Protein simulations combining an all-atom force field with a Go term

2007 J. Phys.: Condens. Matter 19 285215

(http://iopscience.iop.org/0953-8984/19/28/285215)

IOP PUBLISHING JOURNAL OF PHYSICS: CONDENSED MATTER

J. Phys.: Condens. Matter **19** (2007) 285215 (9pp) doi:10.1088/0953-8984/19/28/285215

Protein simulations combining an all-atom force field with a Go term

Jan H Meinke[1] **and Ulrich H E Hansmann**[1,2]

[1] John-von-Neumann Institute for Computing, Forschungszentrum Jülich, D-52425 Jülich, Germany
[2] Department of Physics, Michigan Technological University, Houghton, MI 49931, USA

E-mail: j.meinke@fz-juelich.de and u.hansmann@fz-juelich.de

Received 14 October 2006, in final form 1 March 2007
Published 25 June 2007
Online at stacks.iop.org/JPhysCM/19/285215

Abstract

Using a variant of parallel tempering, we study the changes in sampling within a simulation, when the all-atom model is coupled to a Go-like potential. We find that the native structure is not the lowest-energy configuration in the all-atom force field. Adding a Go term deforms the energy landscape in such a way that the native configuration becomes the global minimum but does not lead to faster thermalization.

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

1. Introduction

Most proteins exist at room temperature in a *unique* structure that one can identify with the lowest *potential* energy conformation [1]. It is now commonly assumed that the energy landscape of a protein is shaped like a funnel with the native state at the bottom [2]. The landscape has, however, many deep local minima and high barriers, because the average protein contains thousands of atoms, and interactions between the atoms can be both repulsive and attractive.

Due to the large number of continuous degrees of freedom and the rough energy landscape, simulating proteins remains a computational challenge. The time to find the native structure of a protein (the bottom of the funnel) depends both on the roughness of the energy landscape and the steepness of the funnel. The more pronounced the funnel is, the more quickly the protein will fold. This is one reason for the popularity of the Go model [3–5]. Its basic assumption is that only interactions present in the native state of a protein are relevant for the folding process. An appropriate energy function then ignores non-native interactions and rewards native interactions. Hence, the Go model represents a perfect funnel model and has none of the roughness normally associated with the protein-folding energy landscape. In their 1981 paper, Abe and Go [3] used a lattice model, where each amino acid occupied a single

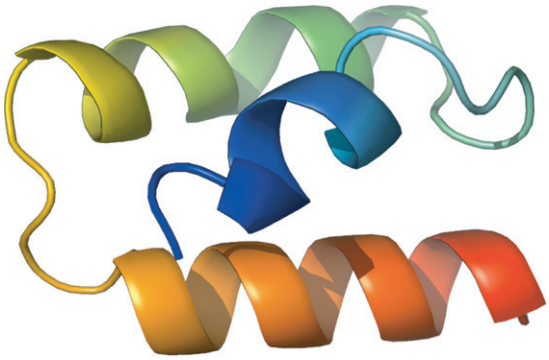

Figure 1. Native structure of the 46 amino-acid long segment of protein A (1bdd) used as a native reference structure. The structure taken from the Protein Data Bank has been adjusted to fit the standard geometry assumed by ECEPP/3, where the bond lengths are fixed. The ground state consists of three helices and two loops connecting the helices.

lattice site. If two amino acids are on adjacent sites that are neighbours in the native state, the system reduces its energy by ϵ. Go-like energy terms are usually only defined between heavy atoms in the protein backbone and therefore lack the detail of all-atom force fields. On the other hand, all-atom simulations relying on present energy functions utilize a number of approximations that may lead to additional spurious minima [6–8] and therefore to an energy landscape with an artificially increased roughness. As a consequence, all-atom simulations are usually too slow to allow an efficient study of the folding of stable domains in proteins, which contain of the order of 50–200 residues.

To speed up all-atom simulations one could deform the energy landscape to obtain a steeper folding funnel. In principle, this can be done by adding a Go-like term to the all-atom energy function. For instance, Pogorelov and Luthey-Schulten used this method to speed up molecular dynamics simulations of the folding of the λ-repressor [9]. It is not clear, however, what the optimal coupling is, how the speed up depends on the coupling, and at what coupling the system is dominated by the Go term. A simple Go model as described above is unable to distinguish the native configuration from its mirror image. More complete descriptions, especially for off-lattice models, take the local backbone geometry into account through, for example, bond or dihedral angle potentials [10] or other potentials that represent the stiffness of the backbone [11–13]. Obviously, the combination of an all-atom energy function with a Go term can also be regarded as a way to suppress mirror images in a simulation of a Go model. Again, the question of the optimal coupling appears.

We have studied these questions using a 46-residue segment of protein A (1bdd in the Protein Data Bank) and a variant of the parallel tempering method that will be introduced in the next section. The structure of the protein is shown in figure 1. The segment consists of three helices and short loops connecting the helices. In the following we will first introduce our method; this will be followed by our results and concluding remarks.

2. Methods

Our investigations rely on simulations of protein A with the ECEPP/3 force field [14, 15]. This force field is implemented in the 2005 version of the program package SMMP [16, 17]. The interactions between the atoms within a protein are approximated by a sum $E_{\mathrm{ECEPP/3}}$ consisting of electrostatic energy E_{C}, a Lennard-Jones term E_{LJ}, a hydrogen-bonding term E_{hb}, and a

torsion energy E_{tor}:

$$\begin{aligned} E_{\text{ECEPP/3}} &= E_{\text{C}} + E_{\text{LJ}} + E_{\text{hb}} + E_{\text{tor}} \\ &= \sum_{(i,j)} \frac{332 q_i q_j}{\epsilon r_{ij}} + \sum_{(i,j)} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + \sum_{(i,j)} \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) \\ &\quad + \sum_{l} U_l (1 \pm \cos(n_l \xi_l)), \end{aligned} \tag{1}$$

where r_{ij} is the distance between the atoms i and j, and ξ_l is the lth torsion angle. The factor 332 sets the scale of the electrostatic energy to kcal mol^{-1}. The charges q_i are partial charges on the atoms. The factors A_{ij}, B_{ij}, C_{ij}, and D_{ij} depend on the type of atoms involved, and the factors U_l depend on the residue and the type of dihedral angle. All of these values have been determined empirically and are given in [14, 15]. The magnitudes are chosen such that energies are measured in kcal mol^{-1}.

The all-atom energy of our molecule is the sum of the intramolecular interactions and the ones between protein and the surrounding solvent:

$$E_{\text{aa}} = E_{\text{ECEPP/3}} + E_{\text{solv}}, \tag{2}$$

where the protein–solvent interaction is approximated by a solvent-accessible surface term:

$$E_{\text{solv}} = \sum_i \sigma_i A_i. \tag{3}$$

The sum goes over the solvent-accessible areas A_i of all atoms i weighted by solvation parameters σ_i as determined in [18], a common choice when the ECEPP/3 force field is utilized. Note that E_{solv} is a rather crude approximation of the interaction between the polypeptide and the surrounding water motivated by the low computational costs compared to simulations with explicit water molecules.

The competing interactions in this detailed energy function lead to an energy landscape that is characterized by a multitude of minima separated by high energy barriers. As the probability to cross an energy barrier of height ΔE is given by $\exp(-\Delta E / k_{\text{B}} T)$ (k_{B} the Boltzmann constant) it follows that extremely long runs are necessary to obtain sufficient statistics in regular canonical simulations at a low temperature T.

One popular method to overcome the resulting extremely slow thermalization at low temperatures is parallel tempering [19, 20] (also known as the replica exchange method or multiple Markov chains), a technique first applied to protein studies in [21]. In its most common form, one considers an artificial system built up of N *non-interacting* replicas of the molecule, each at a different temperature T_i. In addition to standard Monte Carlo or molecular dynamics moves that act only on one replica (i.e., the molecule at a fixed temperature), an exchange of conformations between two copies i and $j = i + 1$ is allowed with probability

$$w(\mathbf{C}^{\text{old}} \to \mathbf{C}^{\text{new}}) = \min(1, \exp(-\beta_i E(C_j) - \beta_j E(C_i) + \beta_i E(C_i) + \beta_j E(C_j))). \tag{4}$$

The exchange of conformations leads to a faster convergence of the Markov chain at low temperatures than is observed in regular canonical simulations with only local moves. The resulting random walk in temperatures allows the configurations to move out of local minima and cross energy barriers.

While parallel tempering is traditionally done in temperature space, it can be used with varying potentials as well. The system could be coarse grained across replicas, or the solvent terms could be varied. This idea was introduced under the name 'model hopping' in [22] and is implemented in this paper by varying the strength of an additional Go-like potential term instead of the temperature. With this we can study the effect of a Go-like potential on

the statistics of a Monte Carlo simulation of a protein. Go-like potentials have their origin in lattice models. They reward native contacts with a reduction in energy. If we assume that long- and short-range interactions cooperatively fold the protein into its native structure—this idea is often depicted as a funnel-like structure of the energy landscape—an additional Go-like potential smoothes the energy landscape, which should lead to faster folding.

With the added Go-like energy our energy function becomes

$$E_{\mathrm{tot}} = E_{\mathrm{aa}} + k_{\mathrm{Go}} E_{\mathrm{Go}}, \tag{5}$$

where E_{aa} is the all-atom energy defined above and k_{Go} a parameter that describes the strength of coupling between the two energies. We use the same form for the Go-like energy term as Pogorelov and Luthey-Schulten [9]. It is based on an associative memory Hamiltonian with a single memory. Associative memory Hamiltonians have been used successfully to recognize tertiary structures in proteins [23] and to study protein folding [24]. They capture the long-range effects of protein folding better than, for example, a square well. The form used here can be viewed as a continuum model of the original Go lattice model.

$$E_{\mathrm{Go}} = \sum_{i}^{N_{c\alpha}} \sum_{j \neq i,\pm 1,\pm 2}^{N_{c\alpha}} \gamma_{ij} \exp\left[-\frac{\left(r_{ij} - r_{ij}^{\mathrm{Nat}}\right)^2}{\left(|i-j|^{0.15}\right)^2} \right]. \tag{6}$$

The values of γ_{ij} were chosen as in [9] as $\gamma_{ij} = 0.4$ if $3 \leqslant |i-j| < 9$ and $\gamma_{ij} = 0.5$ if $|i-j| \geqslant 9$, where i and j are the indices of the residues.

We also define an order parameter Q,

$$Q = \frac{1}{N_{\mathrm{contacts}}} \sum_{i}^{N_{c\alpha}} \sum_{j \neq i,\pm 1}^{N_{c\alpha}} \gamma_{ij} \exp\left[-\frac{\left(r_{ij} - r_{ij}^{\mathrm{Nat}}\right)^2}{\left(|i-j|^{0.15}\right)^2} \right], \tag{7}$$

which measures how native-like the current configuration is. It varies between zero and one, where one is the value of the native configuration.

3. Results and discussion

We start by presenting our results for a regular parallel tempering simulation without any Go potential ($k_{\mathrm{Go}} = 0$). Our simulation used 24 replicas with temperatures varied between 297 and 1429 K. Starting from a stretched configuration we performed 100 000 sweeps. Figure 2 displays the specific heat as a function of temperature. The temperature set was optimized following the suggestions by Trebst and Hansmann [8]. We observe a steep peak in the specific heat at $T_1 = 481$ K followed by a broader saddle at a second and lower temperature $T_2 = 338$ K. The two transitions are also visible in our order parameter Q displayed in the inset. The steep increase at the higher temperature T_1 is correlated with a helix–coil transition at this temperature (data not shown), i.e., the formation of short-range contacts, while the second transition at lower temperature T_2 marks the formation of long-range contacts. Figure 3 displays the configuration with lowest energy obtained in the simulation. It has an all-atom root-mean-square deviation (rmsd) of 3.2 Å for residue 16–46. The N-terminal helix, however, has the wrong orientation and the rmsd over all residues is therefore large at 8.8 Å. The configuration with the highest value of Q is displayed in figure 4. Here, the orientation of the N-terminal helix is correct, leading to an all-atom rmsd of 3.4 Å (over all residues) and a solvent-accessible surface area of 3680 Å^2 that is smaller than the one (4340 Å^2) for the minimal energy configuration of figure 3. However, the energy of this

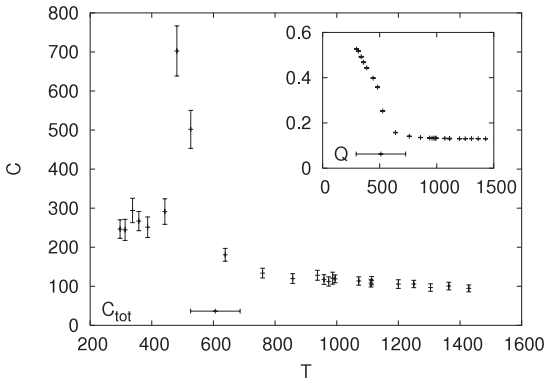

Figure 2. Specific heat versus T of an unbiased parallel tempering run. The sharp specific heat peak at $T_1 = 481$ is correlated with the helix–coil transition.

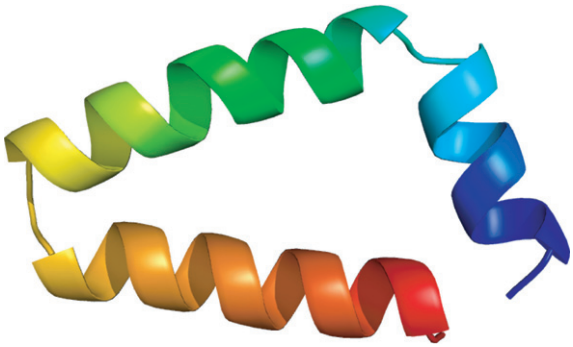

Figure 3. Minimum energy configuration from the unbiased parallel tempering run. The N-terminal helix has the wrong orientation and the rmsd over all residues is therefore large, at 8.8 Å. The all-atom rmsd for residues 16–46 is 3.2 Å.

configuration is, at $E = -567.2$ kcal mol^{-1}, almost 50 kcal mol^{-1} higher than that of the minimal energy configuration ($E = -614.6$ kcal mol^{-1}). This is because the ECEPP force field overemphasizes helix formation. For protein A this leads to the formation of three helices that are more elongated than observed in the native structure and therefore are too stiff to arrange themselves into the correct configuration. Consequently, the higher energy of the configuration with maximal order parameter Q is due to the intramolecular energy term E_{ECEPP} (-378.0 kcal mol^{-1} versus -431.5 kcal mol^{-1}), while the solvation energy E_{solv} is slightly lower (-189.2 kcal mol^{-1} versus -183.1 kcal mol^{-1}). From our result it is not clear whether the global minimum energy configuration would be native-like and just was not found in the simulation, or whether it differs for this force field from the native structure of figure 1. In either case this indicates problems with our energy function that limit its use in protein simulations.

The situation is different in simulations with a Go-energy function. Here, it is ensured by definition of the energy that the global minimum configuration is the native structure (or its mirror configuration). This can be seen in figure 5 which displays the results from a simulation with only the Go term of (6). The replicas differ here in the value of the Go parameter k_{Go}, i.e., the true inverse temperature in the system is βk_{Go} (with β the inverse temperature corresponding to $T = 300$ K). Shown again is the specific heat, and in the inset our order parameter Q. The system does not seem to have any transition. The order parameter increases

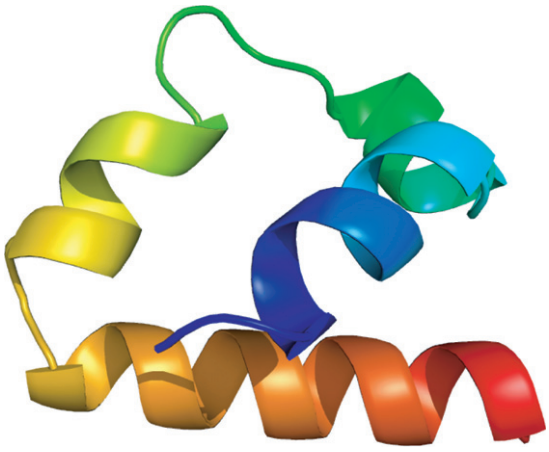

Figure 4. Most native-like configuration from the unbiased parallel tempering run. The orientation of the N-terminal helix is correct, leading to an all-atom rmsd of 3.4 Å over all residues.

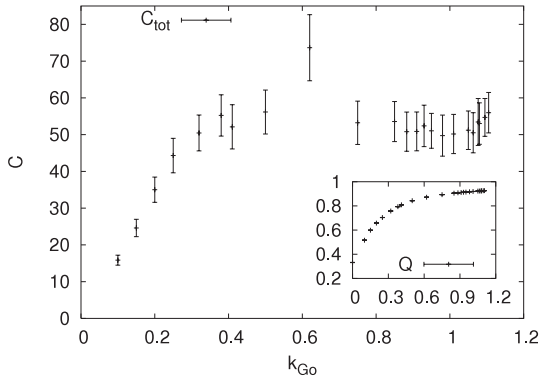

Figure 5. Specific heat C versus Go parameter k_{Go}. There is no apparent specific heat peak. The inset shows a smooth increase in Q. The structures become increasingly native-like as k_{Go} increases.

monotonically. No pronounced peak is observed in the specific heat. By construction of the energy function, the lowest energy configuration is also the one with the largest Q value, and this is shown in figure 6. Note that this structure is actually a mirror configuration and therefore the rmsd is 8 Å, larger than one would expect from visual inspection. If we add a small all-atom energy contribution, it breaks the symmetry of the Go potential, and the lowest energy structure found, with an all-atom rmsd of 3.49 Å (1.87 Å C_α), is very close to the native structure.

In the following we study how the bias introduced by a Go term affects the outcome of an all-atom simulation. For this purpose we study our protein at a temperature $T = 300$ K just below the folding temperature T_2, varying the strength of the contribution of the Go term to the total energy of the system over the ladder of replicas. Figure 7 shows the various energy terms as a function of the coupling strength k_{Go} of the Go term. As expected, the Go energy decreases with increasing strength of coupling. However, the all-atom energy stays constant, i.e., does not change with the introduction of the additional Go term. The superposition of the two energy terms leads to a total energy E_{tot} that sharply decreases for $k_{\mathrm{Go}} \geqslant 0.2$. Hence, for a 'critical' k_{Go}, the contribution from the Go term starts dominating the system. We therefore

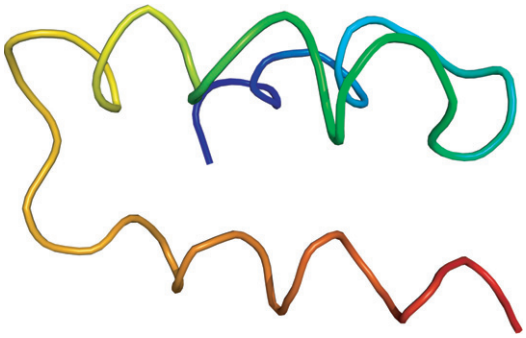

Figure 6. Minimum energy configuration from a parallel tempering with Go energy only. The Go energy does not distinguish between the native structure and its mirror image. In this run we obtained the mirror image of the native structure.

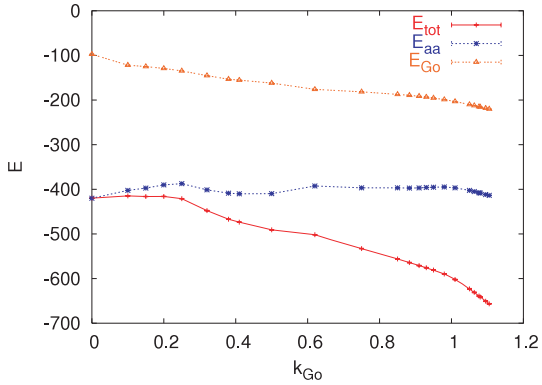

Figure 7. Total, all-atom, and Go energy versus k_{Go} at constant temperature $T = 300$ K. At $k_{\mathrm{Go}} \approx 0.2$ the Go energy starts to dominate the behaviour of the total energy.

conjecture that $k_{\mathrm{Go}} = 0.2$ is the optimal value for coupling of the two energy terms. For a lower value, the influence of the Go term is too weak to be effective, while for a larger value the system behaves as a Go model.

Fixing now $k_{\mathrm{Go}} = 0.2$, we again vary the all-atom temperature across the replica. The resulting lowest energy configuration is shown in figure 8. It has an all-atom rmsd of 4.5 Å over all residues (compared to 8.8 Å for the case without a coupled Go term). When comparing the all-atom energies, we find that the value for this configuration $E_{\mathrm{aa}} = -580.4$ kcal mol^{-1} is higher than that of the free case ($E_{\mathrm{aa}} = -614$ kcal mol^{-1}). Hence, it is not so that the additional Go term solely smooths the energy landscape and increases in this way the chances of finding native-like configurations as the true global minimum. Rather, we conjecture that for protein A the global minimum in our all-atom force field is not the native structure. Only by adding the Go term is the energy landscape deformed in a way such that the native structure (being a sub-optimal competing local minimum in the all-atom energy) becomes the global minimum in the total energy.

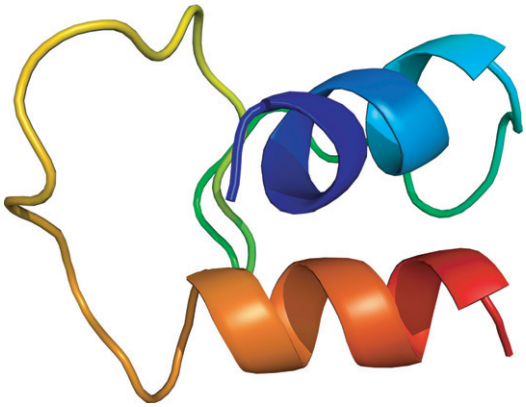

Figure 8. Minimum energy configuration from a biased parallel tempering run with $k_{\mathrm{Go}} = 0.2$.

4. Conclusions

We have performed simulations of the 46 amino-acid long segment of protein A. Simulating the protein with a 'physical' all-atom force field, we find low-energy configurations that are similar to the native structure, but the global minimum configuration differs significantly (by $\approx$8 Å) from this. Addition of a Go term leads to a global minimum (in the combined energy) in the simulation that is close to the native one. However, its all-atom energy is higher than the one found for the global minimum found in a simulation relying only on an all-atom force field. We conclude that the Go term deforms the energy landscape in a way that the native structure becomes the global minimum in the combined energy but that the minimum is not the one for the all-atom force field. The introduction of the Go term does not merely smooth the energy landscape leading to a faster simulation of the system, but causes a large deformation of the energy landscape. The results differ qualitatively from the results obtained from an all-atom simulation. Adding a Go term like the one used in this paper is therefore not a suitable tool for the faster thermalization of all-atom simulations.

Acknowledgments

UH acknowledges support by a research grant (CHE-0313618) of the National Science Foundation (USA). The simulations were done on the Cray XD1 and the JUMP supercomputer of the John von Neumann Institute for Computing at the Research Center Jülich.

References

[1] Anfinsen C B 1973 Principles that govern the folding of proteins *Science* **181** 223–30
[2] 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
[3] Abe H and Go N 1981 Noninteracting local-structure model of folding and unfolding transition in globular proteins: II. Application to two-dimensional lattice proteins *Biopolymers* **20** 1013–31
[4] Go N 1983 Theoretical studies of protein folding *Annu. Rev. Biophys. Bioeng.* **12** 183–210
[5] Takada S 1999 Go-ing for the prediction of protein folding mechanisms *Proc. Natl Acad. Sci. USA* **96** 11698–700
[6] Hansmann U H E 2004 Generalized-ensemble simulations of the human parathyroid hormone fragment pth(1–34) *J. Chem. Phys.* **120** 417–22
[7] Lin C-Y, Hu C-K and Hansmann U H E 2003 Parallel tempering simulations of HP-36 *Proteins* **52** 436–45

[8] Trebst S, Troyer M and Hansmann U H E 2006 Optimized parallel tempering simulations of proteins *J. Chem. Phys.* **124** 174903
[9] Pogorelov T V and Luthey-Schulten Z 2004 Variations in the fast folding rates of the l-repressor: a hybrid molecular dynamics study *Biophys. J.* **87** 207–14
[10] Clementi C, Nymeyer H and Onuchic J N 2000 Topological and energetic factors: What determines the structural details of the transition state ensemble and 'en-route' intermediates for protein folding? An investigation for small globular proteins *J. Mol. Biol.* **298** 937–53
[11] Kwiecinska J I and Cieplak M 2005 Chirality and protein folding *J. Phys.: Condens. Matter* **17** S1565
[12] Hoang T X and Cieplak M 2000 Molecular dynamics of folding of secondary structures in Go-type models of proteins *J. Chem. Phys.* **112** 6851
[13] Hoang T X and Cieplak M 2000 Sequencing of folding events in Go-type proteins *J. Chem. Phys.* **113** 8319–28
[14] Sippl M J, Némthy G and Scheraga H A 1984 Intermolecular potentials from crystal data. 6. Determination of empirical potentials for O–H***O=C hydrogen bonds from packing configurations *J. Phys. Chem.* **88** 6231–3
[15] Nemethy G, Gibson K D, Palmer K A, Yoon C N, Paterlini G, Zagari A, Rumsey S and Scheraga H A 1992 Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides *J. Phys. Chem.* **96** 6472–84
[16] Eisenmenger F, Hansmann U H E, Hayryan S and Hu C-K 2001 [SMMP] A modern package for simulation of proteins *Comput. Phys. Commun.* **138** 192–212
[17] Eisenmenger F, Hansmann U H E, Hayryan S and Hu C-K 2006 An enhanced version of SMMP—an open-source software package for simulation of proteins *Comput. Phys. Commun.* **174** 422–9
[18] Ooi T, Oobatake M, Nemethy G and Scheraga H A 1987 Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides *Proc. Natl Acad. Sci. USA* **84** 3086–90
[19] Hukushima K and Nemoto K 1996 Exchange Monte Carlo method and application to spin glass simulations *J. Phys. Soc. Japan* **65** 1604–8
[20] Geyer C J and Thompson E A 1995 Annealing Markov chain Monte Carlo with applications to ancestral inference *J. Am. Stat. Assoc.* **90** 909
[21] Hansmann U H E 1997 Parallel tempering algorithm for conformational studies of biological molecules *Chem. Phys. Lett.* **281** 140–50
[22] Kwak W and Hansmann U H E 2005 Efficient sampling of protein structures by model hopping *Phys. Rev. Lett.* **95** 138102
[23] Friedrichs M S and Wolynes P G 1989 Towards protein tertiary structure recognition by means of associative memory Hamiltonians *Science* **246** 371–3
[24] Hardin C, Luthey-Schulten Z and Wolynes P G 1999 Backbone dynamics, fast folding, and secondary structure formation in helical proteins and peptides *Proteins* **34** 281–94