All-atom folding studies of a DNA binding protein in a free-energy force field

Srinivasa M Gopal and Wolfgang Wenzel

Institute für Nanotechnologie, Forschungszentrum Karlsruhe, Germany

E-mail: wenzel@int.fzk.de

Abstract

We have recently extended our free-energy force field for the better treatment of beta-sheet structured proteins. The new force field, PFF02, nevertheless stabilizes helical proteins. Here we investigate the folding of the experimentally resolvable fragment of a DNA binding helical protein with a modified evolutionary algorithm. Our simulations converge to a helical ensemble. The energetically best conformation is within 4.4 Åof the experimental conformation, missing a single break in the second helix, which may result from flexible tails at the end of the molecule.

(Some figures in this article are in colour only in the electronic version)

1. Background

All-atom protein tertiary structure prediction and folding from the amino-acid sequence alone still remain challenging problems even for small proteins [1–3]. According to the thermodynamic paradigm [4] of protein folding, the native structure of the protein is the global optimum of a suitable free-energy force field. We have developed an all-atom free-energy force field [5], which predicts the native conformation of various helical proteins (1L2Y [6], 1F4I [7], 1VII [8], 1GYZ [9, 10]) at the global optimum of the force field. We have recently succeeded in generalizing this force field to PFF02 [11], which stabilizes protein of helical, beta-sheet and mixed folds [12]. We have explored several optimization methods to solve the associated optimization problem for a number of small proteins [13]. Among these, an evolutionary strategy [9, 10] hs proven particularly promising, because the optimization problem is solved by a large number of short independent simulations. Here we apply an improved version of this method to fold a 41-amino-acid segment of DNA-binding domain of MafG(PDBID:1K1V) [14]. Starting from a random conformation we find the global optimum structure, which has backbone root mean square deviation (bRMSD) of 4.4 Å.

2. Methods

2.1. The free-energy force field PFF02

We have recently developed all-atom (with the exception of apolar CH_n groups) free-energy protein force fields (PFF01/02) that model the low-energy conformation of proteins with minimal computational demand [5, 11]. The PFF02 force field, which parametrizes the internal free energy of the protein excluding backbone entropy, contains the following non-bonded interactions:

$$V(\{\vec{r}_i\}) = \sum_{ij} V_{ij}\left[\left(\frac{R_{ij}}{r_{ij}}\right)^{12} - 2\left(\frac{R_{ij}}{r_{ij}}\right)^{6}\right] + \sum_{ij}\frac{q_i q_j}{\epsilon_{g(i)g(j)} r_{ij}} + \sum_i \sigma_i A_i + \sum_{\text{hbonds}} V_{\text{hb}} + V_{\text{bb}} + V_{\text{tor}}\,. \tag{1}$$

Here r_{ij} denotes the distance between atoms i and j and $g(i)$ the type of the amino acid i. The Lennard-Jones parameters (V_{ij}, R_{ij}) for potential depths and equilibrium distance depend on the type of the atom pair and were adjusted to satisfy constraints derived from a set of 138 proteins of the PDB database [15–17]. The non-trivial electrostatic interactions in proteins are represented via group-specific and position dependent dielectric constants ($\epsilon_{g(i)g(j)}$), depending on the amino acids to which the atoms i and j belong. Interactions with the solvent were first fitted in a minimal solvent accessible surface model [18] parametrized by free energies per unit area σ_i to reproduce the enthalpies of solvation of the Gly–X–Gly family of peptides [19]. A_i corresponds to the area of atom i that is in contact with a fictitious solvent.

Hydrogen bonds are described via dipole–dipole interactions included in the electrostatic terms and an additional short-range term for backbone–backbone hydrogen bonding (CO to NH) which depends on the OH distance, the angle between N, H and O along the bond and the angle between the CO and NH axes [5]. In comparison to PFF01, the force field PFF02 contains an additional electrostatic term V_{bb} that differentiates between the backbone dipole alignments found in different secondary structure elements and a torsional potential for backbone dihedral angles V_{tor}, which gives a small contribution (about 0.3 kcal mol^{-1}) to stabilize conformations with dihedral angles in the beta-sheet region of the Ramachandran plot [11].

2.2. Evolutionary strategy

The popular basin hopping technique (BHT) for global optimization eliminates high-energy potential-energy surface (PES) [20, 21] by replacing the energy of each conformation with the energy of a nearby local minimum. For protein folding we have replaced the original local minimization by a simulated annealing (SA). In the course of our folding studies, we find that independent BHT simulations often find the identical structures corresponding to the same local (global) minimum. As a result, each independent simulation reconstructs the full folding path independently. It would be very desirable to develop methods where several concurrent simulations exchange information to *learn* from each other. For a PES having many local minima, independent simulations limit the efficient exploration of the PES. Also, occasionally BHT simulations go astray, ending the search in the wrong energy basin of the PES. We have examined a *greedy* version of BHT [22], which overcomes these problems to a certain extent.

We have therefore generalized the BHT approach to a population of size N which is iteratively improved by P concurrent dynamical processes [23]. The population is evolved towards a optimum of the free energy surface with a evolutionary strategy (ES) that balances the energy improvement with population diversity. In the ES, conformations are drawn from the *active* population and subjected to an annealing cycle. At the end of each cycle the resulting

conformation is either integrated into the active population or discarded. The algorithm was implemented as a master–client model in which idle clients request a task from the master. The master maintains the *active* conformation of the population and distributes the work to the clients. Each step in the algorithm has three phases.

(i) *Selection.* A conformation is drawn randomly from the *active* population. We have used a uniform probability distribution with population of 20 conformers.

(ii) *Annealing cycle.* We use a geometric cooling schedule with T_{start} drawn from a exponential distribution and T_{end} fixed at 2 K. The number of steps per cycle is increased as $10^5 \times \sqrt{\text{Cycle}}$.

(iii) *Population update.* We have adjusted the acceptance criterion for newly generated conformations to balance the population diversity and energy enrichment. We define the two structures as *similar* if they have bRMSD less than 3 Å to each other.
We define an *active* population as the pool of 20 lowest energy conformers. The master performs one of the following operations on the complete population.

  (a) *Add.* If the new conformation is not *similar* to any structure in the population, we add it to the population.
  (b) *Replace.* If the new conformation (with energy E_{new}) is *similar to one* existing structure in the population (with energy E_{old}), it replaces that structure provided $E_{\text{new}} < E_{\text{old}} + \Delta$ (see below).
  (c) *Merge.* If the new conformation has *several similar* structures, it replaces this group of structures provided its energy is less than the best one of the group E_{old} plus the acceptance threshold Δ.

In our first BHT simulations we have used a fixed energy threshold (Δ) acceptance criterion. Here we have implemented a *variable* energy threshold, which we define as $\Delta = A \times \tanh D$, where

$$D = \frac{E_{\text{new}} - E_{\text{best}}}{A},$$

where E_{best} is the lowest energy structure in the population. This choice of energy criterion ensures that the conformation with the best energy is never replaced, while conformations higher in energy are more easily replaced in the secure knowledge that they are far from optimal. The rules for the *replace* and *merge* operations ensure the structural diversity of the population and its continued energetic improvement (on average).

3. Results

The original population was seeded with a random conformation which had bRMSD of 14.0 Å to the NMR structure. Forty concurrent processes evolved an *active* population of 20 conformers. Each of the 40 processes ran 200 ES cycles, amounting to a total of 7.5×10^8 function evaluations. Figure 1 shows the average and lowest energy as a function of the ES cycle. It can be seen that there is no change in the energies after 150 cycles. Figure 2 shows the energies of all accepted conformations during the simulation. Four out of the ten best energetically lower conformations have bRMSD less than 5 Å. The secondary structure and energies of best conformations is shown in table 1. The best energy structure has two helices, which is shown in the left panel of figure 3. It differs from the NMR structure by 4.4 Å in bRMSD. There is a good agreement between the predicted and NMR structure as seen from the C_β–C_β distance matrix (figure 4). The second best structure is a three-helix bundle (right panel

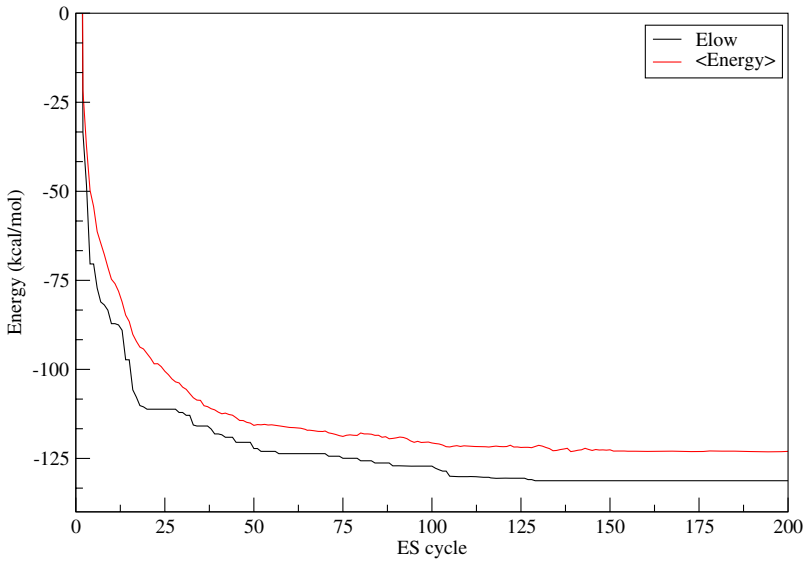

Figure 1. Average energy (red online) and the lowest energy (black) as the function of the ES cycle.

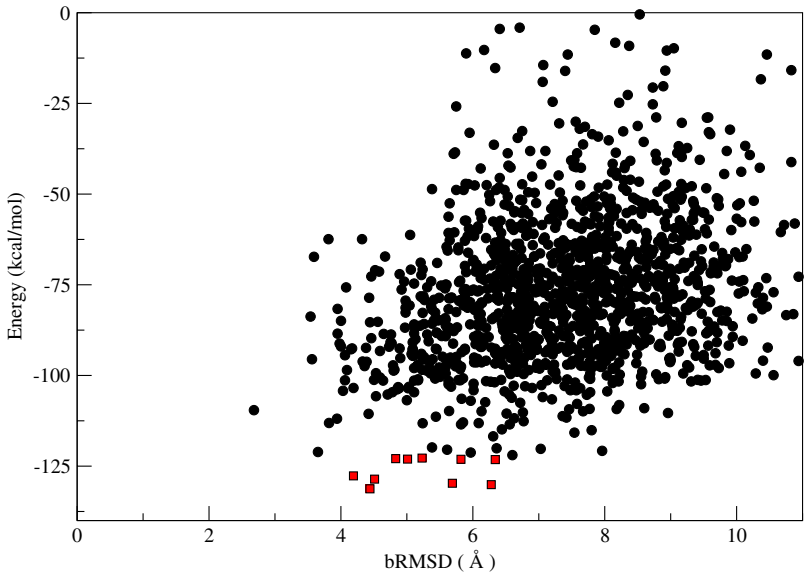

Figure 2. Energy versus bRMSD of all accepted conformations during the simulation. The squares (red online) indicate the ten lowest-energy conformations.

of figure 3), which has bRMSD of 6.28 Å. It has a proper alignment of first and third helices, but the middle helix is shifted with respect to the NMR structure.

The basic region (residues 28–41) along with the extended homology region (1–27) are highly conserved within the Maf family from the DNA binding motif specific to the Maf [14]. Though the basic region consists of 26 residues, the experimental structure is available for 14 residues (the last 14 residues of 1 K 1 V). Nine out of ten low-energy structures have very good

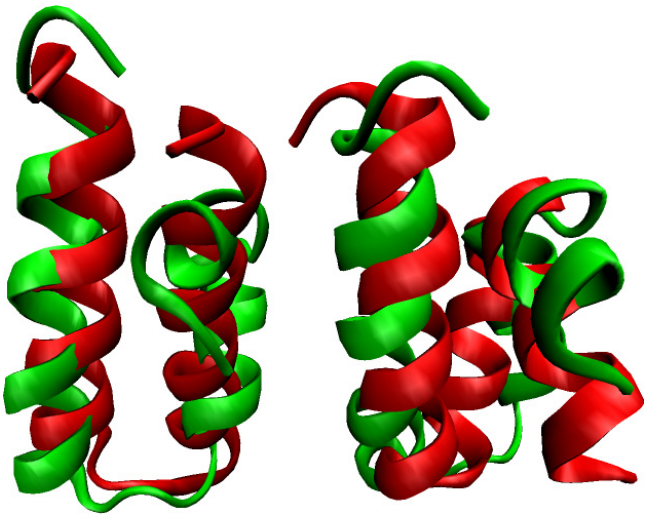

Figure 3. The left panel shows the overlay of best predicted (dark grey, red online) and NMR (light grey, green online) structures. The right panel shows the overlay of the second-best predicted three-helix structure (dark grey, red online) and NMR (light grey, green online). The images are made with VMD [24].

Table 1. Energy, RMSB and secondary structure assignment by DSSP (C = coil, E = sheet, T = turn, S = bend) for the lowest-energy structures from ten independent simulations.

Conf.	Energy (kcal mol^{-1})	bRMSD (Å)	bRMSD^{BR} (Å)	Secondary structure (DSSP)
NMR	—	—	—	CCHHHHHHSCHHHHHHHHTTSCHHHHHHHHHHHHHHTTSCC
1	−131.25	4.40	1.14	CCHHHHHHHHHHHHHHHHHHCCHHHHHHHHHHHHHHHHTC
2	−130.10	6.28	1.24	CCHHHHHHHHHHCHHHHHHHHHCHHHHHHHHHHHHHHTC
3	−129.72	5.69	1.18	CCCHHHHHHHHHHHHHSCCHHHHHHHHHHHHHHHHHHTC
4	−128.58	4.51	1.18	CCHHHHHHHHHSCSHHHHHHHHHSCHHHHHHHHHHHHTC
5	−127.68	4.19	1.12	CCHHHHHHHHHSCCHHHHHHHHHCHHHHHHHHHHHHTC
6	−123.19	6.34	1.37	CCHHHHHHHHHHHHHHHHHHHSCCHHHHHHHHHHHHCC
7	−123.11	5.82	3.28	CCHHHHHHHHHCHHHHHHHHCHHHHHHHHSCHHHHC
8	−123.06	5.01	1.40	CCHHHHHHHTTCHHHHHHHHCTHHHHHHHHHHHHCC
9	−121.93	4.83	1.79	CCHHHHHHHHHHHHHHHHSCSSHHHHHHHHHHHHHHC
10	−122.78	5.23	1.84	CCHHHHHHHHHHHHHSCCTHHHHHHHHSHHHHHHHTC

alignment with the experimentally available basic region (see table 1). The extended homology region is reproduced to 3.04 Å in the best-energy model. We note that here we simulated only the structurally resolved fragment of the entire protein. The existence of tails at both ends of the protein in the experiment could induce the break between the second and third helix that is missing in our lowest-energy structure, because the energetic difference between ensembles with and without such a break is less than 1.5 kcal mol^{-1} in our force field.

4. Conclusions

We have demonstrated that the improved force field PFF02 correctly identifies the tertiary structure for a nontrivial 41-amino-acid helical protein. The global optimum structure for a

Figure 4. C_β–C_β distance matrix between the predicted and the NMR conformations. Rows and columns of this colour-coded distance map indicate the difference in the C_β–C_β distances of the native and the predicted structure. Black/grey squares indicate that the C_β–C_β distances of the native and the predicted structure differ by less than 1.5/2.25 Å, respectively. White squares indicate larger deviations.

DNA binding protein is predicted as a two-helix bundle, which differs from the NMR structure by 4.4 Å in bRMSD. This result further demonstrates that the evolutionary strategy is a robust optimization method to fold small, but nontrivial, proteins in the free-energy force field PFF02.

Acknowledgments

We thank the Deutsche Forschungsgemeinschaft (grants WE 1863/10-2, WE 1863/14-1) and the Kurt Eberhard Bode Stiftung for financial support. We are thankful to the KIST supercomputational materials laboratory for computational facilities. SMG thanks Professor Marek Cieplak for workshop related re-imbursement and the organizing committee for an exciting workshop.

References

[1] Baker D and Sali A 2001 Protein structure prediction and structural genomics *Science* **294** 93–6
[2] Pillardy J, Czaplewski C, Liwo A, Lee J, Ripoll D R, Kamierkiewicz R, Oldziej S, Wedemeyer W J, Gibson K D, Arnautova Y A, Saunders J, Ye Y-J and Scheraga H A 2001 Recent improvements in prediction of protein structure by global optimization of a potential energy function *Proc. Natl Acad. Sci. USA* **98** 2329–33
[3] Schonbrunn J, Wedemeyer W J and Baker D 2002 Protein structure prediction in 2002 *Curr. Opin. Struct. Biol.* **12** 348–52
[4] Anfinsen C B 1973 Principles that govern the folding of protein chains *Science* **181** 223–30
[5] Herges T and Wenzel W 2004 An all-atom force field for tertiary structure prediction of helical proteins *Biophys. J.* **87** 3100–9
[6] Schug A, Herges T and Wenzel W 2003 Reproducible protein folding with the stochastisc tunneling method *Phys. Rev. Lett.* **91** 158102

[7] Schug A, Herges T and Wenzel W 2004 All-atom folding of the three-helix HIV accessory protein with an adaptive parallel tempering method *Proteins* **57** 792–8
[8] Herges T, Schug A and Wenzel W 2004 Protein structure prediction with stochastic optimization methods: Folding and misfolding the villin headpiece *Lect. Notes Comput. Sci.* **3045** 454–64
[9] Schug A and Wenzel W 2004 Predictive *in-silico* all-atom folding of a four helix protein with a free-energy model *J. Am. Chem. Soc.* **126** 16736–7
[10] Schug A and Wenzel W 2006 Evolutionary strategies for all-atom folding of the sixty amino acid bacterial ribosomal proein l20 *Biophys. J.* **90** 4273–80
[11] Verma A and Wenzel W 2006 Stabilization and folding of beta-sheet and alpha-helical proteins in an all-atom free energy model, in preparation
[12] Gopal S M and Wenzel W 2006 *De-novo* folding of the dna-binding atf-2 zinc finger motif in an all-atom free energy forcefield *Angew. Chem. Int. Edn* **45** 7726–8
[13] Schug A, Verma A, Wenzel W and Schoen G 2005 Biomolecular structure prediction with stochastic optimization methods *Adv. Eng. Mater.* **7** 1005–9
[14] Katsuoka F, Morohashi A, Yamamoto M, Kusunoki H, Motohashi H and Tanaka T 2002 Solution structure of the dna-binding domain of Mafg *Nat. Struct. Mol. Biol.* **9** 252–6
[15] Abagyan R A and Totrov M 1994 Biased probability Monte Carlo conformation searches and electrostatic calculations for peptides and proteins *J. Mol. Biol.* **235** 983–1002
[16] Herges T, Merlitz H and Wenzel W 2002 Stochastic optimization methods for biomolecular structure prediction *J. Ass. Lab. Autom.* **7** 98–104
[17] Herges T, Schug A and Wenzel W 2004 Exploration of the free energy surface of a three helix peptide with stochastic optimization methods *Int. J. Quantum Chem.* **99** 854–93
[18] Eisenberg D and McLachlan A D 1986 Solvation energy in protein folding and binding *Nature* **319** 199–203
[19] Sharp K A, Nicholls A, Friedman R and Honig B 1991 Extracting hydrophobic free energies from experimental data: relationship to protein folding and theoretical models *Biochemistry* **30** 9686–97
[20] Nayeem A, Vila J and Scheraga H A 1991 A comparative study of the simulated-annealing and Monte Carlo-with-minimization approaches to the minimum-energy structures of polypeptides: [met]-enkephalin *J. Comput. Chem.* **12** 594–605
[21] Leitner D M, Chakravarty C, Hinde R J and Wales D J 1997 Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms *Phys. Rev.* E **56** 363
[22] Wenzel W 2006 *De novo* folding of two-helix potassium channel blockers, submitted
[23] Schug A, Wenzel W and Hansmann U E H 2005 Energy landscape paving simulations of the trp-cage protein *J. Chem. Phys.* **122** 194711
[24] Humphrey W, Dalke A and Schulten K 1996 VMD—visual molecular dynamics *J. Mol. Graph.* **14** 33–8