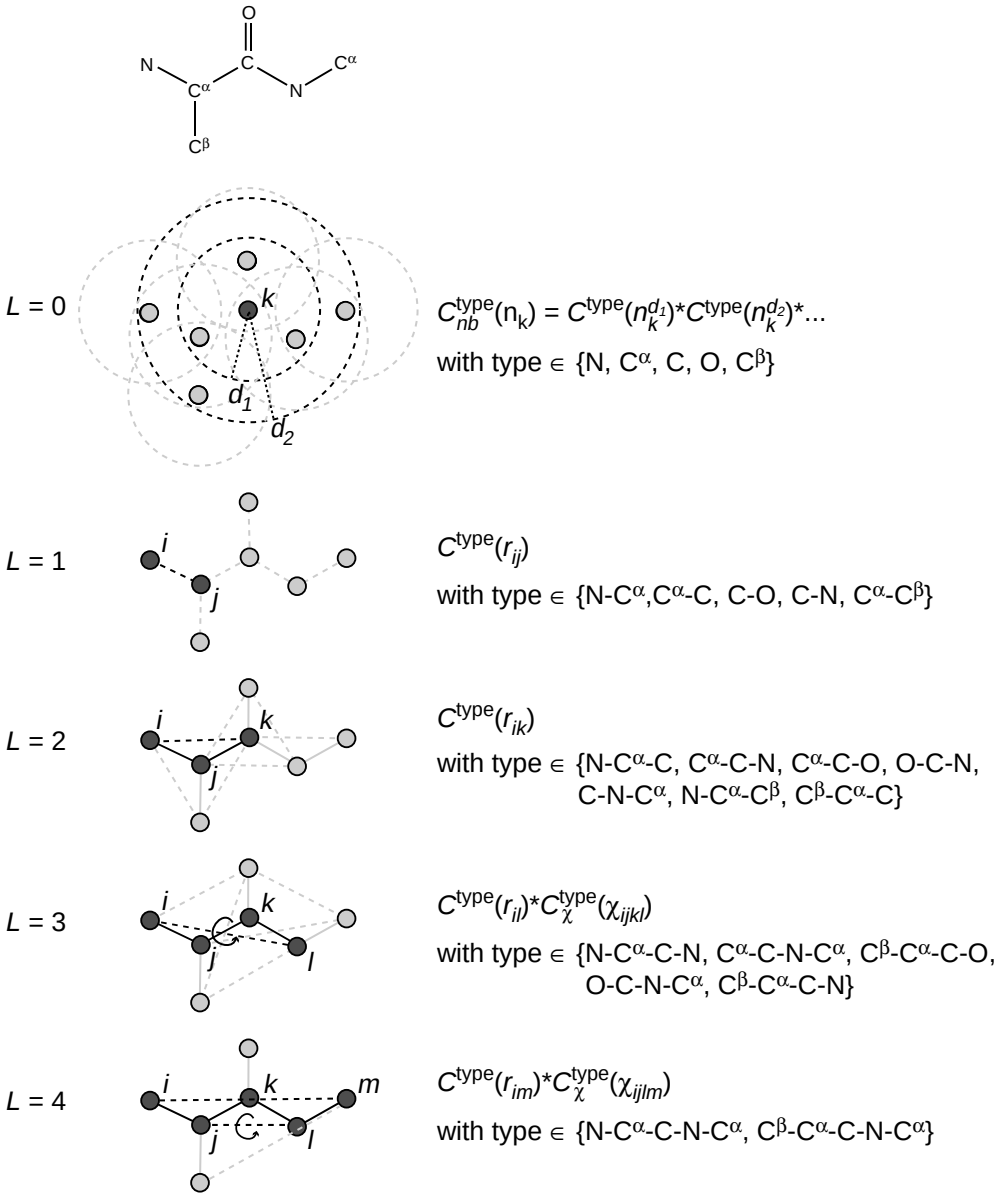

Figure 3.1: *Schematic representation of the conditions defining a linear element N-CA-C-N-CA. Shown are a protein fragment containing this linear element and schematic representations of the conditions involved. These conditions depend on number of neighbouring atoms n (for two shells with radii d_1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

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

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

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

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

3.2.3 Protein-specific force fields

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

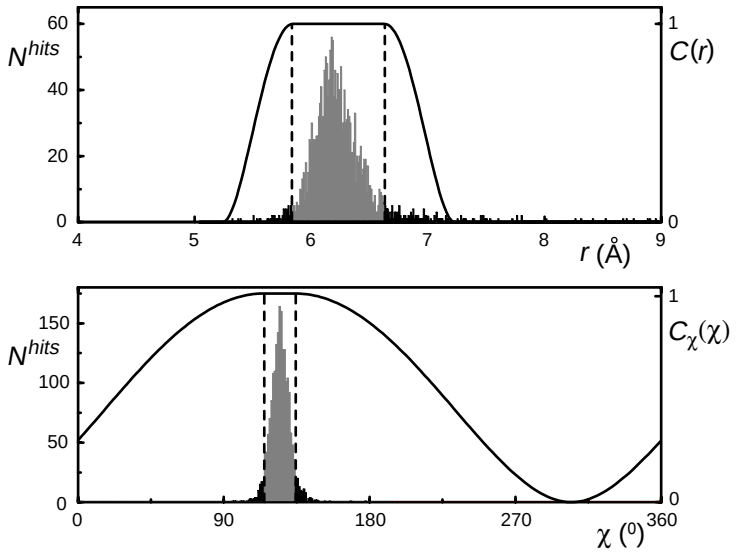

Figure 3.2: *Observed distributions N_{hits} and defined conditions C for (a) inter-atomic distance r and (b) C_χ for torsion angle χ between the outermost atoms of a linear element comprising atoms $C^\alpha(i)$ until $C^\alpha(i+4)$ in an α-helical conformation. Minimum and maximum values for which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

grouped (3.3):

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

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

3.3 Experimental

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

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

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

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

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

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

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

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

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

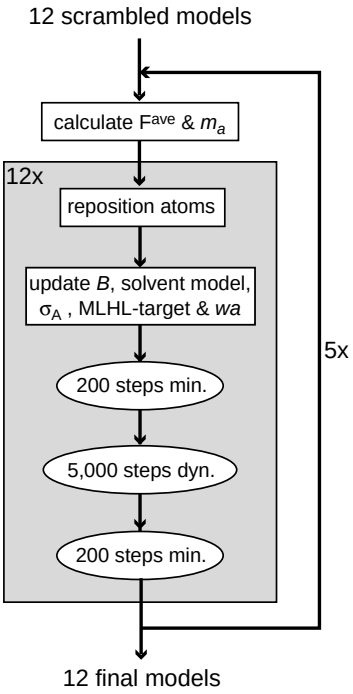

Figure 3.3: *Refinement protocol for the optimization starting from twelve scrambled structures. Prior to every optimization cycle, average structure factor* F^{ave} *and figures of merit* m_a *were calculated from the 12 individual structure factor sets* F^i*. At every cycle a small amount of atoms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

3.4 Results & discussion

3.4.1 Stability of correct structures

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

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

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

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

3.4.2 Searching behaviour in optimization

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

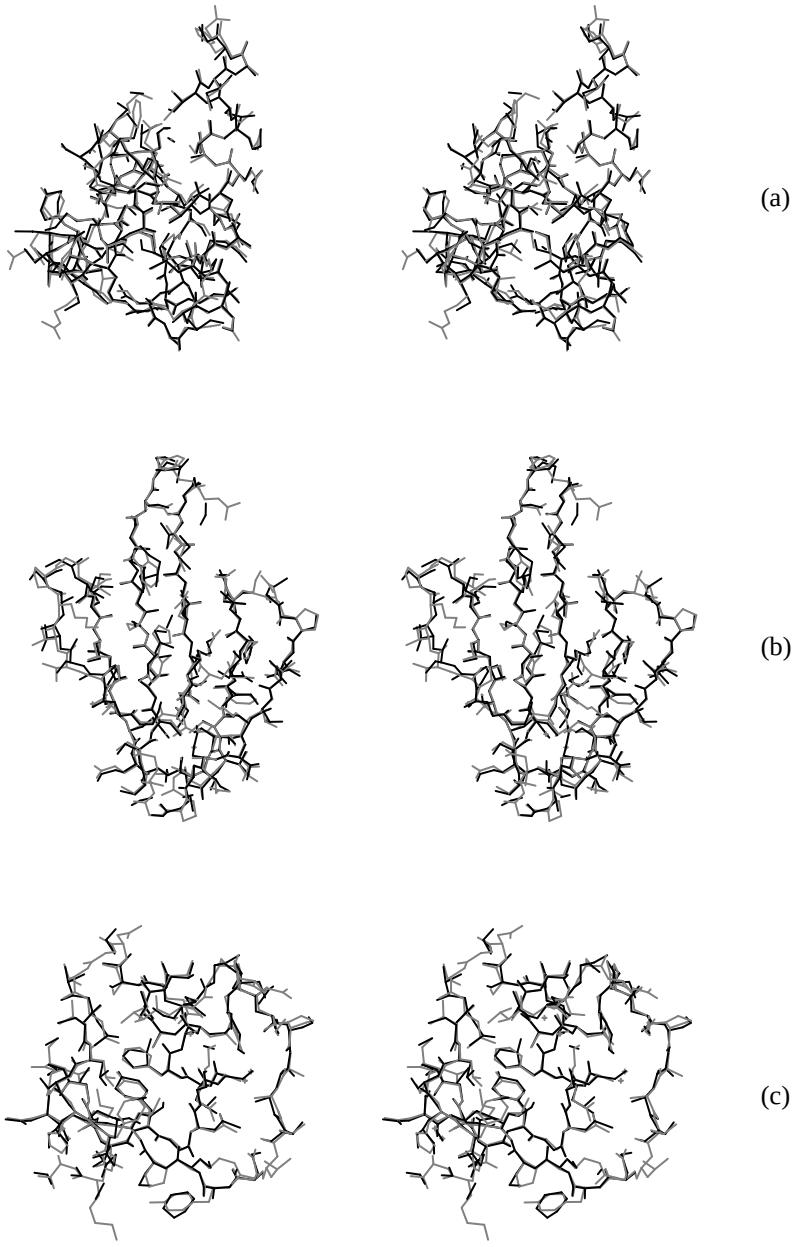

Figure 3.4: *Stereo-views of equilibrated structures in black superimposed on the target structures in gray of the (a) all-α helical case, 1I2T, (b) all-β sheet, 3EBX, and (c) mixed α/β, 1DS3, test cases.*

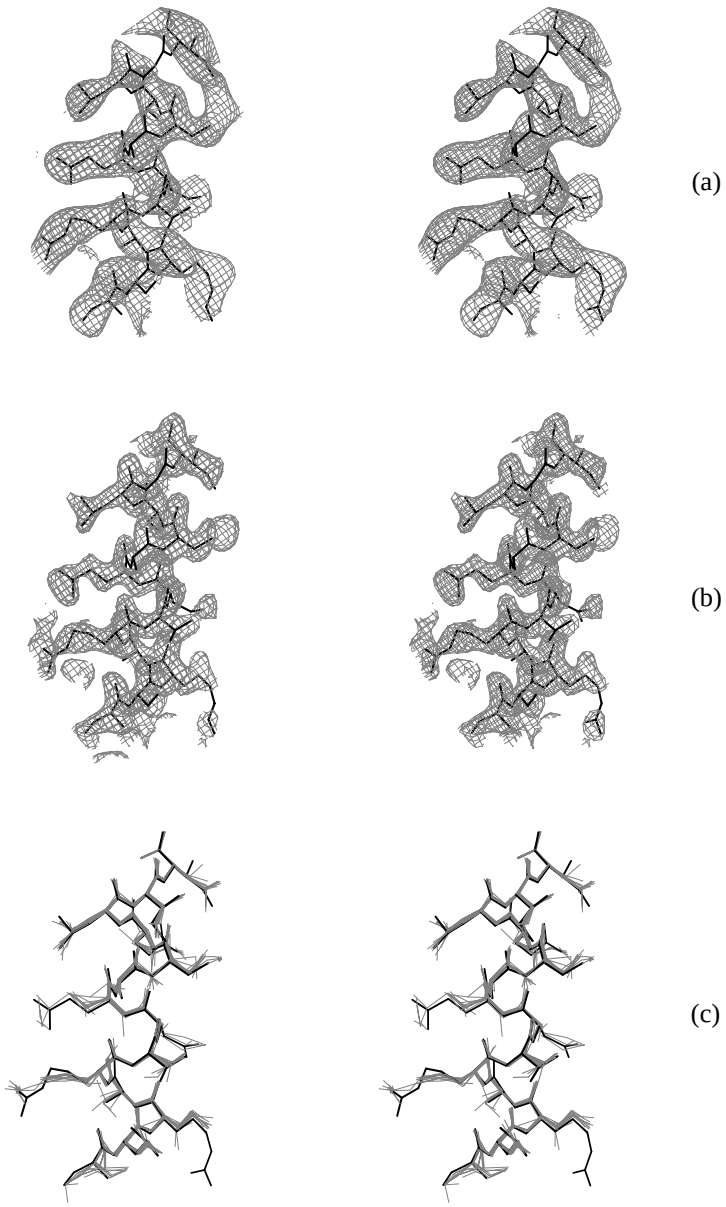

Figure 3.5: *Electron-density maps and models obtained by conditional optimization of 1I2T using twelve scrambled models. Shown are stereo-views of part of the* $m_a|F^{\mathrm{obs}}|\exp(i\varphi^{\mathrm{ave}})$*-electron density maps obtained before (a) and after (b) optimization, and the twelve final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

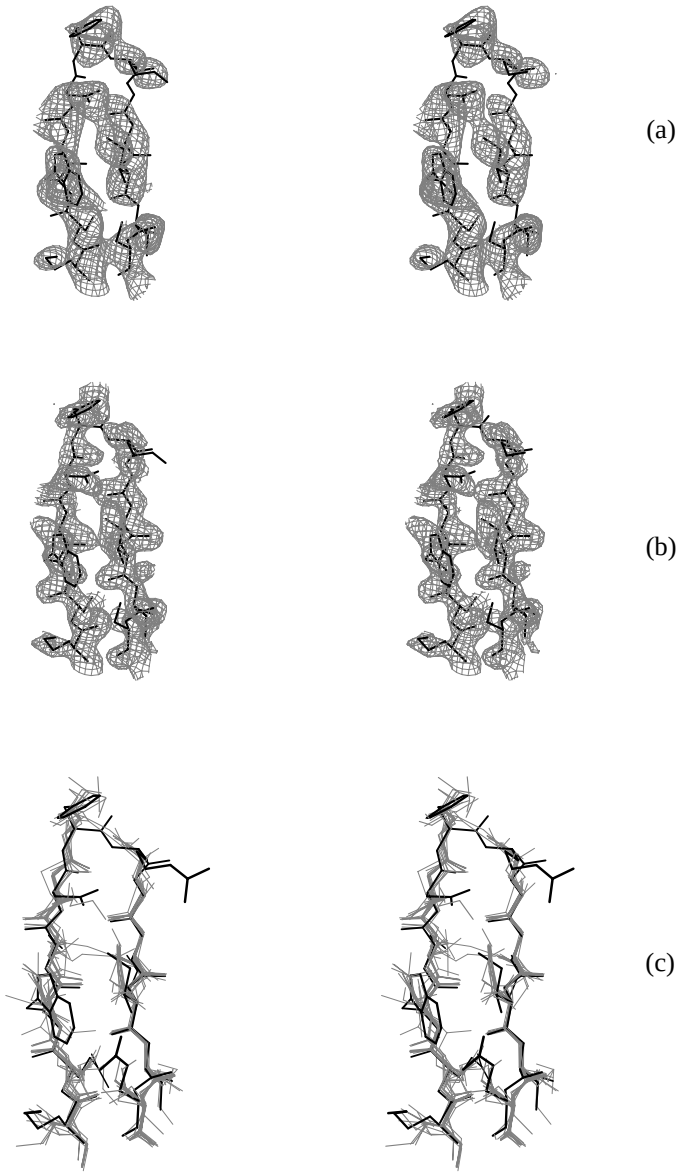

Figure 3.6: *Electron-density maps and models obtained by conditional optimization of 3EBX using twelve scrambled models. Shown are stereo-views of part of the* $m_a|F^{\mathrm{obs}}|\exp(i\varphi^{\mathrm{ave}})$*-electron density maps obtained before (a) and after (b) optimization, and the twelve final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

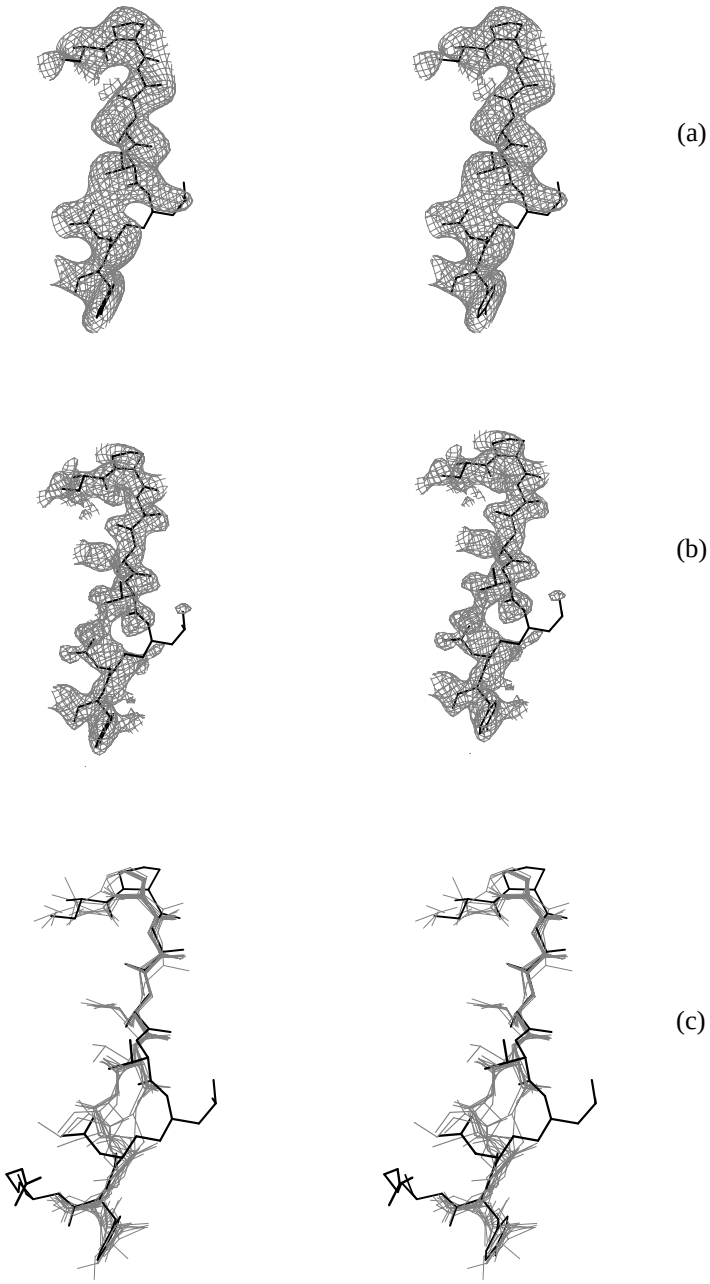

Figure 3.7: *Electron-density maps and models obtained by conditional optimization of 1DS3 using twelve scrambled models. Shown are stereo-views of part of the* $m_a|F^{\mathrm{obs}}|\exp(i\varphi^{\mathrm{ave}})$*-electron density maps obtained before (a) and after (b) optimization, and the twelve final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

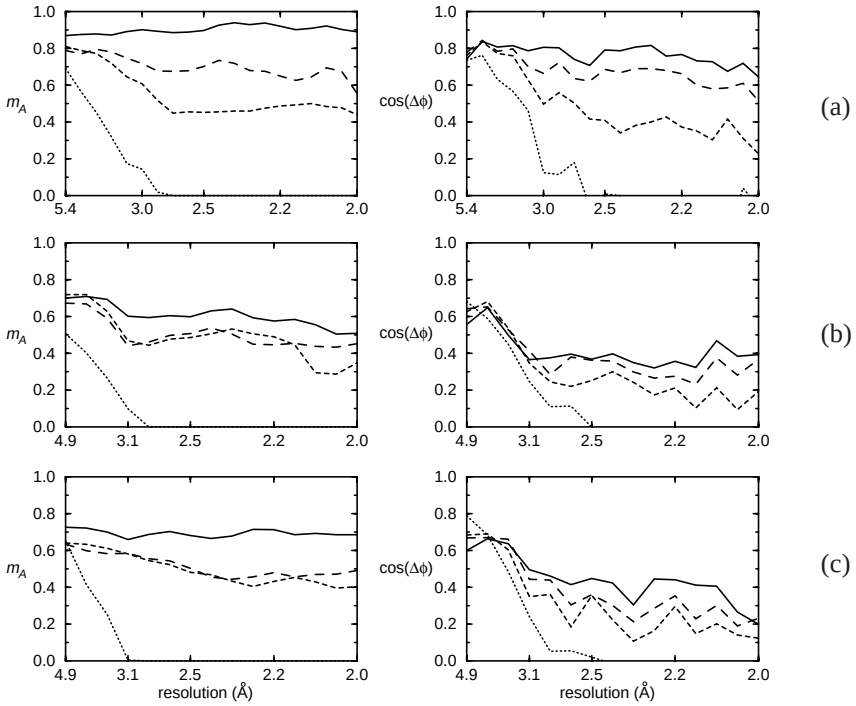

Figure 3.8: *Estimated figures of merit m_a and average cosine of the true phase error, $\langle\cos(\Delta\varphi)\rangle$, for the (a) all-α helical, 1I2T, (b) all-β sheet, 3EBX, and (c) mixed α/β, 1DS3, test cases. Values are shown as calculated for the initial, scrambled structures (dotted lines),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

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

3.5 Conclusions

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

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

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

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

Acknowledgements

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