Predictive and reproducible *de novo* all-atom folding of a β-hairpin loop in an improved free-energy forcefield

2007 J. Phys.: Condens. Matter 19 285213

(http://iopscience.iop.org/0953-8984/19/28/285213)

IOP PUBLISHING JOURNAL OF PHYSICS: CONDENSED MATTER

J. Phys.: Condens. Matter **19** (2007) 285213 (9pp) doi:10.1088/0953-8984/19/28/285213

Predictive and reproducible *de novo* all-atom folding of a β-hairpin loop in an improved free-energy forcefield

Abhinav Verma[1] **and Wolfgang Wenzel**[2]

[1] Institute of Scientific Computing, Forschungszentrum Karlsruhe, Germany
[2] Institute of Nanotechnology, Forschungszentrum Karlsruhe, Germany

E-mail: wenzel@int.fzk.de

Received 20 September 2006, in final form 6 October 2006
Published 25 June 2007
Online at stacks.iop.org/JPhysCM/19/285213

Abstract

We have recently improved our free-energy forcefield for all-atom *de novo* protein folding to address the folding of proteins with beta-sheet secondary structure. The folding of beta strands is nevertheless more difficult than that of helical proteins, because of non-local interactions between regions of the protein chain that are not consecutive in the amino acid sequence. Here we use a greedy version of the basin-hopping technique with our free-energy forcefield PFF02 to reproducibly and predictively fold the structure of a β-hairpin loop. The lowest energy structure found in the simulation has a backbone root mean square deviation of only 2.62 Å to the native conformation. The side-chain alignment is also correctly predicted, as are four of the five backbone hydrogen bonds found in the native structure.

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

1. Background

De novo protein folding and structure prediction from the sequence information alone still remain challenging tasks even for small proteins [1–5]. Atomistic simulations of the folding process remain confined to small peptides, due to their large computational cost [6, 7]. With the development of reliable forcefields and robust simulation techniques, protein folding studies may assist in the understanding of the folding process and protein biological function at an atomistic level. Based on the thermodynamic paradigm [8] of protein folding, free-energy-based methods describe the native structure of the protein as the global optimum of a suitable free-energy forcefield. This approach to protein structure determination is potentially much faster and more predictive than the costly simulation of the folding pathway, but will obviously sacrifice dynamical information.

We have earlier reported the rational development of a transferable free-energy forcefield PFF01 [9] that correctly predicts the native conformation of various helical proteins (1L2Y [4],

1F4I [10], 1VII [11], 1GYZ [12, 13]) at the global minimum. We have recently modified PFF01 to a more generalized free-energy forcefield PFF02 [14], which predicts tertiary structures of proteins with more diverse secondary structural elements, especially beta-sheet structures.

In our experience the folding of alpha-helical conformations requires much less computational resources than that of beta-sheet secondary structure elements. One reason that beta-strand structure prediction is more difficult to fold stems from the high prevalence of non-local interactions between regions of the protein chain that are not-consecutive in the amino acid sequence. These long-range interactions make it difficult to nucleate an appropriate starting point in the protein conformation, from which the whole conformation can grow. While helices can nucleate, at least in principle, anywhere in the helix sequence, beta-sheets require the nucleation of a partially formed structure with the correct hydrogen bonding pattern. Because many more hydrogen bond pairings are possible, the nucleation of such a correct nucleus is much less likely. The optimal choice of the backbone hydrogen bond topology provides an important enthalpic contribution for the stabilization of the native conformation. The search for the global optimum of the free-energy surface of proteins containing beta-sheet secondary structure is therefore much more demanding.

Here we report the predictive and reproducible folding of a designed β-hairpin loop (PDB Code: 2EVQ) in the improved free-energy forcefield PFF02. We used a greedy version of the basin-hopping technique [15], which biases the search process toward lower-energy structures. We find that eight out of ten simulations converge to less than 3 Å backbone root mean square deviations (bRMSDs) from the native conformation; the lowest-energy conformation located during the search process had a bRMSD of 2.62 Å compared to the native conformation. Eight out of ten simulations find hairpin-like conformations (with bRMSDs between 2.4 and 2.8 Å).

2. Methods

2.1. The free-energy forcefield PFF02

We have recently developed all-atom (with the exception of apolar CH_n groups) free-energy protein forcefields (PFF01/02) that model the low-energy conformations of proteins with minimal computational demand [9, 14]. The forcefield, which parameterizes the internal free energy of the protein excluding backbone entropy, contains the following non-bonded interactions:

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

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 [16–18]. 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 [19] parameterized by free energies per unit area σ_i to reproduce the enthalpies of solvation of the Gly–X–Gly family of peptides [20]. 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 axis [9]. In comparison to PFF01, the forcefield PFF02 contains an additional electrostatic term V_{bb} that differentiates between the backbone dipole alignments found in different secondary structure elements (included in the electrostatic potential between atoms i and j via the dielectric constants $\epsilon_{g(i)g(j)}$) [21] 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 [14, 22]. The additional electrostatic term acts as an effective dielectric constant for short-range interactions between the dipole moments of neighbouring amino acids, maximally two residues apart in the sequence [21]. Because the alignment of these dipoles differs in helices and beta-sheets, it effectively discriminates between these two secondary structure elements.

2.2. *Greedy basin hopping*

The basin-hopping technique [23] and the related Monte Carlo-with-minimization method [24] employ a relatively straightforward approach to eliminate high-energy transition states of the potential-energy surface (PES): the original potential-energy surface is simplified by replacing the energy of each conformation with the energy of a nearby local minimum. This replacement eliminates high-energy barriers in the stochastic search that are responsible for the freezing problem in simulated annealing. The basin-hopping technique and derivatives has been used previously to study the potential-energy surface of model proteins and polyalanines using all-atom models [25–27].

We replace the gradient-based minimization step with a simulated annealing (SA) run [28, 15], because local minimization generates only very small steps on the free-energy surface. As in basin hopping, the simulation is always started from the last local minimum found. Within each SA simulation, new configurations are accepted according to the Metropolis criterion, while the temperature is decreased geometrically from its starting value to the final value. The starting temperature and cycle length determine how far the annealing step can deviate from its starting conformation. The final temperature must be small compared to typical energy differences between competing metastable conformations, to ensure convergence to a local minimum. The annealing protocol is thus parameterized by the starting temperature T_{S}, the final temperature T_{F}, and the number of steps.

At the end of one annealing cycle the new conformation is accepted if its energy difference from the current configuration is no higher than a given threshold energy ϵ_{T}, an approach recently proven optimal for certain optimization problems. The greedy version of basin hopping has a varying threshold energy depending upon the best energy found so far in the simulation. Here we calculated the threshold criteria as $(\epsilon_{\mathrm{S}} - \epsilon_{\mathrm{B}})/4$, where ϵ_{S} is the starting energy and ϵ_{B} is the best energy found so far in the simulation. This choice implies that the conformation with the best energy is never replaced with a conformation that is higher in energy, and effectively introduces a 'memory effect' in the simulation. For simulations that are higher in energy, the increased threshold value implies a higher acceptance probability of conformations with high energy. As a result these simulations explore the conformational space faster, in the secure knowledge that they are far from optimal. Overall this leads to an improvement of the convergence of the low-energy simulations, but a loss of convergence for the high-energy conformations.

The starting temperatures were randomly chosen from an exponential distribution which helps to speed up the search process [15]. The number of steps was increased in every cycle by a factor of $\sqrt{n}$, which also helps in accurate location of the minima on the free-energy surface. The final temperatures for each simulation annealing cycle of basin hopping was set at 3 K.

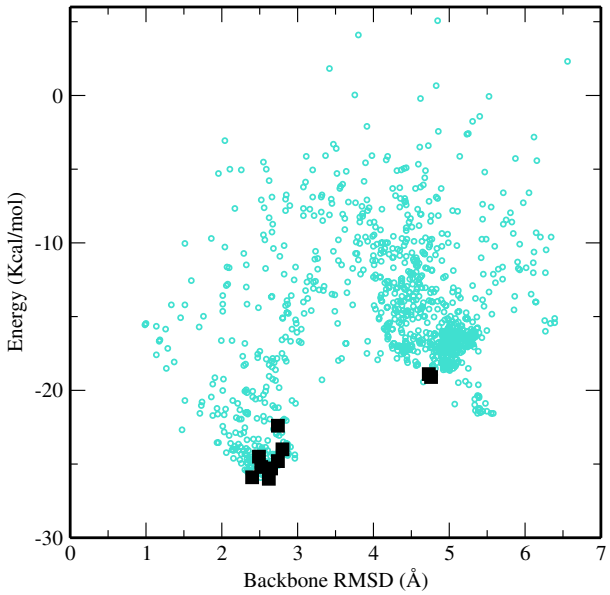

Figure 1. Energy versus bRMSD of all conformations visited during the simulation. The (red) squares indicate the lowest-energy conformation found in each of the ten independent simulations.

3. Results

We started the simulations of a designed β-hairpin loop (pdb-id: 2EVQ, sequence: KTWNPATGKWTE) [30] peptide starting from a completely extended conformation. The starting conformation for the simulation had a bRMSD of 10.5 Å from the native conformation and large positive energy in PFF02. Following the standard definitions for hydrogen bonding in MOLMOL [31] (with maximum distance = 2.4 Å and maximum angle = 35°), 2EVQ folds into a hairpin conformation with backbone hydrogen bonds between residues 2 and 11 and 4 and 9, respectively.

We performed ten independent simulations for 100 cycles, each of which comprised about 7 million energy evaluations, using the greedy basin-hopping algorithm described in section 2. Figure 1 shows the distribution of conformations visited during the search process by energy and bRMSD. The plot of energy as a function of the steps is shown in figure 2. The bottom panel of the figure also shows the first passage times for reaching the native ensemble for the nine converging simulations. The average first passage time was 2.2×10^6 function evaluations, which provides and unbiased comparison with other methods [32].

The minimal energy and bRMSD from the original NMR structure are shown in table 1. Eight out of the ten simulations converged to less than 3 Å bRMSD from the native conformation. The best structure deviated by only 2.62 Å bRMSD. An analysis of the secondary structure content revealed a close similarity of all low-energy structures. All converged structures predict the two strands of the hairpin correctly.

The low-energy part of the free-energy landscape is illustrated in figure 3, which demonstrates the existence of a nearly perfect folding funnel, in agreement with the current paradigm of protein structure formation [33, 34]. In the absence of backbone entropy the low-energy part of the folding funnel appears to be somewhat rugged, and has only a small overall slope towards the native conformation.

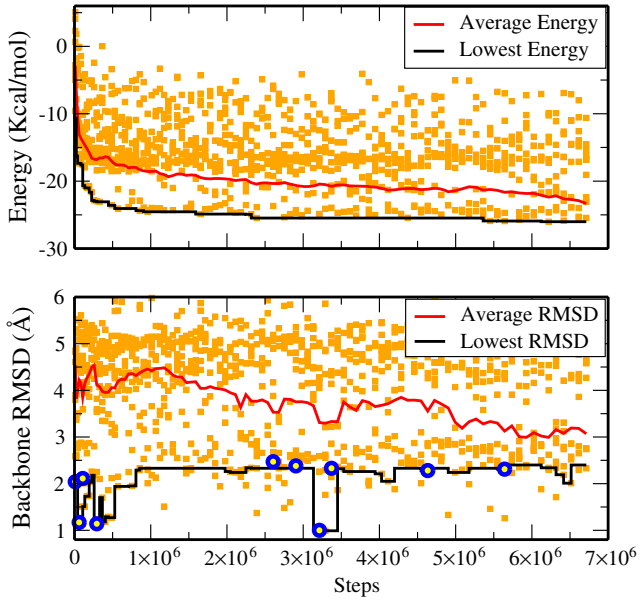

Figure 2. Average energy (top)/bRMSD (bottom) and the lowest energy (top)/bRMSD (bottom) as the function of the number of energy evaluations during the simulations described in the text. The data points are shown as orange squares; the blue circles indicate the first passage times for folding, defined as reaching a near-native conformation within 2.5 Å of the native conformation for simulations that reached the native ensemble.

Table 1. Energy, bRMSD and secondary structure assignment by DSSP [29] (C = coil, E = sheet, T = turn, S = bend) for the lowest energy structures from ten independent simulations.

Name	Energy (kcal mol^{-1})	bRMSD (Å)	Secondary structure (DSSP)
2EVQ	—	—	CEEETTTTEEEC
2EVQ.5	−26.0	2.62	CEEETTTTEEEC
2EVQ.6	−25.9	2.40	CEEETTTTEEEC
2EVQ.2	−25.3	2.65	CEEETTTTEEEC
2EVQ.0	−25.2	2.52	CEEETTTTEEEC
2EVQ.9	−24.8	2.74	CEEETTTTEEEC
2EVQ.4	−24.5	2.49	CEEETTTTEEEC
2EVQ.3	−24.0	2.80	CEEETTTTEEEC
2EVQ.8	−22.4	2.74	CEEETTTTEEEC
2EVQ.7	−19.1	4.76	CCHHHHHTSSCC
2EVQ.1	−18.9	4.73	CCHHHHHTSSCC

The good agreement between the folded and the native conformation in illustrated in figure 4 for both the backbone and important sidechains. To quantify the agreement we also calculated the C_β–C_β distance difference matrix for the best conformation to the NMR conformation (see figure 5). The C_β–C_β distance difference matrix shows the difference between pairs of C_β–C_β distances between two structures. A completely black matrix will mean a complete agreement. The matrix clearly shows the agreement between C_β–C_β distances of the two structures, implying the correct side-chain alignments in the predicted structure.

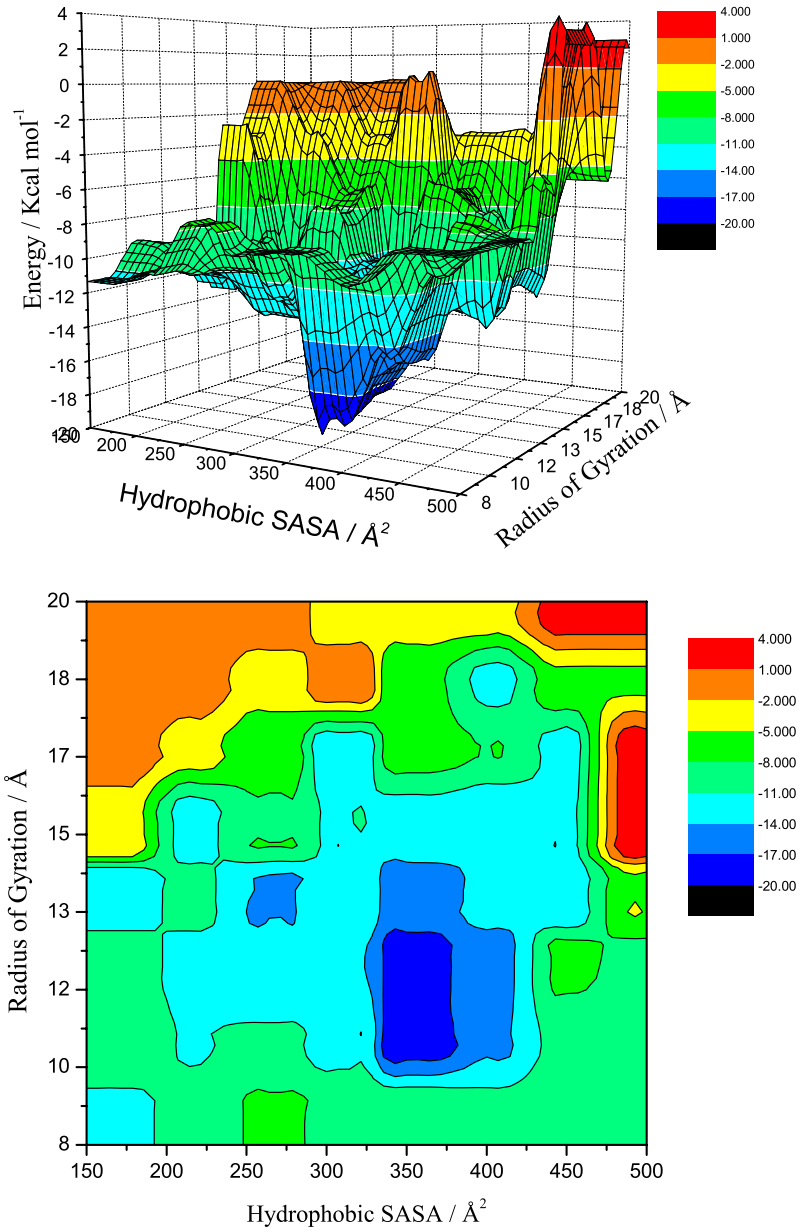

Figure 3. 3D (top) and contour view of the low-energy part of the free-energy landscape assembled from the analysis of all conformations visited in the course of the folding simulations.

Hydrogen bonding is responsible for the formation of secondary structure in proteins and its energetic contribution plays a crucial role in stabilizing protein structures. Especially in beta-sheets, where the hydrogen bond interactions for amino acids is not consecutive in the sequence, a slight change in hydrogen bonding patterns can result in very different structures. Using the same criteria as above we find that four out of five hydrogen bonds present in the native state are found in the lowest-energy conformation (see table 2, calculated using Molmol [31]).

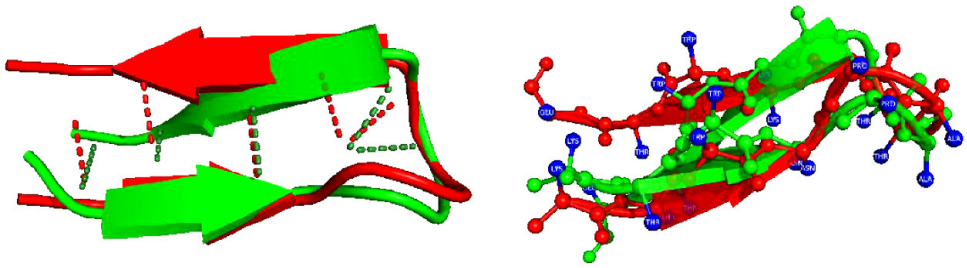

Figure 4. Overlay of folded structure (red) of the hairpin loop on the experimental structure (green). The left panel shows the overlap of the backbone only with the hydrogen bonds, while the right panel shown in addition the overlay of the sidechains (blue spheres).

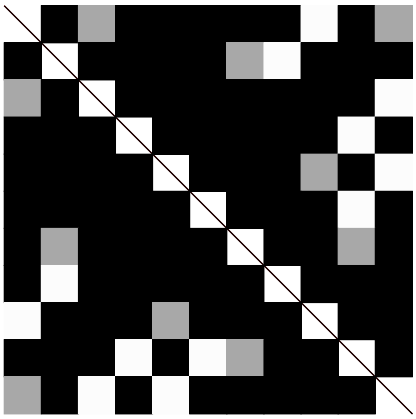

Figure 5. The C_β–C_β pixel in row i and column j of the colour-coded distance map indicates the difference in the C_β–C_β distances of the native and the folded structure. Black/grey squares indicate that the C_β–C_β distances of the native and the other structure differ by less than 1.5/2.25 Å respectively. White squares indicate larger deviations.

Table 2. Hydrogen bond topology for the native conformation and the lowest-energy conformation.

H bond **NH**	Topology **O**	2EVQ	Best (2EVQ.5)
2THR	11THR	Y	Y
4ASN	9LYS$^+$	Y	Y
7THR	4ASN	Y	
8GLY	4ASN	Y	Y
9LYS$^+$	4ASN		Y
11THR	2THR	Y	Y

4. Conclusions

We have demonstrated that the modified forcefield PFF02 correctly predicts the native conformation of the designed β-hairpin loop (PDB:2EVQ) as the global minimum of the its free-energy surface. The best conformation found in the simulation deviates by only 2.62 Å from the NMR conformation. It has the correct hydrogen bonding pattern and side-chain alignment as found in the NMR conformation. Eight out of ten simulations converge to less

than 3 Å bRMSD from the native structure. We find a single folding funnel in the low-energy part of the free-energy surface, which appears very rugged in the absence of backbone entropy. The greedy version of the basin-hopping technique, which was used in this study, is able to efficiently locate the global minimum of this complex free-energy surface to correctly predict the folded state.

Acknowledgments

We thank the Deutsche Forschungsgemeinschaft (grants WE 1863/10-2, WE 1863/14-1) and the Kurt Eberhard Bode Stiftung for financial support. Part of the simulations were performed at the KIST teraflop cluster.

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] Schug A, Herges T and Wenzel W 2003 Reproducible protein folding with the stochastic tunneling method *Phys. Rev. Lett.* **91** 158102
[5] Herges T and Wenzel W 2005 Reproducible *in-silico* folding of a three-helix protein and characterization of its free energy landscape in a transferable all-atom forcefield *Phys. Rev. Lett.* **94** 018101
[6] Pande V S and Rokhsar D S 1999 Molecular dynamics simulations of unfolding and refolding of a beta-hairpin fragment of protein g *Proc. Natl Acad. Sci. USA* **96** 9062–7
[7] Snow C D, Nguyen H, Pande V S and Gruebele M 2002 Absolute comparison of simulated and experimental protein folding dynamics *Nature* **420** 102–6
[8] Anfinsen C B 1973 Principles that govern the folding of protein chains *Science* **181** 223–30
[9] Herges T and Wenzel W 2004 An all-atom force field for tertiary structure prediction of helical proteins *Biophys. J.* **87** 3100–9
[10] 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
[11] Herges T, Schug A and Wenzel W 2004 Protein structure prediction with stochastic optimization methods: folding and misfolding the villin headpiece *Lecture Notes in Computer Science* vol 3045 (Berlin: Springer) pp 454–64
[12] 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
[13] Schug A and Wenzel W 2006 Evolutionary strategies for all-atom folding of the sixty amino acid bacterial ribosomal protein l20 *Biophys. J.* **90** 4273–80
[14] Verma A and Wenzel W 2005 Stabilization and folding of beta-sheet and alpha-helical proteins in an all-atom free energy model, in preparation
[15] Verma A, Schug A, Lee K H and Wenzel W 2006 Basin hopping simulations for all-atom protein folding *J. Chem. Phys.* **124** 044515
[16] 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
[17] Herges T, Merlitz H and Wenzel W 2002 Stochastic optimization methods for biomolecular structure prediction *J. Ass. Lab. Autom.* **7** 98–104
[18] 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
[19] Eisenberg D and McLachlan A D 1986 Solvation energy in protein folding and binding *Nature* **319** 199–203
[20] 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
[21] Avbelj F and Moult J 1995 Role of electrostatic screening in determining protein main chain conformational preferences *Biochemistry* **34** 755–64

[22] Wenzel W 2006 Predictive folding of a β hairpin in an all-atom free-energy model *Europhys. Lett.* **76** 156–62
[23] 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
[24] 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
[25] Mortenson P N and Wales D J 2004 Energy landscapes, global optimization and dynamics of poly-alanine $Ac(ala)_8nhme$ *J. Chem. Phys.* **114** 6443–54
[26] Mortenson P N, Evans D A and Wales D J 2002 Energy landscapes of model polyalanies *J. Chem. Phys.* **117** 1363–76
[27] Wales D J and Dewbury P E J 2004 Effect of salt bridges on the energy landscape of a model protein *J. Chem. Phys.* **121** 10284–90
[28] Schug A, Verma A, Herges T, Lee K H and Wenzel W 2005 Comparison of stochastic optimization methods for all-atom folding of the trp-cage protein *ChemPhysChem* **6** 2640–6
[29] Kabsch W and Sander C 1983 Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features *Biopolymers* **22** 2577–637
[30] Andersen N H, Olsen K A, Fesinmeyer R M, Tan X, Hudson F M, Eidenschink L A and Farazi S R 2006 Minimization and optimization of designed β-hairpin folds *J. Am. Chem. Soc.* **128** 6101–10
[31] Koradi R, Billeter M and Wuethrich K 1996 Molmol: a program for display and analysis of macromolecular structures *J. Mol. Graph.* **14** 51–5
[32] Ripoll D R, Liwo A and Scheraga H A 1998 New developments of the electrostatically driven Monte Carlo method: test on the membrane-bound portion of melittin *Biopolymers* **46** 117–26
[33] Onuchic J N, Luthey-Schulten Z and Wolynes P G 1997 Theory of protein folding: the energy landscape perspective *Annu. Rev. Phys. Chem.* **48** 545–600
[34] Dill K A and Chan H S 1997 From levinthal to pathways to funnels: the 'new view' of protein folding kinetics *Nat. Struct. Biol.* **4** 10–9