Theor Chem Acc (2001) 106:69–75
DOI 10.1007/s002140000221

Theoretical
Chemistry Accounts

Regular article

Ab initio structure predictions using a hierarchical approach applied to 434 cro and the *Drosophila* homeodomain

Dimitri Gilis, Marianne Rooman

Ingénierie Biomoléculaire, Université Libre de Bruxelles, CP 165/64, 50 av. Roosevelt, 1050 Bruxelles, Belgium

Received: 6 July 2000 / Accepted: 8 September 2000 / Published online: 21 December 2000

Abstract. A discrete-state ab initio protein structure prediction procedure is presented, based on the assumption that some proteins fold in an hierarchical way, where the early folding of independent units precedes and helps complete structure formation. It involves a first step predicting, by means of threading algorithms and local structure prediction methods, the location of autonomous protein subunits presenting favorable local and tertiary interactions. The second step consists of predicting the structure of these units by Monte Carlo simulated annealing using several database-derived potentials. In a last step, these predicted structures are used as starting conformations of additional simulations, keeping these structures frozen and including the complete protein sequence. This procedure is applied to two small DNA-binding proteins, 434 cro and the *Drosophila melanogaster* homeodomain that contain 65 and 47 residues, respectively, and is compared to the nonhierarchical procedure where the whole protein is predicted in a single run. The best predicted structures were found to present root-mean-square deviations relative to the native conformation of 2.7 Å in the case of the homeodomain and of 3.9 Å for 434 cro; these structures thus represent low-resolution models of the native structures. Strikingly, not only the helices were correctly predicted but also intervening turn motifs.

Key words: Protein folding – Monte Carlo simulated annealing – Database-derived potentials

1 Introduction

Ab initio protein structure prediction, where the only input is the amino acid sequence, is still a largely unsolved problem, despite many interesting approaches and progresses. The difficulty resides in the enormous size of the conformational space to be explored, which excludes an exhaustive search, and in the imperfectness of the potentials used to evaluate the free energy of the conformations generated. The problem is that the more simplifications are made to reduce the conformational space and thus to render the problem more tractable, the less precise the potentials and thus the predictions become. The most accurate approach, at least in principle, consists of solving numerically the classical equations of motion for all atoms in a protein, using semiempirical potentials to evaluate the interactions between the atoms. However, besides the imperfectness of the potentials and the problems related to numerical integration over large timescales, the main obstacle for such molecular dynamics simulations is the astronomical computer time needed. These simulations are thus, and will probably remain for a long time, restricted to small systems and supercomputers [1].

There is thus a need for faster ab initio prediction methods, whose purpose is to yield low-resolution 3D structures that could be refined afterwards by more accurate procedures, such as restricted molecular dynamics simulations. This need is quite crucial as the number of sequences determined grows much faster than the number of structures determined.

Such ab initio prediction methods have been developed for several decades, devising and exploiting diverse techniques. To reduce the conformational space to be explored, hierarchical approaches have been developed. One of them consists of predicting first the secondary structures and then their assembly into a compact fold [2–4]; however, it suffers from the fundamentally limited scores of secondary structure predictions [5]. The buildup procedure is another type of hierarchical approach. It predicts the lowest free-energy conformations of small segments and assembles them to reconstruct the full protein [6–8]; its limited success is in part due to its assumption that all protein regions are dominated by local interactions along the sequence. Promising extensions of this approach have been developed in which both local

Correspondence to: D. Gilis
e-mail: dgilis@ulb.ac.be

Contribution to the Symposium Proceedings of Computational Biophysics 2000

and nonlocal interactions are taken into account [9] or primitive folding modules constituted of neighboring chain sites further interact in an iterative fashion [10].

Nonhierarchical approaches must sample the conformational space more widely. Molecular dynamics simulations with simplified energy functions can be used [11], but a more often used sampling technique is Monte Carlo (MC) [12], in which trial moves are generated randomly and accepted or rejected according to the Boltzmann weight [13–17]. To limit trapping in local minima, simulated annealing [18], consisting of gradually decreasing the temperature, is often performed [19–22]. An MC-related algorithm that works on populations of conformations is the genetic algorithm [23, 24]. The conformational space can also be reduced by placing the protein on lattices [14, 25, 26], by adding constraints on tertiary contacts or secondary structures [22, 27, 28], or by including restraints derived from multiple sequence alignments [27].

The performances of these ab initio methods are encouraging but remain limited. Indeed, the assessment of the blind predictions performed during the CASP III meeting [29] indicated that except in a small percentage of predictions, the final model was quite distant from the native structures. Only for some protein fragments were good predictions found, as monitored by backbone root mean square (rms) deviations of less than 4 and 5 Å for fragments of 25 and 40 residues, respectively. The easiest targets were mainly small α-helical proteins and the hardest targets were larger proteins containing a β sheet. This score still leaves place for new methods and developments.

In this article we report ab initio predictions on two small proteins using a hierarchical procedure that bears some resemblance to the approach of Ref. [10], MC simulated annealing and database-derived potentials. The proteins considered are DNA-binding proteins of the helix-turn-helix (HTH) type, 434 cro, and a homeodomain. We thus focus on all-α proteins, but the procedure is, in principle, applicable to β and $\alpha\beta$ proteins; this will be considered elsewhere.

2 Prediction protocol

The prediction protocol used here proceeds through three steps. The first is to identify independent folding units, which are more or less stable in the absence of interactions with the rest of the chain and are likely to form at the beginning of the folding process; we shall call the so-defined units foldons [30]. These units typically possess two or three secondary structure elements; they constitute either the whole sequence or a part of it. The second step consists of performing MC simulated annealing simulations on the foldons using combinations of database-derived potentials describing local and tertiary interactions. The feasibility of these first two steps has been demonstrated in Ref. [31]. If the foldon does not cover the whole sequence, there is a last step that consists of lengthening the predicted sequence by adding flanking stretches and performing MC simulated annealing simulations on the lengthened sequence, while keeping the predicted foldon structure frozen. These different steps are detailed later.

This protocol is based on the assumption that at least some proteins fold hierarchically, with some protein subunits folding early and forming compact units, presenting some – yet marginal – stability around which the rest of the chain can assemble. Note that larger, multidomain proteins may contain several nondirectly interacting foldons, but we will not consider such cases here.

For comparison, we also test nonhierarchical predictions, by performing MC simulated annealing directly on the whole protein sequence.

3 Identification of independently stable units (foldons)

To identify potential foldons, we use a combination of local-structure prediction methods and fold recognition procedures, requiring favorable local and tertiary interactions. We typically search for foldons having two to three secondary structure elements, because smaller units might have insufficient stability and lifetime to allow further coalescence and larger domains might possess several foldons. For the simplicity of the procedure, we restrict ourselves to foldons constituted of secondary structure elements that are consecutive along the sequence.

The local-structure prediction methods used are Prelude and Fugue [32], where protein structures are described as strings of (φ, ψ, ω) domains; seven domains A, C, B, P, G, E, and O are considered, defined in the legend to Fig. 1, which allow any type of local structure to be represented. Prelude predicts all the lowest free-

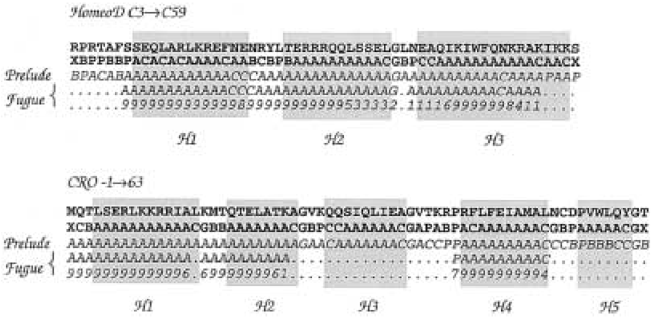
HomeoD C3→C59
RPRTAFSSEQLARLKREFNENRYLTERRRQQLSSELGLNEAQIKIWFQNKRAKIKKS
XBPPBBPACACACAAAACAABCBPBAAAAAAAAAAACGBPCCAAAAAAAAAAAACAACX
Prelude
BPACABAAAAAAAAAAAAACCCAAAAAAAAAAAAAAAAGAAAAAAAAAAAACAAAAPAAP
Fugue
......AAAAAAAAAAAAACCCAAAAAAAAAAAAAAAAG.AAAAAAAAAACAAAA....
......9999999999999899999999999995333332.11116999999998411....
H1
H2
H3
CRO -1→63
MQTLSERLKKRRIALKMTQTELATKAGVKQQSIQLIEAGVTKRPRFLFEIAMALNCDPVWLQYGT
XCBAAAAAAAAAAAACGBBAAAAAAAACGBPCCAAAAAAACGAPABPACAAAAAAAACGBPAAAAACGX
Prelude
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAGAACAAAAAAAACGACCPPAAAAAAAAAACCCBPBBBCCGB
Fugue
AAAAAAAAAAAAAAA.AAAAAAAAAA....................PAAAAAAAAAC...........
999999999999996.6999999961....................79999999994...........
H1
H2
H3
H4
H5

Fig. 1. Structure prediction of CRO and HomeoD using the programs Prelude and Fugue [32]. The amino acid sequence is given in the first line. The next three lines indicate domains in (φ, ψ, ω) torsion angles using the one-letter code A, C, B, P, G, E, and O [32], with X indicating undetermined (φ, ψ, ω) angles. O represents cis conformations ($\omega \approx 0°$) and the other six domains trans conformations ($\omega \approx 180°$). A and C represent α- and 3_{10}-helical conformations, respectively, B and P extended β-type and polyproline-like conformations, and G and E positive φ conformations mirror-symmetrical to A/C and B/P. The first of these three lines contains the (φ, ψ, ω) domains in the native structure, the second the predictions of Prelude and the third the predictions of Fugue. In the Fugue predictions, dots indicate regions without strong predictions; in the subsequent line, the strength of the predictions is given as a number between 0 and 9. *Grey boxes* correspond to α helices; their definition is that of DSSP [45] lengthened at both ends as long as the (φ, ψ, ω) angles remain in the A or the C conformation

energy conformations of a given sequence on the basis of database-derived torsion potentials describing local interactions along the chain (see Sect. 4) and orders them as a function of increasing free energy. Fugue uses the Prelude predictions to identify strongly predicted segments. The strength of the predictions is measured by an integer between 0 and 9 called confidence. Segments with high confidence values are likely to adopt a preferred conformation when excised from the rest of the chain or to be formed at the very beginning of the folding process [32–34]. They typically correspond to a helical or extended stretch of five-to-ten residues or to a turn, involve few or no tertiary interactions, and present only a marginal stability. This stability is nevertheless sufficient to restrict the conformational space and to facilitate the formation of the foldon.

The fold recognition procedure used is called metaFoRe [35]. It proceeds by threading the target sequence onto a dataset of known protein structures and by estimating the sequence–structure compatibility on the basis of database-derived distance and torsion potentials describing tertiary and local interactions along the chain, respectively (see Sect. 4). The predicted structure is that for which the free energy is lowest. MetaFoRe does not allow insertions and deletions in the sequence during threading. This has the advantage of avoiding several error-provoking approximations in the free-energy computations, but the disadvantage of limiting the number of structures tested; however, the latter point is of little importance here. Indeed, as we deal with small protein units, the number of threaded structures is large, i.e. of the order of 22000 on the dataset used, even without insertions and deletions.

Candidate foldons are required to have their local structure well predicted by Prelude and at least two secondary structure elements predicted to present some intrinsic, yet marginal, stability by Fugue. Moreover, they must pick up their native conformation as one of the lowest free-energy conformations out of all alternatives by metaFoRe, on the basis of the local-interaction potential as well as on the basis of the tertiary-interaction potential. The rationale behind the requirement of favorable local and tertiary interactions is that foldons have relatively small sizes, so possible compensations between the different energy terms are limited.

4 Ab initio folding simulations

3D structure predictions are performed by means of MC simulated annealing [18] in discretized (φ, ψ, ω) space [31]. This space is represented by 21 (φ, ψ, ω) triplets, three for each of the seven (φ, ψ, ω) domains A, C, B, P, G, E, and O (see legend to Fig. 1); the three points per domain are obtained by applying a treelike clustering algorithm to the (φ, ψ, ω) values observed in a set of well-resolved protein structures. This representation is sufficient to allow reasonably accurate protein reconstruction [36]. The starting structures are either randomly generated or correspond to the final structure of a previous simulation. The initial temperatures are taken to be 200 K and are gradually decreased by multiplication by 0.9.

At each step of the simulation, one-to-five consecutive residues are randomly selected and their (φ, ψ, ω) angles are modified. This modification is biased to account for the intrinsic preference of single amino acids for (φ, ψ, ω) domains: the average frequency by which an amino acid is assigned to each of the seven domains is equal to the corresponding frequency observed in the protein structure dataset. The protein is then reconstructed on the basis of the new (φ, ψ, ω) angles. If there are steric clashes, the new conformation is rejected. Otherwise, its folding free energy, ΔG, is evaluated using database-derived potentials described in the next paragraph. The new structure is accepted if its ΔG is lower than that of the last accepted structure or if the Metropolis criterion is satisfied at the temperature considered [12]. The new structure, if accepted, is modified in turn; otherwise it is rejected and the last accepted structure is reconsidered and remodified.

Two types of database-derived potentials are used, both derived from a dataset of 141 well-resolved protein structures displaying low sequence identity [37]. The torsion potential [32] describes local interactions along the sequence. It takes into account the propensities of residues to be associated with one of the seven (φ, ψ, ω) domains. It is computed from the frequencies with which an amino acid type a_{i}, at position i along the sequence or pairs of amino acids (a_{i}, a_j), at positions i and j, are associated with one of the seven (φ, ψ, ω) domains t_m at position m, with either $m-1 \le i,j \le m+1$ or $m-8 \le i,j \le m+8$; these two potentials are referred to as torsion[± 1] and torsion[± 8], respectively. The folding free energy $\Delta G^{\mathrm{S}}(\mathrm{C})$ of a sequence, S, in the conformation, C, computed from these frequencies reads as [38]

$$\Delta G^{\mathrm{S}}_{\mathrm{torsion}}(\mathrm{C}) = -kT \sum_{ij,m=1}^{N} \frac{1}{\zeta_m} \ln \frac{P(a_i, a_j, t_m)}{P(a_i, a_j)P(t_m)} ,$$

where N is the number of residues in the sequence S, k is the Boltzmann constant and T is a conformational temperature taken to be room temperature [39]. The normalization factor ζ_m ensures that the contribution of each residue in the windows $[m-1, m+1]$ or $[m-8, m+8]$ is counted once. It is equal to the window width, except near the chain ends.

The $\mathrm{C}^{\mu}-\mathrm{C}^{\mu}$ potential is a distance potential based on propensities of pairs of amino acids (a_i,a_j) at position i and j along the sequence to be separated by a spatial distance d_{ij} [35]. Consecutive residues along the sequence are not taken into account and spatial distances are calculated between average side chain centroids, C^{μ}. The frequencies of the residues separated by one-to-six positions along the sequence are computed separately, whereas the frequencies of the residues separated by seven residues and more are all merged. This yields a potential dominated by non-local, hydrophobic interactions, but which possesses a local component. The folding free energy defined by this potential is

$$\Delta G^{\mathrm{S}}_{\mathrm{C}^{\mu}-\mathrm{C}^{\mu}}(\mathrm{C}) = -kT \sum_{i+1<j}^{N} \ln \frac{P^{|i-j|}(a_i, a_j, d_{ij})}{P^{|i-j|}(a_i, a_j)P^{|i-j|}(d_{ij})},$$

with the probabilities $P^{|\mathrm{i}-\mathrm{j}|}$ independent of $|i-j|$ for $|i-j| > 7$. The discretization of the spatial distances is performed by dividing the distances between 3 and 8 Å into 25 bins of 0.2-Å width and merging the distances greater than 8 Å; the bins are smoothed as described in Ref. [35].

These two potentials are either used separately or in combination. In the latter case, they are simply summed.

5 Choice of the proteins to be predicted

We focus on DNA-binding domains of the HTH type, because most are small and well structured by themselves, even in the absence of DNA. Among them we chose the cro protein from phage 434 and the *Drosophila melanogaster* homeodomain, of the Protein Data Bank [40] codes 2CRO [41] and 1HDD [42], respectively. We shall refer to these proteins as CRO and HomeoD. These two proteins have no sequence identity and are part of two different HTH subfamilies. They contain 65 and 57 residues, respectively. However, the N-termimal arm and C-terminal residues of HomeoD only acquire their structure upon interaction with DNA [43] and will be overlooked here; only the 47 residues [C9, C55] will be considered. CRO has five helices and HomeoD only three, but in both cases the third helix is the recognition helix that enters into the major groove of DNA. In these two proteins, as in roughly half of the HTH proteins, the recognition helix and the preceding helix along the chain are separated by an αGBBα turn [44], where α stands for α helices and G and B refer to torsion angle domains (Table 1). The turn between the first two helices is also an αGBBα turn in CRO, but is characterized by a slightly different orientation between the helices, as in all proteins of the cro subfamily. In HomeoD, it is an αBABBBα turn, as in all homeodomains. Strikingly, this turn is observed in no other protein and appears fully specific to homeodomains [44].

6 Ab initio structure prediction of HomeoD

In a first stage, we applied Prelude and Fugue to the HomeoD sequence. The results are summarized in Fig. 1. The three helices H1, H2, and H3 are well predicted by both Prelude and Fugue, which indicates that they are likely to adopt a helical conformation in the absence of the rest of the chain. The position of the H2-H3 turn is also well predicted. In contrast, the H1-H2 turn is not predicted at all: Prelude and Fugue predict the whole H1-H2 stretch as a single helix.

Table 1. Native and predicted turns. The turns are specified by the (φ,ψ,ω) domain of the residues in the turn, flanked by αs to represent α helices [37, 44]. The (φ,ψ,ω) domains are those given in the legend to Fig. 1, where no distinction is made between the two helical domains A and C and between the two extended domains B and P. The predicted turns are the turns obtained in the ten HomeoD simulations of (type 2; type 3) described in Table 3 and in the ten CRO simulations of type 9 described in Table 4. The number of correctly predicted turns is the number of turns displaying the native motif, out of the ten predictions

	Turn location	Native turn	Correctly predicted turns
HomeoD	H1-H2	αBABBBα	(5;10)
	H2-H3	αGBBα	(2;2)
CRO	H1-H2	αGBBα	0
	H2-H3	αGBBα	10
	H3-H4	αGABABBα	10
	H4-H5	αGBBα	0

Additional information is obtained with the fold recognition program metaFoRe, whose results are given in Table 2. The H1-H2 and H2-H3 fragments are rather well predicted with the torsion[± 1] potential: the native structure appears as the second and the 14th match, respectively. However, neither fragment is well predicted by the torsion[± 8] potential. Using the C^{μ}-C^{μ} potential, the native H1-H2 structure appeared as a top match, whereas the native H2-H3 structure appeared only in the 301st position. These results mean that the first two helices H1-H2 possibly correspond to a foldon; however, as the metaFoRe predictions are not perfect with respect to the torsion potentials, it is also possible that the foldon coincides with the full H1-H2-H3 sequence, especially as this sequence contains only 47 residues. We shall consider the two possibilities.

According to the first possibility, we predicted the structure of H1-H2 by MC simulated annealing. The results are summarized in Table 3. The average rms deviations of the ten structures predicted with respect to the native conformation is 4.9 Å, and the minimum rms deviation is 3.6 Å. The ΔG value of the latter conformation is 0.8 kcal mol^{-1} higher than that of the predicted lowest energy conformation, which has an rms deviation of 5 Å.

The best predicted conformation for H1-H2 is taken as the starting conformation of ten new MC simulated annealing simulations including the whole HomeoD sequence. In these simulations, the predicted H1-H2 conformation is kept frozen except for the first and last three residues, i.e. the ΔG associated with H1-H2 is only allowed to increase by 2.0 kcal mol^{-1}. This leads to structures whose rms deviations with respect to the native structure are 5.9 Å on average; the lowest rms deviation is 4.9 Å.

We then considered the second possibility and performed ten MC simulated annealing simulations on the whole sequence starting from random conformations. The average rms deviation obtained with respect to the native conformation is 5.6 Å and is thus only slightly

Table 2. Predictions of metaFoRe [35]. For each target sequence and each potential, the position of the native structure among all tested alternatives is given; the number of alternatives is listed in the rightmost column. The limits of the CRO H1-H2 and H2-H3-H4 stretches are [−1,24] and [13,52], and those of the HomeoD H1-H2 and H2-H3 stretches are [C8,C40] and [C22,C56]

	Torsion [± 1]	Torsion [± 8]	C^{μ}-C^{μ}	Number of threadings
HomeoD H1-H2	2	390	1	22473
HomeoD H2-H3	14	307	301	22183
CRO H1-H2	964	385	451	23441
CRO H2-H3-H4	1	1	1	21292

lower than that obtained with the hierarchical procedure. Strikingly, the best predicted structure has an rms deviation as low as 2.7 Å. This structure, which is depicted in Fig. 2a, has a ΔG that is only 0.4 kcal mol^{-1} higher than the predicted conformation with lowest free energy out of the 20 simulations that use either the hierarchical or the nonhierarchical procedures.

Hence in this case the procedures for predicting first the H1-H2 stretch or directly the whole sequence are essentially equivalent, with a small advantage towards the second procedure. We can thus consider that the foldon encompasses here the whole sequence, which only comprises three helices.

As seen in Table 1, the H1-H2 and H2-H3 turns are rather well predicted in the simulations, both in the hierarchical and in the nonhierarchical approach. The first turn, the αBABBBα motif typical of all homeodomains, appears correctly in half of the nonhierarchical predictions and in all hierarchical ones. The second turn, the αGBBα HTH motif, is correct in two of the ten simulations, in the two types of procedures. The best predicted structure obviously has both turns correctly predicted. Note that these good turn predictions stem in part from the intrinsic conformational preferences of the residues in the turn and in part from the restraints due to the secondary structures at both sides of the turn.

7 Ab initio structure prediction of CRO

To determine which part of CRO could form a foldon, we first used Prelude and Fugue (Fig. 1). Prelude correctly predicts the helices H1, H2, H3, and H4, but neither helix H5 nor the intervening turns. More precisely, it predicts H5 as extended and H1 and H2 as forming a single long helix. In contrast, Fugue only predicts helix H4 and the stretch encompassing H1 and H2 as presenting some intrinsic stability. Hence, if we require a foldon to contain at least two marginally stable secondary structures, there are two candidates: H1-H2 and H2-H3-H4.

To decide between these candidates, we predicted their structure using the threading algorithm metaFoRe; the results are given in Table 2. The native structure of the H2-H3-H4 stretch is predicted as a top match by the two torsion potentials and by the C$^{\mu}$-C$^{\mu}$ potentials. This stretch is thus predicted to present some stability on the basis of both local and tertiary interactions. In contrast, the H1-H2 stretch is predicted as stable neither with the torsion nor with the C$^{\mu}$-C$^{\mu}$ potentials. This is in agreement with the fact that the H1-H2 turn is strongly predicted as helical by the torsion potentials of Fugue. In conclusion, the predicted CRO foldon is the H2-H3-H4 stretch.

In a second stage, the structure of the foldon is predicted by MC simulated annealing, with a combination of the torsion [± 8] and the C$^{\mu}$-C$^{\mu}$ potential, starting from random conformations. Ten simulations are performed, the results of which are summarized in Table 4. The average rms deviation of the predicted structures with respect to the native conformation is 5 Å, and the minimum and maximum rms deviations are 2.8 and 9.3 Å.

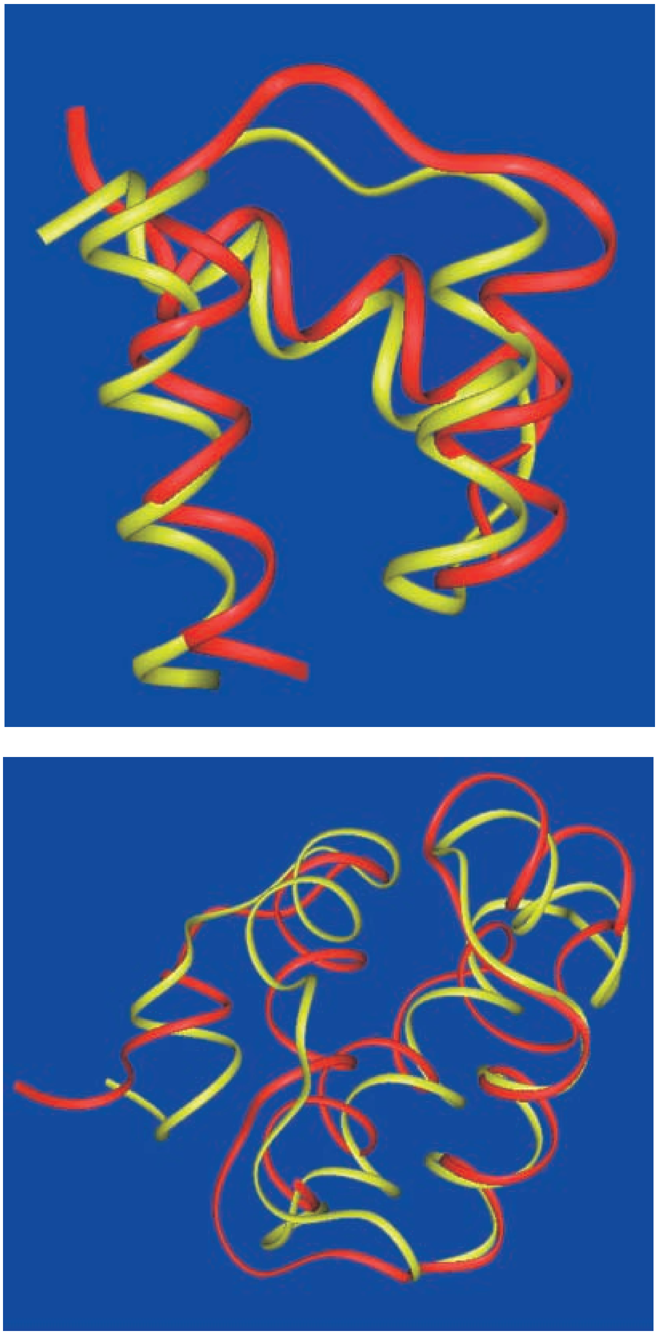

Fig. 2a, b. Representation of the predicted structures of lowest root-mean-square (*rms*) deviation with respect to the native conformation, superimposed on the native structure. The *yellow ribbon* corresponds to the native structure and the *red ribbon* to the predicted structure. The figures were generated by INSIGHT (MSI). **a** HomeoD [C9,C55]. The rms deviation between predicted and native conformations is 2.7 Å. The (φ, ψ, ω) domains of the predicted structure are (cf. native structure in Fig. 1): CAACAAAAAAAACABCBBBAAAAAAAAAACGPPAAAAAAAAAGGGAA. **b** CRO [−1,63]. The rms deviation between predicted and native conformations is 3.9 Å. The (φ, ψ, ω) domains of the predicted structure are (cf. native structure in Fig. 1): CAAAAAAAAAPAAAPGPPAAAAAAAACGPPACAAAAAACGAPABPAAAAAAAACCGBAPPAABBGB

The ΔG values associated with these predicted structures vary between −39 and −37 kcal mol^{-1}. The best

Table 3. Monte Carlo (*MC*) simulations on HomeoD. Three types of simulations, with ten simulations per type, are performed. The starting structures are given in column 4; the *numbers* refer to the type of simulations given in column 1. The potentials used is the sum of the torsion [±1] and C^{μ}-C^{μ} potentials. The average rms deviations and associated standard deviations for these ten simulations, as well as the minimum and maximum rms deviations, are given in the two rightmost columns. The rms deviations are computed on the C^{α} atoms after coordinate superpositions using U3BEST [46]. The limits of the H1-H2 and H1-H2-H3 stretches are [C9,C40] and [C9,C55]

	Predicted stretch	Semifixed stretch	Starting structure	< rms > ± standard deviation (Å)	rms_{min}–rms_{max} (Å)
1	H1-H2		Random	4.9 ± 0.5	3.6–5.4
2	H1-H2-H3		Random	5.6 ± 1.8	2.7–7.9
3	H1-H2-H3	H1-H2	1	5.9 ± 1.1	4.9–7.6

Table 4. MC simulations on CRO. Nine types of simulations, with ten simulations per type, are performed. The starting structures are given in column 4; the *numbers* refer to the type of simulations given in column 1. The combination of potentials used is the sum of the torsion [±8] and C^{μ}-C^{μ} potentials. The average rms deviations and associated standard deviation for these ten simulations, as well as the minimum and maximum rms deviations, are given in the two rightmost columns. The rms deviations are computed on the C^{α} atoms after coordinate superpositions using U3BEST [46]. The limits of the H2-H3-H4, H1-H2-H3-H4, H2-H3-H4-H4 and H1-H2-H3-H4-H5 stretches are [13,53], [−1,53], [13,63], and [−1,63]

	Predicted stretch	Fixed stretch	Starting structure	< rms > ± standard deviation (Å)	rms_{min} – rms_{max} (Å)
1	H2-H3-H4		Random	5.0 ± 2.0	2.8–9.3
2	H1-H2-H3-H4		Random	6.6 ± 2.5	3.8–11.7
3	H2-H3-H4-H5		Random	5.4 ± 1.0	4.7–8.2
4	H1-H2-H3-H4-H5		Random	8.4 ± 2.0	4.8–11.6
5	H1-H2-H3-H4	H2-H3-H4	1	4.8 ± 0.5	3.6–5.2
6	H2-H3-H4-H5	H2-H3-H4	1	4.3 ± 0.5	3.8–5.8
7	H1-H2-H3-H4-H5	H2-H3-H4	1	5.5 ± 0.7	4.0–7.0
8	H1-H2-H3-H4-H5	H2-H3-H4-H5	6	6.1 ± 0.1	6.0–6.3
9	H1-H2-H3-H4-H5	H1-H2-H3-H4	7	4.8 ± 0.3	3.9–5.3

predicted structure is not that with the lowest ΔG, so at this stage we cannot use an energetic criterion to discriminate between the different predicted conformations.

To predict the whole CRO sequence, we tested different procedures. The classical procedure consists of predicting the four or five CRO helices by MC simulated annealing starting from random conformations. The hierarchical procedure consists of adding to H2-H3-H4 the sequence stretch corresponding to helix H1 or helix H5 or both together and performing MC simulated annealing of these enlarged sequences while keeping the predicted structure of the H2-H3-H4 unmodified, except for the first three and last three residues of the H2-H3-H4 stretch.

The results of these two types of procedures are given in Table 4. We found that the rms deviations with respect to the native structure are systematically lower on average when the hierarchical procedure is used: 4.3 and 4.8 Å against 5.4 and 6.6 Å on four helices, and 4.8, 5.5, and 6.1 Å against 8.4 Å on five helices. Moreover the best results are obtained when adding first helix H1 and then helix H5, which is predicted as extended by the torsion potentials and not very well formed in the native structure. The average rms deviation obtained using this method is 4.8 Å. The predicted structure that most resembles the native structure has an rms deviation of 3.9 Å and is depicted in Fig. 2b. Its ΔG is −58.6 kcal mol^{-1}, whereas the predicted conformation of lowest energy, with ΔG of −58.9 kcal mol^{-1}, has an rms deviation of 5.0 Å.

It is noteworthy that several of the turns are well predicted by our procedure (Table 1). The H2-H3 and H3-H4 turns, which are αGBBα and αGABABBα motifs, are correctly predicted in all ten simulations where the helices H1 then H5 are successively added to the foldon. In contrast, the H1-H2 and H4-H5 turns are not well predicted in these ten simulations.

8 Discussion

In the two cases studied, the results of the ab initio simulations were positive. The best predicted structures have rms deviations relative to the native conformation of 2.7 and 3.9 Å, for the 47 residues of HomeoD and the 65 residues of CRO, respectively. The hierarchical procedure, consisting of first identifying and predicting the foldon and then the whole sequence, appeared to be definitely better in the case of CRO. In this case, the predicted foldon comprises the three helices H2-H3-H4, corresponding to the central motif involved in DNA binding. Strikingly, the two turns of this foldon, the αGBBα and αGABABBα turns, were predicted correctly not only in the best predicted conformation, but in all ten simulations. In the case of HomeoD, which is shorter than CRO and possesses only three helices, the folding did not appear hierarchical, and the foldon is the full sequence. This suggests that the minimum foldon size is three secondary structure elements and/or about 40–50 residues and that below this size folding occurs nonhierarchically. Here too, the two intervening turns were well predicted, especially the first, the typical homeodomain αBABBBα turn.

It must be noted that the MC simulated annealing simulations did not converge towards the minimum energy conformation, but yielded generally different conformations in the different runs. Though this can seem a shortcoming, it is actually not in the present context, where the potentials are not totally accurate. Indeed, the different runs yield conformations whose ΔGs are within about 1 kcal mol^{-1}, and the conformation presenting the best rms deviation with respect to the native structure is usually not that with the lowest ΔG. A conformational search algorithm that would pick up only the lowest energy conformation would thus, in general, not pick up the correctly predicted structures. In conclusion, the different simulations yield several structures, which cannot be discriminated on the basis of energetic criteria, but which constitute potentially correct predictions. These predicted structures can be taken as good candidates as starting structures of further analyses, such as ab initio simulations with more accurate representations and force fields.

Acknowledgements. D.G. benefits from a "FIRST-université" grant of the Walloon Region. M.R. is Senior Research Associate at the Belgian National Fund for Scientific Research.

References

1. Duan Y, Kollman PA (1998) Science 282: 740
2. Richmond TJ, Richards FM (1978) J Mol Biol 119: 537
3. Nagano K (1980) J Mol Biol 138: 797
4. (a) Cohen FE, Sternberg MJE, Taylor WR (1980) Nature: 285: 378; (b) Cohen FE, Sternberg MJE, Taylor WR (1981) J Mol Biol 148: 253; (c) Cohen FE, Sternberg MJE, Taylor WR (1982) J Mol Biol 156: 821
5. Eisenhaber F, Frömmel C, Argos P (1996) Proteins Struct Funct Genet 25: 169
6. Vásquez M, Scheraga HA (1985) Biopolymers 24: 1437
7. Simon I, Glasser L, Scheraga HA (1991) Proc Natl Acad Sci USA 88: 3661
8. Sippl MJ, Hendlich M, Lackner P (1992) Protein Sci 1: 625
9. (a) Simons KT, Kooperberg C, Huang E, Baker D (1997) J Mol Biol 268: 209; (b) Simons KT, Ruczinki I, Kooperberg C, Fox BA, Bystroff C, Baker D (1999) Proteins 34: 82; (c) Simons KT, Bonneau R, Ruczinki I, Baker D (1999) Proteins Suppl 3: 171
10. Srinivasan R, Rose GD (1995) Proteins 22: 81
11. Osguthorpe DJ (1999) Proteins Suppl 3: 186
12. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH (1953) J Chem Phys 21: 1087
13. Rey A, Skolnick J (1993) Proteins Struct Funct Genet 16: 8
14. Kolinski A, Skolnick J (1994) Proteins Struct Funct Genet 18: 338
15. Kolinski A, Skolnick J (1994) Proteins Struct Funct Genet 18: 353
16. Hansmann UHE, Okamoto Y, Onuchic JN (1999) Proteins Sruct Funct Genet 34: 472
17. Sung SS (1999) Biophys J 76: 164
18. Kirkpatrick S, Gelatt CD Jr, Vecchi MP (1983) Science 220: 671
19. Kawai H, Kikuchi T, Okamoto Y (1989) Protein Eng 3: 85
20. Wilson C, Doniach S (1989) Proteins Sruct Funct Genet 6: 193
21. Simons KT, Kooperberg C, Huang E, Baker D (1997) J Mol Biol 268: 209
22. Standley DM, Gunn JR, Friesner RA, McDermott AE (1998) Proteins Struct Funct Genet 33: 240
23. Goldberg DE (1989) Genetic algorithm in search, optmization and machine learning. Addison-Wesley, Reading, Massachussets
24. Pedersen JT, Moult J (1996) Curr Opin Struct Biol 6: 227
25. Skolnick J, Kolinski A (1990) Science 250: 1121
26. Jernigan RL (1992) Curr Opin Struct Biol 2: 248
27. (a) Ortiz AR, Kolinski A, Skolnick J (1998) Proc Natl Acad Sci USA 95: 1020; (b) Ortiz AR, Kolinski A, Skolnick J (1998) J Mol Biol 277: 419
28. Ortiz AR, Kolinski A, Rotkiewicz P, Ilkowski B, Skolnick J (1999) Proteins Suppl 3: 177
29. Orengo CA, Bray JE, Hubbard T, LoConte L, Sillitoe I (1999) Proteins Suppl 3: 149
30. Panchenko AR, Luthey-Schulten Z, Cole R, Wolynes PG (1997) J Mol Biol 272: 95
31. Gilis D, Rooman M (2001) Proteins Struct Funct Genet (in press)
32. (a) Rooman MJ, Kocher J-PA, Wodak SJ (1991) J Mol Biol 221: 961; (b) Rooman MJ, Kocher J-PA, Wodak SJ (1992) Biochemistry 31: 10226
33. Rooman MJ, Wodak SJ (1992) Biochemistry 31: 10239
34. Pintar A, Chollet A, Bradshaw C, Chaffotte A, Cadieux C, Rooman MJ, Hallenga K, Knowles J, Goldberg M, Wodak SJ (1994) Biochemistry 33: 11158
35. Kocher J-PA, Rooman MJ, Wodak SJ (1994) J Mol Biol 235: 1598
36. Park BH, Levitt M (1995) J Mol Biol 249: 493
37. Wintjens RT, Rooman MJ, Wodak SJ (1996) J Mol Biol 255: 235
38. (a) Rooman MJ, Wodak SJ (1995) Protein Eng 8: 849; (b) Rooman M, Gilis D (1998) Eur J Biochem 254: 135
39. Pohl FM (1971) Nature 234: 277
40. (a) Bernstein FC, Koetzle TF, Williams GJB, Meywe EFJr, Brice MD, Rodgers JR, Kennard O, Shimanoushi T, Tasumi M (1977) J Mol Biol 112: 535; (b) Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) Nucl Acids Res 28: 235
41. Mondragon A, Wolberger C, Harrison SC (1989) J Mol Biol 205: 179
42. Kissinger CR, Liu B, Martin-Blanco E, Kornberg TB, Pabo CO (1990) Cell 63: 579
43. (a) Tsao DH, Gruschus JM, Wang LH, Nirenberg M, Ferretti JA (1994) Biochemistry 33: 15053; (b) Cox M, van Tilborg PJ, de Laat W, Boelens R, van Leeuwen HC, van der Vliet PC, Kaptein RJ (1995) Biomol NMR 6: 23; (c) Carra JH, Privalov PL (1997) Biochemistry 36: 526
44. Wintjens R, Rooman M (1996) J Mol Biol 262: 294
45. Kabsch W, Sander C (1983) Biopolymers 22: 2577
46. Kabsch W (1978) Acta Crystallogr A 34: 827