Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Molecular Graphics and Modelling 22 (2004) 451–463
Journal of Molecular Graphics and Modelling
www.elsevier.com/locate/JMGM

Exploring the protein folding free energy landscape: coupling replica exchange method with P3ME/RESPA algorithm

Ruhong Zhou

Computational Biology Center, IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA
Department of Chemistry, Columbia University, New York, NY 10027, USA

Abstract

A highly parallel replica exchange method (REM) that couples with a newly developed molecular dynamics algorithm particle–particle particle–mesh Ewald (P3ME)/RESPA has been proposed for efficient sampling of protein folding free energy landscape. The algorithm is then applied to two separate protein systems, β-hairpin and a designed protein Trp-cage. The all-atom OPLSAA force field with an explicit solvent model is used for both protein folding simulations. Up to 64 replicas of solvated protein systems are simulated in parallel over a wide range of temperatures. The combined trajectories in temperature and configurational space allow a replica to overcome free energy barriers present at low temperatures. These large scale simulations reveal detailed results on folding mechanisms, intermediate state structures, thermodynamic properties and the temperature dependences for both protein systems.

Keywords: Protein folding; Free energy landscape; Conformation space sampling; Multiple time step; P3ME

1. Introduction

How to efficiently sample the conformational space of complex systems, such as protein folding, still remains a great challenge. The free energy landscape of protein folding is believed to be at least partially rugged. At room temperature, protein systems get trapped in many local minima. This trapping limits the capacity to effectively sample configurational space. Many methods have been proposed to enhance the conformation space sampling [1–7]. These methods include multicanonical sampling, simulated tempering, parallel tempering, catalytic tempering, and expanded ensembles. Despite the enormous effort of many groups, it is still difficult, even with today's supercomputers, to perform realistic all-atom explicit solvent simulations for protein folding. In terms of molecular dynamics (MD), microsecond to millisecond simulations, which are believed to be the folding time of most proteins, are still beyond the current capacity [8,9]. The only microsecond simulation with an all-atom model and explicit solvent was done by Duan and Kollman on a villin head piece [10]

Here, we propose a replica exchange method (REM) or parallel tempering [6,11] coupled with a newly developed MD algorithm particle–particle particle–mesh–Ewald (P3ME)/RESPA for efficient sampling of the protein folding conformation space. In the REM method, replicas are run in parallel at a sequence of temperatures ranging from the desired temperature to a high temperature at which the replica can easily surmount the energy barriers. From time to time the configurations of neighboring replicas are exchanged and this exchange is accepted by a Metropolis acceptance criterion that guarantees the detailed balance. Thus, REM is essentially a Monte Carlo (MC) method, and in fact, the early implementations were based on MC simulations. Because the high temperature replica can traverse high energy barriers there is a mechanism for the low temperature replicas to overcome the quasi ergodicity they would encounter in a single temperature replica. The replicas can be generated by MC, by Hybrid Monte Carlo (HMC) as was used in our recent implementation of J-Walking and S-Walking [12], or by MD with velocity rescaling. Sugita and Okamoto [13] have developed a temperature rescaling scheme for coupling MD with REM. The good thing with the MD approach is that new advances in MD algorithms can be naturally coupled to the REM method. In this study, we will combine REM with a newly developed MD algorithm [14] which efficiently couples the multiple time step algorithm RESPA (reference system propagator algorithm) [15] with the P3ME [16]. The combined method will then be applied to the folding study of two separate small protein systems in an explicit solvent. One is the C-terminal β-hairpin of protein G, and the other is a designed mini-protein Trp-cage.

The β-hairpin has received much attention recently from both experimental and theoretical fronts [17–26]. Parallel

E-mail address: ruhongz@us.ibm.com (R. Zhou).

1093-3263/$ – see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2003.12.011

and anti-parallel beta sheets and alpha helices are the key secondary structures in proteins. It is believed that understanding the folding of these elements will be a foundation for investigating larger and more complex structures. The study of isolated beta sheets has for a long time been limited by the lack of an amenable experimental system. The breakthrough experiments by the Serrano and coworkers [19] and Eaton and coworkers [17] have recently established the β-hairpin from the C-terminus of protein G as the system of choice to study beta sheets in isolation. These pioneering experiments inspired a number of theoretical works on this system with various models [20–23,27–29]. However, there are still a number of important aspects that remain controversial. For instance, the folding pathway intermediates are observed in some but not all studies; the relative importance of the intra-strand hydrogen bonds compared with hydrophobic core formation; and the existence of helical structures during the folding process. The current study will use an all-atom OPLSAA force field and an explicit solvent SPC model to explore the folding free energy landscape of this β-hairpin. A total of 64 replicas of the solvated system consisting of 4342 atoms will be simulated with temperatures spanning from 270 to 695 K. Some of the β-hairpin results have been published previously [28], so we only use it to demonstrate the power of REM–P3ME/RESPA algorithm. The main focus will be on the following Trp-cage folding.

The 20-residue mini-protein Trp-cage (NLYIQ WLKDG GPSSG RPPPS) is the fastest folding protein known so far [30,31]. This designed mini-protein folds spontaneously and co-operatively into a Trp-cage in about 4 μs [31]. It contains a short alpha-helix from residue 2–9, a 3_{10}-helix from residues 11–14, and a C-terminal poly-proline II helix packing against the central tryptophan (Trp6) [30,31]. Stability data suggests that a salt-bridge connects Asp9 with Arg16 in the folded state. The folding appears highly cooperative, with circular dichroism (CD), fluorescence, and chemical shift deviations (CSD) generating virtually identical thermal denaturation profiles [30,31]. The small size, high stability and fast folding time of this Trp-cage make it an ideal choice for protein folding simulations. There are several simulations done so far [32–34] all using a continuum solvent model GBSA. To our knowledge, this is the first explicit solvent simulation. Again, the all-atom OPLSAA force field and SPC explicit water model are used with periodic boundary condition. A total of 50 replicas of the solvated system consisting of 12,242 atoms are simulated with temperatures spanning from 282 to 598 K.

2. Methodology and systems

2.1. Replica exchange method

The replica exchange method (REM) has been implemented in the context of the molecular modeling package IMPACT [12,35]. As mentioned above, the original REM was implemented with MC methods, and here we follow a similar approach as Sugita and Okamoto's [13] to combine MD with REM. We have also implemented a combination of HMC with REM, which is more efficient for smaller systems. A brief discussion of the REM method based on molecular dynamics is described in the following.

Suppose there is a system of N atoms with masses m_k ($k = 1, 2, \ldots, N$), and coordinates and momenta $q \equiv \{\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_N\}$ and $p \equiv \{\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_N\}$, the Hamiltonian $H(p, q)$ of the system can be expressed as:

$$H(p,q) = \sum_{k=1}^{N} \frac{p_k^2}{2m_k} + U(q), \tag{1}$$

where $U(q)$ is the potential energy of the N atom system. In the canonical ensemble at temperature T, each state $x \equiv (p, q)$ with the Hamiltonian $H(p, q)$ is weighted by the Boltzmann factor:

$$\rho(x;T) = \frac{1}{Z} \exp^{[-\beta H(p,q)]}, \tag{2}$$

where $\beta = 1/k_B T$ (k_B is the Boltzmann constant), and Z is the partition function $Z = \int \exp^{[-\beta H(p,q)]} \mathrm{d}p\, \mathrm{d}q$. The generalized ensemble for REM consists of M non-interacting replicas of the original system at M different temperatures T_m ($m = 1, 2, \ldots, M$). The replicas are arranged such that there is always exactly one replica at each temperature. Then there is a one-to-one correspondence between replicas and temperatures, the label i ($i = 1, 2, \ldots, M$) for replicas is a permutation of the label m ($m = 1, 2, \ldots, M$) for temperatures and vice versa,

$$i = i(m) \equiv f(m), \qquad m = m(i) \equiv f^{-1}(i), \tag{3}$$

where $f(m)$ is a permutation function of m, and $f^{-1}(i)$ is its inverse.

The meta state X of this generalized ensemble will be a collection of all the M sets of coordinates $q^{[i]}$ and momenta $p^{[i]}$ of the N atoms in replica i at temperature T_m : $x_m^{[i]} \equiv (p^{[i]}, q^{[i]})_m$:

$$X = (x_1^{[i(1)]}, \ldots, x_M^{[i(M)]}) = (x_{m(1)}^{[1]}, \ldots, x_{m(M)}^{[M]}), \tag{4}$$

where the superscript and the subscript in $x_m^{[i]}$ label the replica and the temperature indices, respectively, which have a one-to-one correspondence. Because the replicas are non-interacting, the weight factor for the state X in this generalized ensemble is given by the product of Boltzmann factors for each replica or each temperature:

$$\begin{aligned}\rho_{\mathrm{REM}}(X) &= \exp\left\{-\sum_{i=1}^{M} \beta_{m(i)} H(p^{[i]}, q^{[i]})\right\} \\ &= \exp\left\{-\sum_{m=1}^{M} \beta_m H(p^{[i(m)]}, q^{[i(m)]})\right\},\end{aligned} \tag{5}$$

where $i(m)$ and $m(i)$ are the permutation functions defined in Eq. (3). Now suppose a pair of replicas is exchanged. For

generality, we assume the pair being swapped is (i, j) which are at temperatures (T_m, T_n), respectively,

$$X = (\ldots, x_m^{[i]}, \ldots, x_n^{[j]}, \ldots,) \rightarrow X' = (\ldots, x_m^{[j]'}, \ldots, x_n^{[i]'}, \ldots). \quad (6)$$

The indices i, j, and m, n are related by the permutation function. Upon the exchange, the permutation function will be updated, let us rename it f':

$$i = f(m) \rightarrow j = f'(m), \qquad j = f(n) \rightarrow i = f'(n). \quad (7)$$

The above exchange of replicas can be rewritten in more detail as:

$$x_m^{[i]} = (p^{[i]}, q^{[i]})_m \rightarrow x_m^{[j]'} = (p^{[j]'}, q^{[j]})_m,$$
$$x_n^{[j]} = (p^{[j]}, q^{[j]})_n \rightarrow x_n^{[i]'} = (p^{[i]'}, q^{[i]})_n, \quad (8)$$

where the new momenta $p^{[i]'}$ and $p^{[j]'}$ will be defined below. It is easy to see that this process of exchanging a pair of replicas (i, j) is equivalent to exchanging the two corresponding temperatures T_m and T_n:

$$x_m^{[i]} = (p^{[i]}, q^{[i]})_m \rightarrow x_n^{[i]'} = (p^{[i]'}, q^{[i]})_n,$$
$$x_n^{[j]} = (p^{[j]}, q^{[j]})_n \rightarrow x_m^{[j]'} = (p^{[j]'}, q^{[j]})_m. \quad (9)$$

This mathematical equivalence is very useful in practice, since it can be used to reduce the communication costs in REM, i.e., rather than exchanging the two full sets of coordinates and momenta, one can just swap the two temperatures for the two replicas and then update the permutation function. In the original implementations of the REM [6,11], Monte Carlo algorithms were used, thus only the coordinates q and the potential energy $U(q)$ need to be taken into account. In order for this exchange process to generate the equilibrium canonical distribution functions, it is necessary and sufficient to impose the detailed balance condition on the transition probability $T(X \rightarrow X')$ from meta state X to X',

$$\rho_{\mathrm{REM}}(X) T(X \rightarrow X') = \rho_{\mathrm{REM}}(X') T(X' \rightarrow X). \quad (10)$$

From Eqs. (1), (5) and (10), one can easily derive that

$$\frac{T(X \rightarrow X')}{T(X' \rightarrow X)} = \exp(-\Delta), \quad (11)$$

where

$$\Delta = (\beta_m - \beta_n)(H(x^{[j]}) - H(x^{[i]})). \quad (12)$$

For molecular dynamics simulations, both the potential energy and kinetic energy are present in the Hamiltonian, thus, Sugita and Okamoto [13] introduced a momenta rescaling scheme to simplify the detailed balance condition,

$$p_n^{[i]'} = \sqrt{\frac{T_n}{T_m}} p_m^{[i]}, \qquad p_m^{[j]'} = \sqrt{\frac{T_m}{T_n}} p_n^{[j]}. \quad (13)$$

With the above velocity rescaling scheme, the detailed balance equation can be reduced into

$$\Delta = (\beta_m - \beta_n)(U(q^{[j]}) - U(q^{[i]})). \quad (14)$$

Note that because of the velocity rescaling in Eq. (13) the kinetic energy terms are cancelled out in the above detailed balance condition, and that the same criterion, Eq. (14), which was originally derived for Monte Carlo algorithm, is recovered. It should also be noted that this detailed balance criterion is exactly the same as in Jump Walking methods [4,12]. The above detailed balance condition can be easily satisfied, for instance, by the usual Metropolis criterion,

$$T(X \rightarrow X') \equiv T(x_m^{[i]} | x_n^{[j]}) = \begin{cases} 1, & \text{for } \Delta \leq 0, \\ \exp(-\Delta), & \text{for } \Delta \geq 0. \end{cases} \quad (15)$$

Thus, the replica exchange method can be summarized as the following two-step algorithm:

(1) Each replica i $(i = 1, 2, \ldots, M)$, which is in a canonical ensemble of the fixed temperature T_m $(m = 1, 2, \ldots, M)$, is simulated simultaneously and independently for a certain MC or MD steps.
(2) Pick some pairs of replicas, for example $x_m^{[i]}$ and $x_n^{[j]}$, and exchange the replicas with the probability $T(x_m^{[i]} | x_n^{[j]})$ as defined in Eq. (15), and then go back to step (1).

In the present work, the MD algorithm is used in step (1) and all the replicas are run in parallel; and in step (2), only the replicas in neighboring temperatures are attempted for exchanges because the acceptance ratio of the exchange decreases exponentially with the difference of the two beta. Note that whenever a replica exchange is accepted in step (2), the permutation functions in Eq. (3) must be updated.

The major advantage of REM over other generalized ensemble methods such as multicanonical algorithm and simulated tempering lies in the fact that the weight factor is a priori known (see Eq. (5)), while in the latter algorithms the determination of the weight factors can be non-trivial for complex systems and very time consuming. In REM method, a random walk in the "temperature" space is realized for each replica, which in turn induces a random walk in potential energy space. This alleviates the problem of being trapped in states of energy local minimum.

2.2. P3ME/RESPA algorithm

The technique of Ewald sums is very useful for complex systems, such as solvated proteins, which have a large number of partial charges, since the long-ranged Coulomb interactions do not converge sufficiently when summed over a single unit cell. The slowly and conditionally converging sum of electrostatic interactions

$$U^{\mathrm{elec}} = \sum_{\boldsymbol{n}} \sum_{i<j}{}' \frac{q_i q_j}{|\boldsymbol{r}_{ij} + \boldsymbol{n}L|}, \quad (16)$$

is rearranged so that part of it is summed in real space, and the rest is summed in Fourier space,

$$U^{\mathrm{elec}} = \sum_{i<j} q_i q_j \frac{\mathrm{erfc}(\alpha r_{ij})}{r_{ij}} + \sum_{i<j}\sum_{k\neq 0} \frac{1}{\pi L^3}\frac{4\pi^2}{k^2} q_i q_j \mathrm{e}^{-k^2/4\alpha^2} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}_{ij}} - \frac{-\alpha}{\sqrt{\pi}}\sum_i q_i^2, \tag{17}$$

where the metallic boundary condition is used.

With a suitable choice for the screening parameter alpha (α), both sums can be made to converge reasonably quickly. More specifically, α is usually chosen so that the first term in the expression above (the real-space sum) is adequately converged within a radius of no more than $r = L/2$, where L is the side length of the cubic unit cell. Therefore, the first term includes primarily short-ranged interactions. The second term (the k-space sum), on the other hand, results from a Fourier expansion of the potential due to an infinite array of Gaussian charges, many of which are considerably longer-ranged than in the real-space sum.

Mesh-based Ewald summation methods, including particle mesh Ewald (PME) [36] and P3ME [16], provide an approximation to the reciprocal space term of the Ewald sum by assigning the point charges to a finite sized grid. The other terms in the Ewald sum are left unchanged. The mesh allows the k-space sum to be evaluated using the fast Fourier transform (FFT), which scales as $\mathrm{O}(N \log N)$. If one chooses a large enough value for the Ewald parameter α, (sufficiently small real space cut-off), the $N \log N$ scaling extends to the entire calculation. Since the details of the procedure for calculating the electrostatics using P3ME has been described before [14,16,37], we only give a brief overview here. Typically, the procedure consists of four steps: (1) assigning charges to the grid; (2) solving Poisson's equation on the grid; (3) differentiation to determine the forces; and (4) interpolating the forces on the grid back to particles. We have also improved the optimal influence function $\tilde{I}_{\mathrm{opt}}(k)$ used in step (2) for the Fourier transform of charge densities. Interested readers can consult the previous reference [14] for more details.

On the other hand, the multiple time step algorithms such as RESPA are based on subdividing the inter-particle forces into a hierarchy of fast and slow forces, which allows the slow forces to be integrated with a larger timestep, and the fast forces with a smaller timestep. The increase in efficiency springs from the fact that the slowest parts of the force, usually the long-range pairwise interactions in force fields, accounts for the vast majority of the computation. Various implementations of the RESPA method have been applied to a wide variety of systems resulting in speedups by factors of 4–15 [14,15,35].

In general, the choice of how to subdivide the forces is critical, and the most useful split is often dictated by the physics of the problem at hand. Occasionally, however, several different choices seem appropriate, and sometimes the most obvious factorization does not turn out to be the most efficient. The aim here is to outline the most efficient split for systems with long range electrostatic forces treated by a class of Ewald-type methods, such as Ewald, PME, and P3ME, etc. Without losing generality, we use the Ewald method as an example to illustrate the coupling with RESPA, since only the k-space part is treated differently in all these Ewald-type methods.

Since Ewald sums are used to evaluate long-ranged Coulombic interactions, it seems natural to use them as a basis for separating near (real-space) and far (k-space) forces in a RESPA split. Under the usual assumption that long-ranged forces may be updated less frequently than short-ranged forces, it thus seems reasonable to separate the real- and k-space sums in a RESPA split. For example, if we rewrite the Ewald sum in the form

$$U^{\mathrm{elec}} = \frac{1}{2}\sum_i \sum_j U_{ij}^{\mathrm{elec}}, \tag{18}$$

where

$$U_{ij}^{\mathrm{elec}} = q_i q_j \left[(1-\delta_{ij}) \frac{\mathrm{erfc}(\alpha r_{ij})}{r_{ij}} + \frac{1}{\pi L^3}\sum_{k\neq 0} \frac{4\pi^2}{k^2} \mathrm{e}^{-k2/4\alpha^2} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}_{ij}} - \delta_{ij}\frac{2\alpha}{\sqrt{\pi}} \right], \tag{19}$$

then we can separate the real- and k-space parts of the potential,

$$U_{ij}^{\mathrm{elec}} = U_{ij}^{\mathrm{rs}} + U_{ij}^{\mathrm{ks}}, \tag{20}$$

with

$$U_{ij}^{\mathrm{rs}} = (1-\delta_{ij}) q_i q_j \frac{\mathrm{erfc}(\alpha r_{ij})}{r_{ij}}, \tag{21}$$

and

$$U_{ij}^{\mathrm{ks}} = q_i q_j \left[\frac{1}{\pi L^3}\sum_{k\neq 0} \frac{4\pi^2}{k^2} \mathrm{e}^{-k2/4\alpha^2} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}_{ij}} - \delta_{ij}\frac{2\alpha}{\sqrt{\pi}} \right]. \tag{22}$$

The forces can be easily obtained by taking derivatives of the above potential. Then, we can use the RESPA integrator to propagate the dynamics. The real-space forces could also be further subdivided into distance classes, if desired. Such an approach seems perfectly reasonable, given the disparity in distances over which the terms in the real- and k-space sums act. Indeed, an approach very similar to this has been used recently in large-scale Ewald simulations of proteins [38].

Although this particular RESPA split is moderately successful, it is not necessarily the best choice. The reason for this is that the "long-ranged" k-space sum still contains some fraction of every pair interaction, even the most short-ranged. This can be seen by re-expressing the Coulomb term as

$$\frac{q_i q_j}{r_{ij}} = \frac{q_i q_j}{r_{ij}}\{\mathrm{erfc}(\alpha r_{ij}) + \mathrm{erf}(\alpha r_{ij})\}, \tag{23}$$

where we have used the identity,

$$\mathrm{erfc}(x) + \mathrm{erf}(x) = 1, \tag{24}$$

Comparison of Eq. (23) with Eq. (17) shows that the reciprocal space part given in Eq. (22) implicitly contains the $\mathrm{erf}(\alpha r_{ij})$ terms for every pair, even those very close ones within the real space cut-off. Thus the breakup of the forces suggested in Eqs. (21) and (22) is not optimal as the reciprocal space part of the force will vary rapidly for pairs that are close to each other. The presence of these short-ranged interactions in the *k*-space sum will limit the size of the large timestep Δt more than would be necessary if the slow piece of the propagator were truly long-ranged. Indeed, in the published report by Procacci and Marchi [38] which uses this propagator, the *k*-space forces required a timestep which was even shorter than that used for the outer shell of the real-space forces (the real space forces are splitted into two shells [38]).

A better alternative would be to remove the fast part of the $\mathrm{erf}(\alpha r_{ij})$ contributions from the reciprocal space terms in Eq. (22), and add it back to the real space term in Eq. (21). The term to be subtracted and added is

$$\Delta = (1 - \delta_{ij}) \frac{q_i q_j}{r_{ij}} \mathrm{erf}(\alpha r_{ij}). \tag{25}$$

Thus, the new breakup will be purely based on the "fast" and "slow" motions:

$$U_{ij}^{\mathrm{elec}} = U_{ij}^{(\mathrm{f})} + U_{ij}^{(\mathrm{s})}, \tag{26}$$

where the new "real space" (fast force) part contains the rapidly varying part of the potential,

$$U_{ij}^{(\mathrm{f})} = (1 - \delta_{ij}) q_i q_j \frac{1}{r_{ij}}. \tag{27}$$

It should be noted that Eq. (27) is equivalent to the usual minimum image real space energy with a short range cut-off. The new "*k*-space" (slow force) term becomes,

$$U_{ij}^{(\mathrm{s})} = q_i q_j \left[- (1 - \delta_{ij}) \frac{\mathrm{erf}(\alpha r_{ij})}{r_{ij}} + \frac{1}{\pi L^3} \sum_{k \neq 0} \frac{4\pi^2}{k^2} \mathrm{e}^{-k^2/4\alpha^2} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}_{ij}} - \delta_{ij} \frac{2\alpha}{\sqrt{\pi}} \right]. \tag{28}$$

This new breakup of the potential leads to a subdivision of the forces on the basis of the distance over which they act, regardless of whether they are real-space or *k*-space forces.

For complex systems, such as the solvated protein systems in the present study, the potential (force) is more complicated and it contains several terms in a typical force field, such as OPLSAA [39],

$$F(x) = F_{\mathrm{stret}}(x) + F_{\mathrm{bend}}(x) + F_{\mathrm{tors}}(x) + F_{\mathrm{vdW}}(x) + F_{\mathrm{elec}}(x), \tag{29}$$

where F_{stret}, F_{bend}, F_{tors}; F_{vdW}, and F_{elec} represent the forces for stretching, bending, torsion, van der Waals, and electrostatic interactions, respectively. The forces are then splitted in RESPA according to their intrinsic time scales and the above special treatment for electrostatics:

$$F_0(x) = F_{\mathrm{stret}}(x) + F_{\mathrm{bend}}(x) + F_{\mathrm{tors}}(x), \tag{30}$$

$$F_1(x) = F_{\mathrm{vdW}}^{\mathrm{near}}(x) + F_{\mathrm{elec}}^{\mathrm{near}}(x), \tag{31}$$

$$F_2(x) = F_{\mathrm{vdW}}^{\mathrm{med}}(x) + F_{\mathrm{elec}}^{\mathrm{med}}(x), \tag{32}$$

$$F_3(x) = F_{\mathrm{elec}}^{\mathrm{far}}(x). \tag{33}$$

The fast varying bonded forces are included in $F_0(x)$. The SHAKE/RATTLE [40] algorithm is used to constrain the bond lengths in this study. If no constraints used, the bond stretching force should be separated from the other bonded forces since it varies fastest [35]. This is part of the innermost "reference" propagator. The non-bonded forces are separated into three different time scales according to pair distances, near-range ($F_1(x)$), intermediate-range ($F_2(x)$), and far-range ($F_3(x)$). $F_1(x)$ is defined as normal van der Waals and modified Ewald "real space" electronic forces defined in Eq. (27) with a pair distance less than 7 Å; $F_2(x)$ as van der Waals and "real space" electrostatic forces with pair distances between 6 and 10 Å (there is some overlap due to a switching function applied between 6 and 7 Å to make the forces smooth); and $F_3(x)$ as the modified "*k*-space" electrostatic forces defined in Eq. (28).

The RESPA split proposed here for the Ewald method is equally applicable to the P3ME and PME methods. The P3ME or PME approximations apply only to the reciprocal space term, thus nothing special is needed for the RESPA division. The only difference will lie in choice of alpha and appropriate cut-offs that are to be optimized for each method.

Finally, the coupling of P3ME/RESPA with REM is straightforward. The P3ME/RESPA algorithm efficiently speeds up the molecular dynamics simulation by about an order of magnitude for a normal-size solvated protein system, which is then utilized as the underlying sampling engine for REM through either velocity rescaling or Hybrid Monte Carlo. The Hybrid Monte Carlo is more efficient for smaller systems, such as protein in a continuum solvent [29], while the velocity rescaling approach is more appropriate for the explicit solvent simulations. Either way, the good thing about using MD as underlying sampling engine in REM is that the new advances in molecular dynamics algorithms can be easily incorporated into replica exchange Monte Carlo scheme. As shown below, the combined method is a powerful tool for protein folding conformational space sampling.

2.3. Protein systems

The β-hairpin system of this study is taken from the C terminus (residues 41–56) of protein G (PDB entry

2gb1). The β-hairpin is capped with the normal Ace and Nme groups, resulting in a blocked peptide sequence of Ace-GEWTYDDATKTFTVTE-Nme. The folding time for this hairpin is about 6 μs, which is about two to three orders longer than what we can routinely simulate in a single MD simulation. The solvated system has 1361 water molecules (SPC water, with density 1.0 g/cm^3) and also three counter ions ($3Na^+$) for neutralizing the molecular system, which results in a total of 4342 atoms in each replica. A total of 64 replicas are simulated with temperatures spanning from 270 to 695 K. All the MD (canonical ensemble, constant N, V, T) simulations are carried out with the molecular modeling package IMPACT [12,35], with the OPLSAA force field [39]. Periodic boundary conditions are used in all simulations. The long-range electrostatic interactions are calculated by the P3ME method, with a mesh size of 36 × 36 × 36 (grid spacing about 1.0 Å). A time step of 4.0 fs (outer timestep) is used for all temperatures using the RESPA [15,35] algorithm. A standard equilibration protocol is used for each replica. It starts with a conjugate gradient minimization for each solvated system. Then a two-stage equilibration, each consisting of a 100 ps MD, is followed. In the first stage, the β-hairpin is frozen in space, so only the solvent molecules are equilibrated; and in the second stage, all atoms are equilibrated. The final configurations of the above equilibrations are then used as the starting points for the 64 replicas. Each replica is run for 3.0 ns for data collection, with replica exchanges attempted every 0.4 ps. Protein configurations were saved every 0.08 ps, giving a total of 2.4 million configurations. The aggregate MD integration time of all replicas is 0.192 μs.

The Trp-cage structure under study is taken from the NMR structure (PDB 1l2y.pdb). The 20 residue mini-protein (structure 1 out of the 38 NMR structures) is then solvated in a 50 Å × 50 Å × 50 Å water box using the SPC model with a density of 1.0 g/cm^3. This results in a total of 12,242 atoms for each replica, with 305 protein atoms and one Cl-counter ion to neutralize the solvated protein system. The mesh size of P3ME is set to 50 × 50 × 50 (grid spacing 1.0 Å). A time step of 4.0 fs (outer time step) is again used at each temperature. A total of 50 replicas are simulated with temperatures ranging from 282 to 598 K. A similar equilibration protocol as described above is used. Each replica is run for 5.0 ns for data collection. Similarly, the replica exchanges were attempted every 0.4 ps, and the protein configurations were saved every 0.08 ps. This results in a total of 3.125 million configurations and an aggregate MD integration time 0.25 μs.

3. Results and discussion

3.1. β-Hairpin folding

The optimal temperature distributions in the replica exchange method can be obtained by running a few trial replicas with short MD simulations. The temperature gap can be easily determined by monitoring the acceptance ratio desired between neighboring temperatures. As mentioned above, we only allow exchanges between neighboring replicas. The rest of the temperature list can be easily interpolated, since the optimal temperature distribution should be roughly exponential assuming the heat capacity is relatively a constant (there might be a heat capacity spike near the melting transition temperature in protein folding, but for simplicity we assume it to be a constant). For this β-hairpin, we set an exponentially distributed temperature series of 270, 274, 278,. . . , 685, 695 K, which results in an acceptance ratio of about 30–40%. The temperature gaps between these replicas range from 4 to 10 K. The detailed temperatures for each replica and neighboring exchange acceptance ratios are listed in Table 1. As one can see, a relatively uniform acceptance ratio is obtained for this temperature series, which is desired for a smooth random walk in temperature space. The random walk in temperature space will then result in a random walk in potential space.

Table 1
Acceptance ratio for replica exchanges between neighboring replicas in REM for the solvated β-hairpin system

Temperature pairs	Acceptance ratio	Temperature pairs	Acceptance ratio
270–274	0.370	437–443	0.340
274–278	0.310	443–450	0.350
278–282	0.270	450–457	0.370
282–287	0.300	457–464	0.420
287–291	0.190	464–471	0.360
291–295	0.280	471–478	0.410
295–300	0.320	478–485	0.450
300–305	0.250	485–492	0.420
305–310	0.240	492–500	0.320
310–314	0.330	500–507	0.330
314–318	0.200	507–515	0.490
318–323	0.330	515–523	0.340
323–328	0.370	523–531	0.330
328–333	0.340	531–539	0.430
333–338	0.370	539–547	0.440
338–343	0.310	547–555	0.500
343–348	0.250	555–563	0.340
348–354	0.360	563–572	0.400
354–359	0.310	572–581	0.400
359–365	0.360	581–589	0.440
365–370	0.260	589–598	0.340
370–376	0.300	598–607	0.440
376–381	0.320	607–617	0.420
381–387	0.370	617–626	0.370
387–393	0.350	626–635	0.390
393–399	0.300	635–645	0.410
399–405	0.260	645–655	0.470
405–411	0.350	655–665	0.440
411–417	0.400	665–675	0.480
417–424	0.370	675–685	0.470
424–430	0.420	685–695	0.320
430–437	0.300	–	–

The temperature series is distributed exponentially, which gives a relatively constant acceptance ratio for all neighboring replica exchanges.

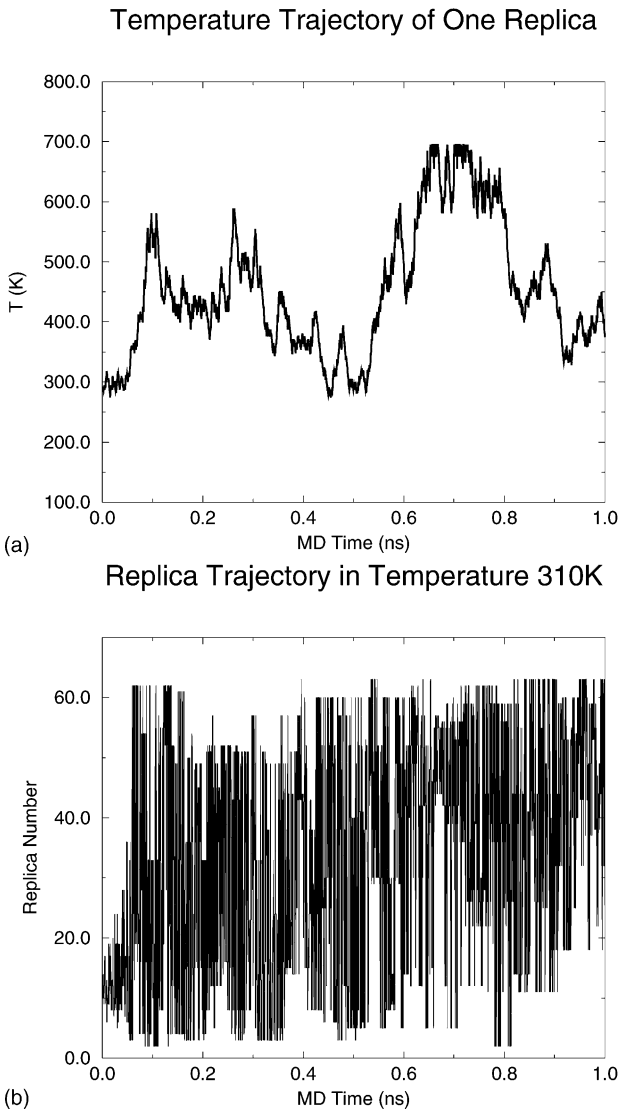

Fig. 1. The replica exchange trajectory from the β-hairpin simulation: (a) the top one is the temperature trajectory for one replica (started at 310 K); (b) the bottom one is the replica trajectory in temperature 310 K. Both show the replica exchange method is "surfing" the temperature space effectively.

Fig. 1a shows the "temperature trajectory" for one replica (replica 9), which started at 310 K. It is clear that this replica walks freely in the temperature space. A similar graph is to monitor the "replica trajectory" at one particular temperature. Fig. 1b plots the replica trajectory at temperature 310 K. It shows that various replicas visit this temperature randomly. These two plots basically show the time trajectory of the permutation function and its inverse in Eq. (3) as we discussed in the REM methodology section. The results indicate that our temperature series is reasonably optimized for this system with sufficiently high acceptance ratios for replica exchanges. In REM, this random walk in temperature space is correlated with the inter-basin jumps in the free energy landscapes.

The free energy landscape of the β-hairpin folding in water is obtained by the usual histogram analysis [23,41],

$$P(X) = \frac{1}{Z}\exp(-\beta W(X)), \tag{34}$$

and

$$W(X_2) - W(X_1) = -RT\log\left(\frac{P(X_2)}{P(X_1)}\right), \tag{35}$$

where $P(X)$ is the normalized probability obtained from a histogram analysis as a function of X. X is any set of reaction coordinates or any parameters describing the conformational space. $W(X_2) - W(X_1)$ is thus the relative free energy, or so-called potential of mean force (PMF) discussed by Garcia and Sanbonmatsu [23]. We have previously tried many reaction coordinates for this hairpin [28], including the number of β strand hydrogen bonds, hydrophobic core radius of gyration, fraction of native contacts, radius gyration of the entire peptide, RMSD from the native structure, and principal components (PC) from PCA analysis [23,42]. Here, we will use the number of β strand hydrogen bonds (N_{HB}^{β}) and the radius of gyration of the hydrophobic core ($\mathrm{Rg}^{\mathrm{core}}$) as the reaction coordinates for the free energy contour map, which turns out to be very informative for this particular peptide. N_{HB}^{β} is defined as the number of backbone-backbone hydrogen bonds excluding the two at the turn of the hairpin (five out of total seven backbone–backbone hydrogen bonds). A hydrogen bond is counted if the distance between the two heavy atoms (N and O in this case) is less than 3.5 Å and the angle N–H...O is larger than 120.0°. $\mathrm{Rg}^{\mathrm{core}}$ is the radius of gyration of the side chain atoms on the four hydrophobic residues, Trp43, Tyr45, Phe52, and Val54.

Fig. 2 shows the free energy contour map (in units of RT) and representative structures at each local free energy basin. The free energy landscape reveals a few interesting points: (1) the overall free energy contour map has an "L" shape, indicating that the folding mechanism is likely driven by the hydrophobic core collapse. If it were driven by a "hydrogen bond zipping" mechanism, the shape would have been a more "diagonal" one in the 2D graph. (2) There are four states, or local energy minima, at biological temperature: the native folded state (F), the unfolded state (U) and two intermediates, a "molten globule" state, which is similar to Pande and coworkers state H [20,21] and a partially folded state (P). (3) The intermediate state H shows a compact hydrophobic core but with no beta-strand hydrogen bonds. (4) The representative structures from the four states also indicate that the folding pathway is driven by a hydrophobic core collapse followed by a local beta-strand hydrogen bond reformation to make up some energy losses due to the breakup of protein–water hydrogen bonds. These results in general are consistent with what has been found by others [17,20,22,23,27]. However, there are also some significant differences. One important difference is in the folding mechanism. Eaton and coworkers [17,18] proposed a "hydrogen bond zipping" mechanism in which folding initiates

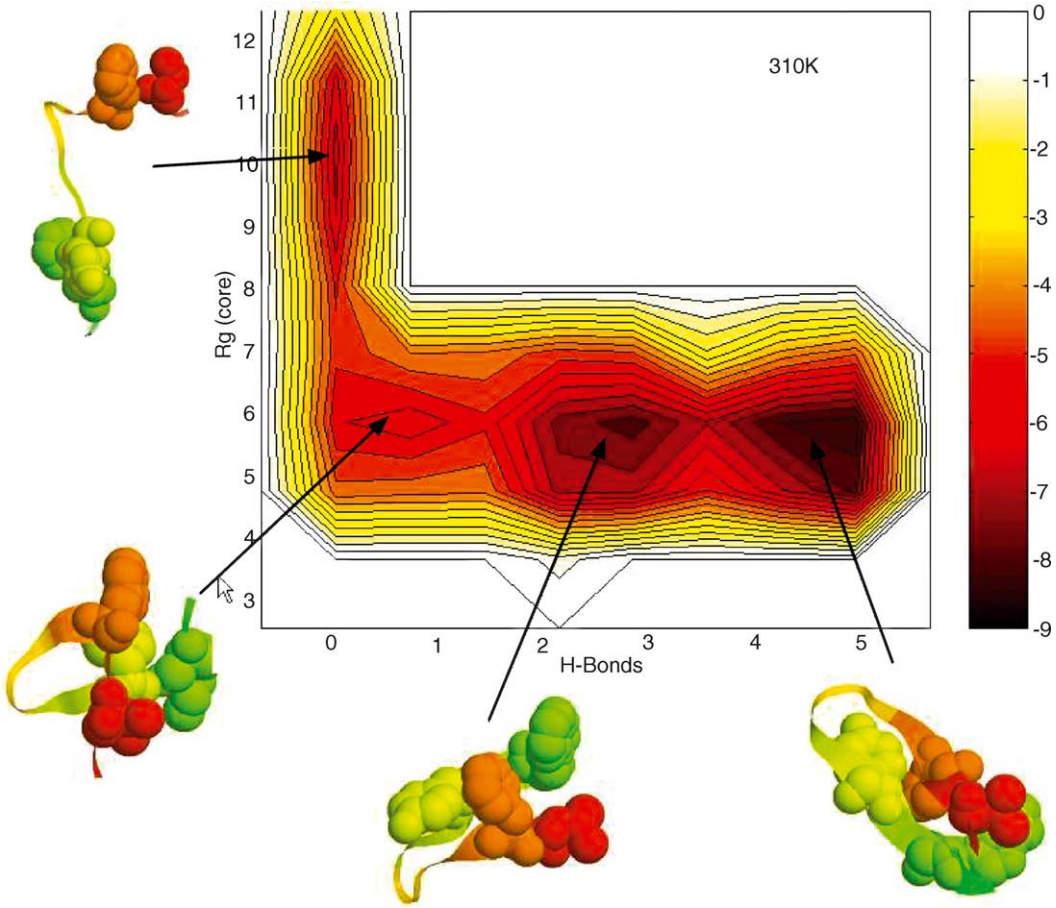

Fig. 2. Folding free energy contour map of the β-hairpin vs. the two reaction coordinates, the number of β-sheet H-bonds N_{HB}^{β} and the radius gyration of the hydrophobic core residues Rg^{core}. The contours are spaced at intervals of 0.5*RT*. The representative structures are shown for each energy state. The hydrophobic core residues Trp43, Tyr45, Phe52, and Val54 are shown in space-fill mode, while all other residues are shown as ribbons (see text for more details).

at the turn and propagates toward the tails, so that the hydrophobic clusters, from which most of the stabilization derives, form relatively late during the reaction. Our analysis shows no evidence of this hydrogen bond zipping model. Pande and Rokhsar [20] and Garcia and Sanbonmatsu [23] found similar results using the CHARMM and AMBER force fields, namely, the β-hairpin system first folds into a compact H state before it folds into the native structure. It is also found that both H and P states have a well-formed hydrophobic core, but the P state (two to three H-bonds) has a significantly higher population than the state H (zero to one H-bonds). The heavy population in the partially folded state P and a low free energy barrier from the H state to the P state (≈0.8*RT*) implies that the final β-strand hydrogen bonds could be formed nearly simultaneously with the hydrophobic core. Thus, it seems that after an initial core collapse of the peptide, it quickly adopts a partially folded state with two to three hydrogen bonds before folding into the native state.

Another interesting question regarding the folding intermediates is: "to what extent do α-helical structures form during the folding process?" Early experiments and theoretical simulations have not found significant helical content [17,19,20,22]. However, very recently, Garcia and Sanbonmatsu have found that significant helical content exists (15–20%) at low temperatures (282 K) using the AMBER94 force field. These conformations are only slightly unfavorable energetically with respect to hairpin formation near biological temperatures. Pande and coworkers also found significant helical intermediates at 300 K from their continuum solvation simulations with an old version of the OPLS united atom force field [21]. It is of interest to see if this remains the case for the all-atom OPLSAA model with an explicit solvent and a more rigorous treatment of the long-range electrostatic interactions. The number of residues in the beta-strands and helices are calculated with program STRIDE [43], which uses both hydrogen bond energies and dihedral angles when assigning secondary structure to each residue in the sequence [43]. Fig. 3 shows the number of residues in the beta sheet (N_β) and the alpha-helix (N_α, including both alpha-helix and 3_{10}-helix) at various temperatures. As one can see, the number of helical residues is typically less than or equal to three, and more importantly, only one to two percent of the conformations

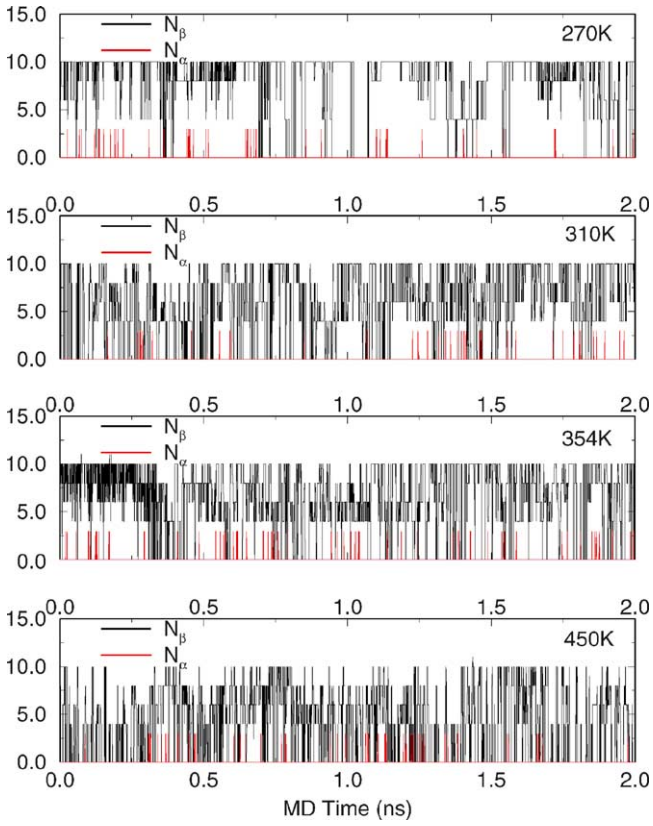

Fig. 3. Number of residues in helix and beta sheet format for the β-hairpin at various temperatures as determined by the program STRIDE [43]. It is obvious that no significant helix content is found in our simulation.

show any helical content for all of the temperatures examined. Furthermore, almost all of the helices found are 3_{10}-helices near the original beta-turn (residues (47–49)). This is in contrast to what had been found previously by others [23]. However, there is evidence showing that for AMBER simulations the significant α-helical content might be due to the artifacts in AMBER force fields (AMBER94, AMBER99) which favor the α-helices [44–46].

The simulation also reveals the folding transition temperature, hydrogen bond distributions at various temperatures, temperature dependence of the native hairpin population, and NOE distance constraints, etc. [28]. These results are reported previously, so we are not including them here anymore. As mentioned earlier, this β-hairpin system is mainly used to demonstrate the power of the methodology.

3.2. Trp-cage folding

Fig. 4 shows the free energy contour map and representative structures for the Trp-cage The two reaction coordinates used are the fraction of native contacts and the radius of gyration of the entire protein. Unfortunately, for this Trp-cage it is not obvious to identify some unique and informative reaction coordinates, like the two used for the β-hairpin, thus, two more general reaction coordinates are adopted, instead. Searching for better reaction coordinates is still of great interest in protein folding studies. In this study, a native contact is defined as a C_α–C_α distance less than 6.5 Å for non-adjacent residues, and the radius of gyration is based on all heavy atoms with unit mass. Similarly, the free energy contour maps reveal several interesting features for this Trp-cage folding. (1) The folding free energy landscapes are in general very smooth, even smoother than that of the β-hairpin. The landscape becomes completely smooth and funnel-like [47,48] above 340 K (for the β-hairpin it is about 360 K), indicating a stable two-state like folding at higher temperatures. (2) At low temperatures, however, such as 300 K or lower, there exists an intermediate state “I”, or a molten globule state, near reaction coordinates (9.4, 0.42), showing about 15% population at 300 K. The intermediate state structures are found to have two partially formed hydrophobic cores separated by a salt-bridge between residues Asp9 and Arg16 near the center of the molecule (more discussions below). The free energy barrier between the intermediate state “I” and the folded state “F” is low though, for example, at 300 K the free energy barrier is only about 1.2*RT*, indicating that it is easy for the peptide to cross over the barrier by thermal fluctuations. (3) The folded state “F” has a much lower free energy than the intermediate state, which means that it dominates the population at equilibrium, again it is estimated to be about 70% at 300 K. The lowest free energy structure agrees well with the native NMR structure, with only a 1.50 Å C_α-RMSD from the native structure.

These results agree well with the experimental results [30,31], which also show a stable two-state folder with a very high native state population. The low free energy barrier between the intermediate state “I” and the folded state “F” at 300 K indicates that the intermediate state structures might be short lived. The native population at 300 K from our simulation is somewhat lower than that of the fluorescence or chemical shift deviation experiment. This might be related to the fact that we estimate the population with a tougher criterion, i.e., each native contact has to be within 6.5 Å to be counted in the simulated ensemble conformations, while the Trp fluorescence experiment (using residue Trp6) might collect the fluorescence signal if Trp6 is buried but not perfectly folded with other residues. In other words, partially folded structures with Trp6 buried might exhibit the Trp fluorescence as well. The CSD experiment, on the other hand, recalibrates the signal to be “100% folded” for the largest values observed for the “C-cap” and “cage formation” measures [30]. Thus, our lower population at low temperatures might be reasonable. Similar results were also found in the β-hairpin population estimation. The fraction of native contacts shows a 72% β-hairpin population while fluorescence experiments show about 80% population at 282 K.

It is worth taking a closer look at the structures from the intermediate state for a better understanding of the folding mechanism. The representative structure shown in Fig. 4 is the most popular structure from the intermediate state at 300 K. The representative structure is selected from the

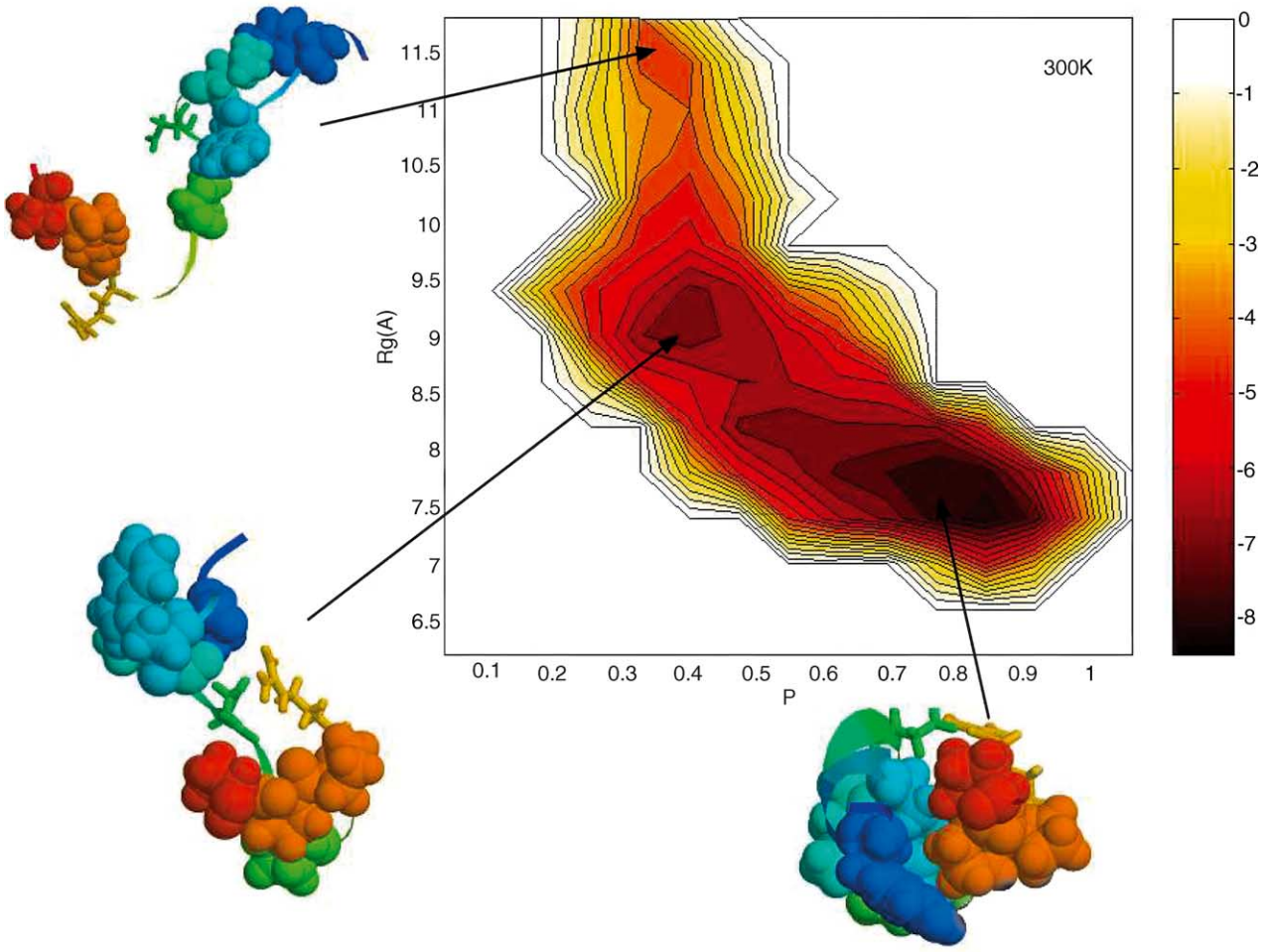

Fig. 4. Folding free energy contour map of Trp-cage vs. the fraction of native contacts ρ and the radius gyration of the entire protein Rg. The contours are spaced at intervals of 0.5*RT*. The representative structures are shown for each energy state. The key hydrophobic residues forming the Trp-cage core, Tyr3, Trp6, Leu7, Pro12, Pro17, Pro18, and Pro19 are represented in the space-fill mode, and the two charged residues Asp9 and Arg16 are represented by sticks (see text for more details).

clustering analysis [28], which allows us to determine the unique structures in a free energy basin and also the populations in each cluster. The key hydrophobic residues forming the Trp-cage core, Tyr3, Trp6, Leu7, Pro12, Pro17, Pro18, and Pro19 are represented in the space-fill mode, and the two charged residues Asp9 and Arg16 are represented by sticks. The intermediate state structure shows two partially formed hydrophobic cores, one by residues Tyr3, Trp6, and Leu7, and the other by the four Pro residues. The two charged residues Asp9 and Arg16 form a salt-bridge, which is located near the center of the structure. For comparison, the native structure also shows a salt-bridge between Asp9 and Arg16, but it is formed outside the central hydrophobic core region and located on the molecular surface, i.e., it is exposed to the solvent.

This prematurely formed salt-bridge near the center of the molecule creates a metastable state, the intermediate state "I", since it takes energy to break this salt-bridge to make the final hydrophobic core. Breaking a salt-bridge might take up to 3–4 kcal/mol free energy for buried ones [49] and about 1.0 kcal/mol for surface ones [50] (smaller than the buried ones, since strong hydrogen bonds with water can make up some differences). In this case, the salt-bridge between Asp9 and Arg16 probably falls in-between these two categories, so a free energy loss of about one to two kilocalories per molecule might be expected for breaking this prematurely formed salt-bridge. The overall free energy barrier of 1.2*RT* ($\approx$0.8 kcal/mol) from the intermediate state to the native state in Fig. 4 is in line with this analysis. Thus, the folding process seems to involve an intermediate state where the peptide quickly forms two partial hydrophobic cores separated by a salt-bridge between residues Asp9 and Arg16. The two partially formed hydrophobic cores then collapse into a larger final one, and the salt-bridge reforms on the molecular surface to further stabilize the protein system. It will be interesting to see exactly how the intermediate state structure folds into the final native structure. Preliminary 10–30 ns kinetics runs starting from the intermediate state show that the structure can stay with the salt-bridge formed and open and reformed in the entire trajectory. Very

few runs show the salt-bridge broken completely. We are currently investigating this and the results will be reported elsewhere [51]. To our knowledge, this is the first intermediate state identified which might provide an explanation to the super fast folding rate of this protein, since it is easier to make a correct packing for each partial core or subunit, in this case, the Try3, Trp6, Leu7 unit and the four Pro residues unit. This is seen in the intermediate state structure in Fig. 4, where the correct hydrophobic packing for each subunit are compared with the final structure. Thus, a two-step folding mechanism emerges: first, the peptide is separated into two regions for easier partial hydrophobic core packing by forming a meta-stable salt-bridge near the center; and second, the two correctly packed partial cores are assembled into final larger one. More experiments might be helpful here to study the intermediates and their structural and dynamical properties. In addition, no meaningful α-helix is found in the intermediate state, which indicates that the α-helix is formed at the late stage with the Trp-cage core.

Fig. 5 shows the comparison of the lowest free energy structure from the free energy landscape in Fig. 4 and the native NMR structure (structure 1 of the 38 NMR structures). The NMR structure (Fig. 5b) shows an α-helix at residues 2–9, and a 3_{10}-helix at residues [11–14]. It also reveals a compact hydrophobic core where four proline residues (Pro12, Pro17, Pro18, and Pro19) and a leucine (Leu7) pack against the aromatic side chains of Try3 and Trp6. The lowest free energy structure (Fig. 5a) shows a 1.50 Å C_α-RMSD from the native structure, with the major deviations from residues N1, G10 and S20, which all show a C_α-RMSD larger than 2.2 Å. If one ignores the two terminal residues, which are poorly defined in NMR anyway, the C_α-RMSD is reduced to 1.31 Å. The noticeable differences between our lowest free energy structure and the NMR structure include:

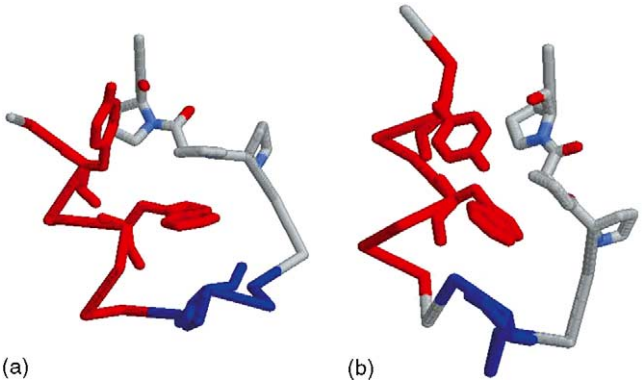

Fig. 5. Comparison of the lowest free energy structure (a, left) and the native NMR structure (b, right) for Trp-cage. The key hydrophobic residues packing against the central Trp6 residue (Tyr 3, Trp6, Leu7, Pro 12, Pro 17, Pro18, and Pro19) are shown with sidechains and all other residues are shown only backbones. The lowest free energy structure shows an C_α-RMSD of 1.50 Å from the native structure, with the major differences in 3_{10}-helix region (residues 11–14) and residue Tyr3, where the phenyl ring is not as closely packed to the Trp6 as the native structure.

(1) the 3_{10}-helix in residues [11–14] is no longer apparent in the simulated structure, but instead, it is classified as beta-turn residues by the STRIDE program [43]; (2) the side chain (phenyl ring) in residue Try3 of the lowest free energy structure is not as closely packed to the central Trp6 as in the native structure. Instead, it extends more into the solvent to fully expose the hydroxyl (–OH) group. Interestingly, this is also seen in the best structure from the stimulation of Pande et al. (Fig. 2B in ref. [6]) with a united atom OPLS force field and a continuum solvent GBSA model. Simmerling et al. [32] first simulated this Trp-cage system using a modified AMBER99 force field with the GBSA model, and Pitera and Swope repeated some of the calculations using AMBER99 force field and the same GBSA continuum solvent model. Both of them found remarkably low RMSD (<1 Å C_α-RMSD) structures from their simulations. While this 1.50 Å C_α-RMSD seems slightly worse than the best structure from Simmerling et al. and Pitera et al.'s simulations, it is noticeably better than the best structure obtained by Pande et al. using mean structure analysis. The best structure of Pande et al. shows a larger than 2.0 Å C_α-RMSD, and most importantly, the central Trp6 residue seems not well packed within the "cage", but it instead sticks out from the "cage" (Fig. 2B in [33]). On the other hand, at least one reason why the best structures from AMBER simulations [32,34] have a lower RMSD than ours might be because the final NMR structures are minimized with the AMBER force field [30]. Furthermore, these best structures are either from lowest potential energies or from the lowest C_α-RMSDs in the entire ensemble, they are not necessarily the lowest free energy structures; while the structure we reported here is from the lowest free energy basin in the free energy landscape. In general, the potential energy profiles from AMBER (and other force fields as well) also degenerate with respect to RMSD, in other words, very different structures can share the same low potential energy. Therefore, it is usually difficult to select the "best" structure based just on the potential energy, for example, in the simulation of Simmerling et al. [32], there are also many structures having an backbone RMSD greater than 4 Å but with potential energies comparable with the lowest one (Fig. 1 in [32]). In principal, the lowest free energy should be used instead of the lowest potential energy for picking the best structure.

4. Conclusion

A replica exchange method combined with a newly developed molecular dynamics algorithm P3ME/RESPA has been proposed for efficient sampling of protein folding conformation space. The replica exchange method uses a velocity rescaling scheme to couple molecular dynamics trajectories with a temperature exchange Monte Carlo process. The P3ME/RESPA algorithm optimally combines the P3ME method for the long range electrostatic interactions and the RESPA algorithm for multiple time steps

in MD simulations. The final combined method has been applied to two separate protein systems, β-hairpin and a designed Trp-cage. Both simulations are performed with an explicit solvent model and a periodic boundary condition. The OPLSAA force field and SPC water model are used for simulation. A total of 64 and 50 replicas are simulated in parallel over a wide range of temperatures for the β-hairpin and Trp-cage, respectively. The main conclusions from folding simulations are summarized in the following.

For the β-hairpin, an "L" shaped free energy contour map versus the number of beta-strand hydrogen bonds and the radius gyration of hydrophobic core has been found, which indicates that the folding mechanism of this β-hairpin is mainly driven by hydrophobic core collapse. This is different from the "hydrogen bond zipping" mechanism proposed by some early experiments and simulations [17,27]. The low free energy barrier between the H state (hydrophobic core formed but with no beta-strand hydrogen bonds) and partially folded P state also indicates that the final beta-strand hydrogen bonds could be formed nearly simultaneously with the hydrophobic core. In contrast to some recent simulations [21,23], no meaningful helical content has been found in our simulation at any temperature, which appears to agree with NMR experiments better. We speculate that the alpha-helix content found in these simulations are probably due to the artifacts of some force fields.

For the Trp-cage folding, a new intermediate state has been identified from the free energy contour map versus the fraction of native contacts and the radius of gyration. At room temperature 300 K, the Trp-cage quickly undergoes an intermediate state which has two correctly packed partial cores separated by an essential salt-bridge between residues Asp9 and Arg16 near the center of the molecule. The free energy barrier to break this prematurely formed salt-bridge makes it a meta-stable state, which provides an explanation to the super fast folding rate for this mini-protein, since it is easier to pack partial hydrophobic cores in each subunit. Thus we propose a following two-step folding mechanism: first the peptide is separated into two regions by forming a meta-stable salt-bridge near the center; then the two correctly pre-packed hydrophobic cores are assembled into the final larger core and also reform the salt-bridge on the molecular surface to gain further stability. The lowest free energy structure is found to show only a 1.50 Å C_α-RMSD from the NMR structure at 300 K. No meaningful α-helix is found in the intermediate state, which indicates that the α-helix is formed in the final stage along with the Trp-cage core.

Acknowledgements

I would like to thank Bruce Berne, Jed Pitera, Bill Swope, Angel Garcia, and Carlos Simmerling for many helpful discussions and comments.

References

[1] B.J. Berne, J.E. Straub, Curr. Top. Struct. Biol. 7 (2) (1997) 181.
[2] R.H. Swendsen, J.S. Wang, Phys. Rev. Lett. 58 (1987) 86.
[3] A. Nayeem, J. Vila, H.A. Scheraga, J. Comp. Chem. 12 (1991) 594.
[4] D.L. Freeman, D.D. Frantz, J.D. Doll, J. Chem. Phys. 97 (1992) 5713.
[5] E. Marinari, G. Parisi, Europhys. Lett. 19 (1992) 451.
[6] K. Hukushima, K. Nemoto, J. Phys. Soc. Jpn. 65 (1996) 1604.
[7] D.B. Faken, A.F. Voter, D.L. Freeman, J.D. Doll, J. Phys. Chem. 103 (1999) 9521.
[8] C.M. Dobson, A. Sali, M. Karplus, Angrew Chem. Int. Ed. Engl. 37 (1998) 868.
[9] C.L. Brooks, M. Gruebele, J.N. Onuchic, P.G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. 95 (1998) 11037.
[10] Y. Duan, P.A. Kollman, Science 282 (1998) 740.
[11] E. Marinari, G. Parisi, J. Ruiz-Lorenzo, World Scientific, Singapore, 1998, p. 59.
[12] R. Zhou, B.J. Berne, J. Chem. Phys. 107 (1997) 9185.
[13] Y. Sugita, Y. Okamoto, Chem. Phys. Lett. 329 (2000) 261.
[14] R. Zhou, E. Harder, H. Xu, B.J. Berne, J. Chem. Phys. 115 (2001) 2348.
[15] M. Tuckerman, B.J. Berne, G.J. Martyna, J. Chem. Phys. 97 (1992) 1990.
[16] R.W. Hockney, J.W. Eastwood, Computer Simulations Using Particles, IOP, Bristol, 1988.
[17] V. Munoz, P.A. Thompson, J. Hofrichter, W.A. Eaton, Nature 390 (1997) 196.
[18] V. Munoz, E.R. Henry, J. Hofrichter, W.A. Eaton, Proc. Natl. Acad. Sci. U.S.A. 95 (1998) 5872.
[19] F.J. Blanco, G. Rivas, L. Serrano, Nat. Struct. Biol. 1 (1994) 584.
[20] V.S. Pande, D.S. Rokhsar, Proc. Natl. Acad. Sci. U.S.A. 96 (1999) 9062.
[21] B. Zagrovic, E.J. Sorin, V.S. Pande, J. Mol. Biol. 313 (2001) 151.
[22] A.R. Dinner, T. Lazaridis, M. Karplus, Proc. Natl. Acad. Sci. U.S.A. 96 (1999) 9068.
[23] A.E. Garcia, K.Y. Sanbonmatsu, Proteins 42 (2001) 345.
[24] D. Roccatano, A. Amadei, A.Di Nola, H.J. Berendsen, Protein Sci. 10 (1999) 2130.
[25] A. Kolinski, B. Ilkowski, J. Skolnick, Biophys. J. 77 (1999) 2942.
[26] B. Ma, R. Nussinov, J. Mol. Biol. 296 (2000) 1091.
[27] D.K. Klimov, D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. 97 (2000) 2544.
[28] R. Zhou, B.J. Berne, R. Germain, Proc. Natl. Acad. Sci. U.S.A. 98 (2001) 14931.
[29] R. Zhou, B.J. Berne, Proc. Natl. Acad. Sci. U.S.A. 99 (2002) 12777.
[30] J.W. Neidigh, R.M. Fesinmeyer, N.H. Andersen, Nat. Struct. Biol. 9 (2002) 425.
[31] L. Qiu, S.A. Pabit, A.E. Roitberg, S.J. Hagen, J. Am. Chem. Soc. 124 (2002) 12952.
[32] C. Simmerling, B. Strockbine, A.E. Roitberg, J. Am. Chem. Soc. 124 (2002) 11258.
[33] C.D. Snow, B. Zargovic, V.S. Pande, J. Am. Chem. Soc. 124 (2002) 14548.
[34] J. Pitera, W. Swope, Proc. Natl. Sci. Acad. U.S.A. 100 (2003) 7587.
[35] R. Zhou, B.J. Berne, J. Chem. Phys. 103 (1995) 9444.
[36] A. Tom, M. Darrin, G. Pedersen, J. Chem. Phys. 98 (1993) 10089.
[37] M. Deserno, C. Holm, J. Chem. Phys. 109 (1998) 7678.
[38] P. Procacci, M. Marchi, J. Chem. Phys. 104 (1996) 3003.
[39] W.L. Jorgensen, D. Maxwell, J. Tirado-Rives, J. Am. Chem. Soc. 118 (1996) 11225.
[40] H.C. Andersen, J. Comp. Phys. 52 (1983) 24.
[41] A.M. Ferrenberg, R.H. Swendsen, Phys. Rev. Lett. 63 (1989) 1195.
[42] A.E. Garcia, Phys. Rev. Lett. 68 (1992) 2696.
[43] D. Frishman, P. Argos, Proteins 23 (1995) 566.
[44] M. Beachy, D. Chasman, R. Murphy, T. Halgren, R. Friesner, J. Am. Chem. Soc. 119 (1997) 5908.

[45] A.E. Garcia, K.Y. Sanbonmatsu, Proc. Natl. Acad. Sci. U.S.A. 99 (2002) 2782.
[46] R. Zhou, Proteins 53 (2003) 148.
[47] P.G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. 94 (1997) 6170.
[48] K.A. Dill, H.S. Chan, Nat. Struct. Biol. 4 (1997) 10.
[49] C. Waldburger, J. Schildbach, R. Sauer, Nat. Struct. Biol. (1995) 2.
[50] A. Horovitz, L. Serrano, B. Avron, M. Bycroft, A. Fersht, J. Mol. Biol. (1990) 216.
[51] R. Zhou, Proc. Natl. Sci. Acad. U.S.A. 100 (2003) 13280.