

# Balancing Simulation Accuracy and Efficiency with the Amber United Atom Force Field

*Meng-Juei Hsieh and Ray Luo\**

Department of Molecular Biology and Biochemistry  
University of California, Irvine, California, USA

\*E-mail: rluo@uci.edu

## Abstract

We have analyzed the quality of a recently proposed Amber united-atom model and its overall efficiency in *ab initio* folding of two stable  $\beta$ -hairpins. It is found that the simulated native  $\beta$ -hairpin stabilities are similar to those derived from measured chemical-shift deviations. Detailed analysis of simulated conformations shows that the majority of the sampled conformations are with high fractions of native  $\beta$  and native tertiary contacts. Given the reasonable quality of the united-atom model with respect to experimental data, we have further studied the simulation efficiency of the united-atom model over its corresponding all-atom model in Amber. Our data shows that the united-atom model is a factor of six to eight faster than the all-atom model as measured with the *ab initio* first pass folding time for the two tested  $\beta$ -hairpins. Detailed structural analysis shows that all *ab initio* folded trajectories enter the native basin whether the united-atom model or the all-atom model is used. Finally, we have also studied the simulation efficiency of the united-atom model as measured in term of how fast thermodynamic convergence can be achieved. It is apparent that the united-atom simulations reach convergence faster than the all-atom simulations with respect to both potential energy and backbone structural distributions. These findings show that the efficiency of the united-atom model is clearly beyond the per-step dynamics simulation of about two over the all-atom model. Thus reasonable reduction of protein model can be achieved with improved sampling efficiency

while still preserve a high level of accuracy for *ab initio* protein folding simulations. This study motivates us to develop more simplified protein models with sufficient consistency with the all-atom models for enhanced conformational sampling.

## 1. Introduction

*Ab initio* protein folding has remained to be a challenging problem in biophysical chemistry due to the enormous sampling space and the atomistic resolution that are needed for consistent accuracy in theoretical prediction. The sampling space of typical protein domains (a.k.a. ~200 residues) is exponentially large. If each residue can only adopt two possible conformations, a protein domain of 200 residues can adopt a total of  $1.6 \times 10^{60}$  possible conformations, which cannot be adequately sampled at the atomistic resolution within reasonable time. Quite often accuracy at the atomistic resolution is a must for further functional analyses of proteins. Addressing the challenges in sampling and accuracy requires significant computational resources. Increasing computing power has made it possible to simulate *ab initio* folding of small proteins,<sup>1</sup> but it has remained impractical for typical protein domains. The computational difficulties in *ab initio* protein folding can be roughly grouped into two categories: overcoming enthalpic barriers, i.e. how to escape from local minima; and overcoming entropic barriers, i.e. how to sample the exponentially large conformational space with reasonable scalability.

Most previous developmental efforts have been focused on crossing enthalpic barriers. Monte Carlo based methods, including simulated annealing,<sup>2</sup> and/or genetic algorithms,<sup>3</sup> are widely used.<sup>4-12</sup> This is across different protein models, either knowledge-based or physics-based potentials, though its application to all-atom potentials based on molecular mechanics force fields is challenging.

Potential energy surface smoothing is also widely used. Various methods such as the diffusion equation method,<sup>13</sup> Gaussian “coarse-graining”,<sup>14</sup> and packet annealing<sup>15,16</sup> have been proposed. Smoothing potential energy surface in principle reduces the number of minima, thus reduces the number of energy barriers. With the hope that the global minimum of the deformed potential energy surface shares a similar topology with the global minimum of the original surface,<sup>14</sup> the resulting global minimum is then mapped back to the original surface by a reversing procedure. However, the reversing procedure has been a major challenge in these approaches.<sup>17</sup>

Generalized-ensemble methods such as multi-canonical algorithm, simulated tempering, and replica-exchange methods have recently emerged to be effective conformational sampling methods.<sup>18-45</sup> Their applications to *ab initio* protein folding simulation have also been reported. Multi-canonical algorithm relies on the idea of artificially eliminating energy barriers, thereby circumventing the problems associated with the traditional sampling methods. In simulated tempering, temperature becomes a dynamical variable that takes values ranging over a definite set. In this method, it is possible to utilize the fact that at higher temperatures potential energy barriers are effectively lower. Many other methods have been proposed to enhance conformational sampling for systems of high dimensionality.

It seems that overcoming entropic barriers is much harder without sacrificing the resolution of protein models. Previous approaches can be roughly grouped into two categories. The first strategy utilizes discrete conformational space, such as the use of lattice models<sup>9,46,47</sup> or discrete torsion-angle values in Monte Carlo sampling<sup>7</sup>, to overcome entropic barriers. The second strategy aims at reducing the dimensionality of sampling space: this leads to reduced protein models at residue-level representations. The applications of reduced protein models to *ab initio*

protein structure predictions are very encouraging;<sup>8,10</sup> These models have also been used to understand the physical principles in protein folding.<sup>48-56</sup> The reduced models can use only one particle per residue with either protein-specific Gō potentials derived from specific protein structures or statistical potentials derived from a database of high-quality native structures.<sup>57,58</sup> More complex reduced protein models represent each residue by more than one particle.<sup>59-64</sup> Compared with the minimal models with one particle per residue, these reduced models improve the predictive power with a better description of the side-chain geometry and interactions. These models also perform gracefully well when combined with post-refinement using other atomistic models<sup>65-67</sup> in *ab initio* protein folding simulations. Nevertheless, use of reduced models at lower resolutions apparently results in low-resolution theoretical predictions. Similarly use of discrete space results in limited coverage of sampling space that in turn also leads to low-resolution predictions. Given these limitations, new strategies such as resolution or Hamiltonian exchange<sup>38-45</sup> can be utilized to couple high-resolution all-atom models for accuracy and low-resolution reduced models to enhancing sampling efficiency without sacrificing the much needed all-atom accuracy during simulations.

The reduction of model complexity can also be achieved by using implicit solvents because the bottleneck of all-atom simulations lies in the expensive solvent-solvent interactions, though there are still issues remaining in their development, especially when they are compared with explicit solvents.<sup>68,69</sup> With the growing popularity of implicit solvents in protein simulations, the advantages of united-atom models become apparent because these models can be used to achieve modest reduction of complexity without much sacrifice of atomistic accuracy. Indeed, it is subject to debate whether it is necessary to represent hydrogen atoms in atomistic simulations when we have given up explicit representation of solvent molecules. The idea of using united-

atom models for efficient simulations goes back to the 1970's when Dunfield *et al* developed the UNICEPP force field.<sup>70</sup> In UNICEPP nonpolar hydrogens are not represented explicitly, but are included implicitly by representing nonpolar carbons and their bonded hydrogens as single particles.<sup>70</sup> Compared with all-atom models, the advantages in using united-atom models are apparent even if only raw efficiency gain in simulations is considered. First they can significantly reduce the size of most problems, since roughly half of the atoms in biological or other organic macromolecules are hydrogens. Thus there are fewer nonbonded interactions and internal degrees of freedom in united-atom models. Second, larger dynamics integration step sizes can be used by not including hydrogens since their small mass requires a smaller time step for accurate integration. However, an often overlooked and more important advantage in adopting united-atom models is the efficiency gain in conformational sampling. The simpler energy model also reduces the noise of potential energy ( $E_p$ ) landscape, for example the room-temperature  $\sigma E_p/E_p$  for the two tested systems to be described in Methods is 3.58% in the united-atom model versus 6.57% for the all-atom model. With fewer degrees of freedom ( $N$ ), the total energy ( $E$ ) fluctuation is also higher, since  $\sigma E/E$  is in the order of  $N^{-1/2}$ .<sup>71</sup> This may provide just enough fluctuation to overcome potential energy barriers between local minima, which is an added benefit for conformational sampling in molecular dynamics-based methods. It should be pointed out that these advantages become less apparent when the systems are solvated in explicit solvent.

Earlier comparisons between the all-atom and the united-atom simulations show that, the united-atom model is a satisfactory representation of internal vibrations and bulk properties of small molecules and short peptides.<sup>70,72</sup> However, limitations were also revealed in previous studies:<sup>72</sup> 1) explicit representation of hydrogens was found to be necessary for accurate

treatment of hydrogen bonding; 2)  $\pi$ -stacking could not be represented without including hydrogens in aromatic groups explicitly; 3) dipole and quadrupole moments were found inaccurate when uniting hydrogens with polar heavy atoms. New approaches were found to overcome the limitations of united-atom models. For example, only aliphatic hydrogens, which are not significantly charged and do not participate in hydrogen bonds, are represented as united-atoms while other hydrogens are represented explicitly. In this way, the limitations of the united-atom model are partially mitigated while preserving most of the benefits of the united-atom model. Of course, larger dynamics time step can no longer be used due to the use of polar and aromatic hydrogens. However, with increasing computing power, a factor of about two saving from using a larger time step becomes less important.

Recently a new united-atom force field, termed *ff03ua* in the Amber package,<sup>73,74</sup> was reported to achieve a high-level agreement with its all-atom counterpart *ff03*<sup>75</sup> in both structures and dynamics for tested proteins. A major difference of *ff03ua* with earlier united-atom models lies in its all-atom representation of protein main chain. In addition its parameter development was tightly coupled with that of its all-atom counterpart *ff03*. The previous study showed that the new united-atom model was more efficient than the Duan et al. all-atom force field for the tested system of ALA18 in the distance-dependent dielectric, which is known to fold into a stable helical structure.<sup>73</sup> In this study, we have investigated the overall simulation efficiency of *ff03ua* for *ab initio* folding simulations of the more challenging  $\beta$ -hairpins and in the more realistic generalized-Born implicit solvent treatment.

## **2. Methods**

### **2.1 Model Molecules**

To examine the overall sampling efficiency of the new united-atom (UA) model over the all-atom (AA) model, a set of model molecules was used as benchmarks. Due to the number of independent trajectories needed for a meaningful comparison and our limited computing resources, we chose two short but stable  $\beta$ -hairpin peptides designed for protein folding studies. These are HP5w4 and HP5F.<sup>76</sup> Their sequences are shown in Table 1. Both the UA and the AA models for the two  $\beta$ -hairpins were built with the LEaP program in Amber 9.<sup>74</sup> All initial structures were relaxed with a brief steepest descent minimization of 1000 steps.

(Table 1)

## 2.2 Simulation Methods

The generalized-Born implicit solvent<sup>77</sup> with the default options in Amber was used to treat polar solvation. Our previous analysis of solvent models shows that the classical SA term of nonpolar solvation overstabilizes the hairpin in both GB and PB solvents,<sup>78</sup> thus the nonpolar solvent accessible surface area term was turned off. Also it is highly inaccurate when compared with the TIP3P solvent.<sup>69</sup> All non-bonded interactions were computed without cutoff. SHAKE<sup>79</sup> was used to constrain bonds containing hydrogen atoms. The SANDER program in Amber 9 was used to perform Langevin dynamics simulation at 300K with a friction constant of  $\gamma = 1 \text{ ps}^{-1}$ .

Three different sets of trajectories were collected. The first set, termed *native* set, consists of 10 independent trajectories (with different random seeds for initial velocity assignments) starting from the native conformation for 50 ns Langevin dynamics at 300K. The *native* set is used as reference for following *ab initio* simulations. The second set, termed *ab initio* set, consists of 20 independent trajectories with randomized initial structures and different random seeds for up to 1,000 ns Langevin dynamics. The randomized initial structures for the *ab initio* set were generated from the saved snapshots (every 5 ps) of a high-temperature Langevin dynamics

simulation at 600K from the linear all-trans conformation. The third set, termed *convergence* set and used to check thermodynamic convergence, consists of 20 independent trajectories for 60 ns Langevin dynamics. The randomized initial structures for the *convergence* set were generated from the saved snapshots (every 5 ps) of a high-temperature Langevin dynamics simulation at 600K from the native conformation.

### 2.3 Native State Analysis

The quality of the UA model was studied by comparing the populations of folded peptides at 300K from both simulation and experiment. Simulated chemical shift deviation (CSD) was used to estimate the populations of folded peptides as in the NMR experiment for the two designed stable  $\beta$ -hairpins.<sup>76</sup> In NMR the populations of folded peptides were assumed to be 100% at 280K.<sup>76</sup> Thus to compute the population of folded peptide for a given system at any higher temperature, two CSDs are needed, one at 280K and one at the higher temperature (300K). The ratio of the two CSDs is used to estimate the population of folded peptide.<sup>76</sup> Here only residues 2 to 15 were used in the computation of chemical shifts.<sup>76</sup> In this study, chemical shifts of all trajectories were computed with the SHIFTS<sup>80</sup> program by Case and co-workers (<http://www.scripps.edu/mb/case/shifts.html>). The CSDb database by Andersen and co-workers (<http://andersenlab.chem.washington.edu/CSDb/>) was then used to estimate CSD as in the NMR experiment.<sup>81</sup>

Thus to compute CSD<sup>81</sup>, a fourth set of simulations at 280K is needed. Here 10 independent trajectories of 20 ns each from the native conformation were simulated. It was found that 10 ns per trajectory is enough to observe converged CSD values for the simulated peptides in both UA and AA models (see Figure 1), so that the second 10 ns per trajectory, a total of 100 ns, were used to compute the reference CSD values at 280K.

(Figure 1)

## 2.4 *Ab Initio* Folding Analysis

The first-pass folding times were used to quantify the sampling efficiency of UA and AA models. The first-pass folding time is defined as the simulation time needed to reach the native basin. Three criteria are used to detect whether a trajectory reaches the native basin: 1) the running average potential energy should reach within one standard deviation from the mean potential energy of the reference *native* set; 2) the backbone RMSD from the mean structure of the reference *native* set is less than 1Å; and 3) the peptide has to stay at the native basin so defined in 1) and 2) for at least 5 ns.

## 2.5 Convergence Analysis

The sampling efficiency of the UA model can also be measured in term of how fast thermodynamic convergence can be achieved. Here potential energy distributions and the structural distributions of backbone RMSD from different simulation sets are analyzed to see whether, after long simulation time, they can converge to the same distribution. We have compared the potential energy and structural distributions of the last 10 ns from the *ab initio* set with those of the last 10ns from the *convergence* set. The structural reference for measuring the RMSD is the average structure from the last 10 ns frames of *ab initio* trajectories.

## 3. Results and Discussion

The first goal of this project is to see if the UA model is good enough by comparing the experimental data with the simulation data. Second, we want to know whether the UA model is more efficient than the AA model under identical simulation conditions. We answered the question by kinetic analysis and then by thermodynamic analysis.

### 3.1 Quality of United-Atom Model

We first compared simulated percentages of  $\beta$ -hairpin in both UA and AA models with experiment.<sup>76</sup> It can be found that both UA and AA simulations produce percentages of  $\beta$ -hairpin similar to experiment as measured CSD (Table 2). In both simulation and experiment HP5W4 is more stable. Interestingly, the AA model generates slightly more stable  $\beta$  hairpins than both the UA model and experiment.

(Table 2)

Besides analyzing  $\beta$ -hairpin stability from experiment, we also computed the secondary structure distributions in the UA model. Figure 2 shows that HP5w4 stays in the native  $\beta$  structure in 87.8% of the time, and HP5F stays in the native  $\beta$  structure in 78.26% of the time. The distribution of the fraction of native  $\beta$  structure was also examined. As can be seen in Figure 3, majority of the population is with high native  $\beta$  structure.

(Figures 2 and 3)

The native tertiary contact distribution and salt-bridge distribution were also analyzed. Figure 4 shows that majority of the population is with high fraction of native tertiary contact for both peptides. Figure 5 shows that majority of the population is with relatively low occupancy of any salt bridge, whether it is native or not. However it is still apparent that the native salt bridge, Lys2/Glu16, is indeed more populated than the two nonnative salt bridges.

(Figures 4 and 5)

### **3.2 First-Pass Folding Times**

Given the reasonable quality of the UA model with respect to experimental data, we went ahead to study the efficiency gain of the UA model over the AA model. Here the first pass folding times for the two peptides were analyzed. As defined in Methods, a folding event into the native basin consists of both potential energy falling in one standard deviation from the mean

potential energy of the reference *native* set simulations, and main chain RMSD within 1 Å from the mean structure of the reference *native* set simulations. Figure 6 shows the *ab initio* simulations of HP5w4 for both UA and AA models for the first 60 ns simulated. It can be seen that 6 *ab initio* folding events out of 20 independent trajectories occur during the 60 ns simulated for the UA model, while there is only 3 *ab initio* folding events out of 20 independent trajectories during the same amount of simulation time. Similarly for HP5F in Figure 7, there are 6 *ab initio* folding events up to 60 ns, while 4 events in the AA model within the same amount of simulation time.

(Figures 6 and 7)

All *ab initio* folding simulations were continued until the eleventh *ab initio* folding event in UA and AA was reached. Specifically, 200 ns were simulated in UA and 600 ns were simulated in AA for HP5w4, respectively. For HP5F 140 ns were simulated in UA and 600 ns were simulated in AA, respectively. All available *ab initio* first pass folding times are also listed in Table 3. These data show that the median first pass folding time is 184.15 ns for HP5w4 in UA and 120.95 ns for HP5F in UA. The median first pass folding time is 442.30 ns for HP5w4 in AA and 481.60 ns for HP5F in AA. Thus, the efficiency of the UA model is clearly beyond the per-step dynamics simulation of about 2 over the AA model. We have also analyzed the *ab initio* folded mean structures for both peptides. The representative snapshots are shown in Figure 8. Apparently, all *ab initio* folded trajectories indeed enter the native  $\beta$  hairpin basin whether the UA model or the AA model is used. The testing data confirms our hypothesis that a reasonable reduction in protein model can be achieved with noticeable improvement of sampling while still preserve a high level of accuracy for theoretical prediction in *ab initio* protein folding simulations.

(Table 3)

(Figure 8)

### 3.3 Thermodynamic Convergence

In this paper, the sampling efficiency of the UA model was also measured in term of how fast thermodynamic convergence can be achieved. It should be pointed out that ergodic measures were proposed to assess the sampling efficiency,<sup>82,83</sup> based on the ergodic theorem that the time average of an observable is equal to the configuration space average for an ergodic system. Usually in practice, squared variants of the observable between two different simulations can be calculated as ergodic measures.<sup>38</sup> However the ergodic measures suitable for protein systems are yet to be developed. Thus in this study, potential energy distributions and the structural distributions from different initial structures are analyzed to see whether, after long simulation time, they converge so that the same distributions can be obtained. We have compared the potential energy and structural distributions of the last 10 ns from the *ab initio* set with those of the last 10 ns from the *convergence* set. Recall that the two sets of 20 independent trajectories were started from two completely different sets of random structures. The convergence of potential energy and structural distributions from the UA and AA simulations are shown in Figure 9 and Figure 10, respectively. It is apparent that the UA simulations have reached convergence while the AA simulations from HP5w4 have not in the time allocated for simulation.

(Figures 9 and 10)

## 4. Conclusions

In this study, we have analyzed the quality of the recently proposed Amber united-atom model and its overall efficiency in *ab initio* folding of two  $\beta$ -hairpins. It is found that the

percentages of both  $\beta$ -hairpins are similar to those derived from measured CSD data whether the Amber united-atom or the all-atom model is used. Detailed analysis of simulated conformations shows that the majority of the sampled conformations are with high fractions of native  $\beta$  and native tertiary contacts. Interestingly, all salt bridges are only populated with low occupancy, around 2~12%, whether it is native or not. Nevertheless it is still apparent that the native salt bridges are indeed more populated than the two nonnative salt bridges.

Given the reasonable quality of the united-atom model with respect to experimental data, we have studied the efficiency gain of the united-atom model over the all-atom model in Amber. Analysis of the *ab initio* first pass folding time data shows that the median first pass folding time is 184.15 ns for HP5w4 and 120.95 ns for HP5F in the united-atom model, while the median first pass folding time is 442.30 ns for HP5w4 and 481.60 ns for HP5F in the all-atom model. Detailed structural analysis shows that all *ab initio* folded trajectories indeed enter the native  $\beta$  hairpin basin whether the united-atom model or the all-atom model is used. Finally, we have also studied the sampling efficiency of the united-atom model as measured in term of how fast thermodynamic convergence can be achieved. It is apparent that the united-atom simulations reach convergence faster with respect to both potential energy and backbone structural distributions than the all-atom simulations.

The findings are consistent with our previous test of a poly-alanine helix in the distance-dependent dielectric treatment. Thus the efficiency of the united-atom model is clearly beyond the per-step dynamics simulation of about 2 over the all-atom model: it is about 6 to 8 times faster as measured with the first-pass folding times for the tested peptides. Therefore, the testing data confirms our hypothesis that reasonable reduction of protein model can be achieved with noticeable improvement of sampling while still preserve a high level of accuracy for theoretical

prediction in *ab initio* protein folding simulations. This study motivates us to develop more simplified protein models to further enhance conformational sampling while still maintain sufficient accuracy in *ab initio* protein folding simulations.

## **Acknowledgements**

This work is supported in part by NIH (GM069620 & GM079383).

## References

- (1) Duan, Y.; Kollman, P. A. *Science* **1998**, *282*, 740-4.
- (2) Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. *Science* **1983**, *220*, 671-80.
- (3) Holland, J. H. *SIAM Journal on Computing* **1973**, *2*, 88-105.
- (4) Pedersen, J. T.; Moult, J. *Current Opinion in Structural Biology* **1996**, *6*, 227-31.
- (5) Lomize, A. L.; Pogozheva, I. D.; Mosberg, H. I. *Proteins-Structure Function and Genetics* **1999**, 199-203.
- (6) Osguthorpe, D. J. *Proteins-Structure Function and Genetics* **1999**, 186-93.
- (7) Samudrala, R.; Xia, Y.; Huang, E.; Levitt, M. *Proteins-Structure Function and Genetics* **1999**, 194-8.
- (8) Scheraga, H. A.; Lee, J.; Pillardy, J.; Ye, Y. J.; Liwo, A.; Ripoll, D. *Journal of Global Optimization* **1999**, *15*, 235-60.
- (9) Skolnick, J.; Kolinski, A. *Computing in Science & Engineering* **2001**, *3*, 40-50.
- (10) Standley, D. M.; Eyrich, V. A.; An, Y. L.; Pincus, D. L.; Gunn, J. R.; Friesner, R. A. *Proteins-Structure Function and Genetics* **2001**, 133-9.
- (11) Baker, D. *Abstracts of Papers of the American Chemical Society* **2001**, *221*, U432-U.
- (12) Jones, D. T. *Proteins-Structure Function and Genetics* **2001**, 127-32.

- (13) Piela, L.; Kostrowicki, J.; Scheraga, H. A. *Journal of Physical Chemistry* **1989**, *93*, 3339-46.
- (14) Stillinger, F. H.; Stillinger, D. K. *Journal of Chemical Physics* **1990**, *93*, 6106-7.
- (15) Church, B. W.; Oresic, M.; Shalloway, D. Tracking Metastable States to Free-Energy Global Minima. In *Dimacs Series in Discrete Mathematics and Theoretical Computer Science*; American Mathematical Society, 1996; Vol. 23; pp 41-64.
- (16) Shalloway, D. Packet Annealing: A Deterministic Method for Global Minimization Application to Molecular Conformation. In *Recent Advances in Global Optimization*; Princeton University Press, 1992; pp 433-77.
- (17) Wales, D. J.; Scheraga, H. A. *Science* **1999**, *285*, 1368-72.
- (18) Hansmann, U. H. E.; Okamoto, Y. The Generalized-Ensemble Approach for Protein Folding Simulations. In *Annual Review Computational Physics*; World Scientific: Singapore, 1999; Vol. 6; pp 129-57.
- (19) Okamoto, Y. *Journal of Molecular Graphics and Modelling* **2004**, *22*, 425-39.
- (20) Mitsutake, A.; Sugita, Y.; Okamoto, Y. *Biopolymers* **2001**, *60*, 96-123.
- (21) Yang, W.; Nymeyer, H.; Zhou, H. X.; Berg, B.; Brüschweiler, R. *Journal of Computational Chemistry* **2008**, *29*, 668-72.
- (22) Berg, B. A.; Neuhaus, T. *Physics Letters B* **1991**, *267*, 249-53.
- (23) Berg, B. A.; Neuhaus, T. *Physical Review Letters* **1992**, *68*, 9-12.

- (24) Nadler, W.; Hansmann, U. H. E. *Physical Review E* **2007**, *75*, 026109.
- (25) Lyubartsev, A. P.; Martsinovski, A. A.; Shevkunov, S. V.; Vorontsovvelaminov, P. N. *Journal of Chemical Physics* **1992**, *96*, 1776-83.
- (26) Marinari, E.; Parisi, G. *Europhysics Letters* **1992**, *19*, 451-8.
- (27) Irbäck, A.; Potthast, F. *Journal of Chemical Physics* **1995**, *103*, 10298-305.
- (28) Hansmann, U. H. E.; Okamoto, Y. *Physical Review E* **1996**, *54*, 5863-5.
- (29) Hansmann, U. H. E.; Okamoto, Y. *Journal of Computational Chemistry* **1997**, *18*, 920-33.
- (30) Marinari, E.; Parisi, G.; Ruiz-Lorenzo, J. J. Numerical Simulations of Spin Glass Systems. In *Spin Glasses and Random Fields*; World Scientific: Singapore, 1998; pp 59-98.
- (31) Irbäck, A.; Sandelin, E. *Journal of Chemical Physics* **1999**, *110*, 12256-62.
- (32) Swendsen, R. H.; Wang, J. S. *Physical Review Letters* **1986**, *57*, 2607-9.
- (33) Geyer, C. J. *Statistical Science* **1992**, *7*, 473-83.
- (34) Hukushima, K.; Nemoto, K. *Journal of the Physical Society of Japan* **1996**, *65*, 1604-8.
- (35) Hansmann, U. H. E. *Chemical Physics Letters* **1997**, *281*, 140-50.
- (36) Hansmann, U. H. E.; Okamoto, Y. *Current Opinion in Structural Biology* **1999**, *9*, 177-83.

- (37) Sanbonmatsu, K. Y.; García, A. E. *Proteins* **2002**, *46*, 225-34.
- (38) Lwin, T. Z.; Luo, R. *Journal of Chemical Physics* **2005**, *123*, 194904-10.
- (39) Lyman, E.; Ytreberg, F. M.; Zuckerman, D. M. *Physics Review Letters* **2006**, *96*, 028105.
- (40) Li, H.; Yang, W. *Journal of Chemical Physics* **2007**, *126*, 114104.
- (41) Liu, P.; Voth, G. A. *Journal of Chemical Physics* **2007**, *126*, 045106.
- (42) Mu, Y.; Yang, Y.; Xu, W. *Journal of Chemical Physics* **2007**, *127*, 384119.
- (43) Bandyopadhyay, P. *Journal of Chemical Physics* **2008**, *128*, 134103.
- (44) Christen, M.; Van Gunsteren, W. F. *Journal of Computational Chemistry* **2008**, *29*, 157-66.
- (45) Li, W.; Takada, S. *Journal of Chemical Physics* **2009**, *130*, 214108.
- (46) Skolnick, J.; Kolinski, A. *Monte Carlo Methods in Chemical Physics* **1999**, *105*, 203-42.
- (47) Kolinski, A.; Galazka, W.; Skolnick, J. *Journal of Chemical Physics* **1998**, *108*, 2608-17.
- (48) Gō, N. *Annual Review of Biophysics and Bioengineering* **1983**, *12*, 183-210.
- (49) Wolynes, P. G.; Onuchic, J. N.; Thirumalai, D. *Science* **1995**, *267*, 1619-20.

- (50) Onuchic, J. N.; Wolynes, P. G.; Lutheyschulten, Z.; Socci, N. D. *Proceedings of the National Academy of Sciences of the United States of America* **1995**, *92*, 3626-30.
- (51) Karplus, M.; Sali, A. *Current Opinion in Structural Biology* **1995**, *5*, 58-73.
- (52) Dill, K. A.; Chan, H. S. *Nature Structural Biology* **1997**, *4*, 10-9.
- (53) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. *Proteins-Structure Function and Genetics* **1995**, *21*, 167-95.
- (54) Dill, K. A.; Bromberg, S.; Yue, K. Z.; Fiebig, K. M.; Yee, D. P.; Thomas, P. D.; Chan, H. S. *Protein Science* **1995**, *4*, 561-602.
- (55) Li, A. J.; Daggett, V. *Journal of Molecular Biology* **1996**, *257*, 412-29.
- (56) Shakhnovich, E. I. *Current Opinion in Structural Biology* **1997**, *7*, 29-40.
- (57) Crippen, G. M.; Ponnuswamy, P. K. *Journal of Computational Chemistry* **1987**, *8*, 972-81.
- (58) Skolnick, J.; Kolinski, A. *Journal of Molecular Biology* **1990**, *212*, 787-817.
- (59) Levitt, M.; Warshel, A. *Nature* **1975**, *253*, 694-8.
- (60) Miyazawa, S.; Jernigan, R. L. *Macromolecules* **1985**, *18*, 534-52.
- (61) Zhou, Y. Q.; Hall, C. K.; Karplus, M. *Protein Science* **1999**, *8*, 1064-74.
- (62) Munoz, V.; Eaton, W. A. *Proceedings of the National Academy of Sciences of the United States of America* **1999**, *96*, 11311-6.

- (63) Thirumalai, D.; Klimov, D. K. *Current Opinion in Structural Biology* **1999**, *9*, 197-207.
- (64) Klimov, D. K.; Thirumalai, D. *Proceedings of the National Academy of Sciences of the United States of America* **2000**, *97*, 2544-9.
- (65) Liwo, A.; Oldziej, S.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *Journal of Computational Chemistry* **1997**, *18*, 849-73.
- (66) Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Oldziej, S.; Scheraga, H. A. *Journal of Computational Chemistry* **1997**, *18*, 874-87.
- (67) Liwo, A.; Kazmierkiewicz, R.; Czaplewski, C.; Groth, M.; Oldziej, S.; Wawak, R. J.; Rackovsky, S.; Pincus, M. R.; Scheraga, H. A. *Journal of Computational Chemistry* **1998**, *19*, 259-76.
- (68) Tan, C.; Yang, L.; Luo, R. *Journal of Physical Chemistry B* **2006**, *110*, 18680-7.
- (69) Tan, C.; Tan, Y. H.; Luo, R. *Journal of Physical Chemistry B* **2007**, *111*, 12263-74.
- (70) Dunfield, L. G.; Burgess, A. W.; Scheraga, H. A. *Journal of Physical Chemistry* **1978**, *82*, 2609-16.
- (71) Mcquarrie, D. A. Fluctuations. In *Statistical Mechanics*; University Science Books, 2000; pp 62.
- (72) Brooks, B. R.; Brucoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *Journal of Computational Chemistry* **1983**, *4*, 187-217.

- (73) Yang, L. J.; Tan, C. H.; Hsieh, M. J.; Wang, J. M.; Duan, Y.; Cieplak, P.; Caldwell, J.; Kollman, P. A.; Luo, R. *Journal of Physical Chemistry B* **2006**, *110*, 13166-76.
- (74) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *Journal of Computational Chemistry* **2005**, *26*, 1668-88.
- (75) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G. M.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J. M.; Kollman, P. *Journal of Computational Chemistry* **2003**, *24*, 1999-2012.
- (76) Fesinmeyer, R. M.; Hudson, F. M.; Andersen, N. H. *Journal of the American Chemical Society* **2004**, *126*, 7238-43.
- (77) Onufriev, A.; Bashford, D.; Case, D. A. *Proteins-Structure Function and Bioinformatics* **2004**, *55*, 383-94.
- (78) Lwin, T. Z.; Zhou, R. H.; Luo, R. *Journal of Chemical Physics* **2006**, *124*, -.
- (79) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *Journal of Computational Physics* **1977**, *23*, 327-41.
- (80) Sitkoff, D.; Case, D. A. *Journal of the American Chemical Society* **1997**, *119*, 12262-73.
- (81) Andersen, N. H.; Neidigh, J. W.; Harris, S. M.; Lee, G. M.; Liu, Z. H.; Tong, H. *Journal of the American Chemical Society* **1997**, *119*, 8547-61.

- (82) Thirumalai, D.; Mountain, R. D.; Kirkpatrick, T. R. *Physical Review A* **1989**, *39*, 3563-74.
- (83) Thirumalai, D.; Mountain, R. D. *Physical Review A* **1990**, *42*, 4574-87.
- (84) Hu, H.; Elstner, M.; Hermans, J. *Proteins-Structure Function and Genetics* **2003**, *50*, 451-63.

## Tables

**Table 1.** Sequences of the two tested  $\beta$ -hairpin model molecules HP5w4 and HP5F for folding simulations.

Molecule	Sequence
HP5w4	KKWTWNPATGKWTWQE
HP5F	KKYTWNPATGKFTVQE

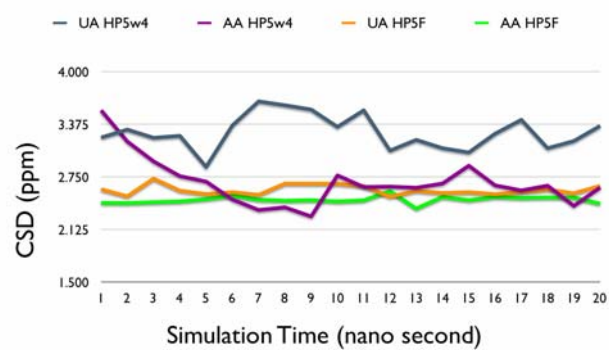
**Table 2.** Folded percentages of two simulated  $\beta$ -hairpins. Folded percentages are either measured in the NMR experiments or estimated from the molecular simulations of both united-atom models and all-atom models

	%folded (UA)	%folded (AA)	% folded (Fesinmeyer et al.)
HP5w4	94%	>99%	>96%
HP5F	83%	89%	82 $\pm$ 4%

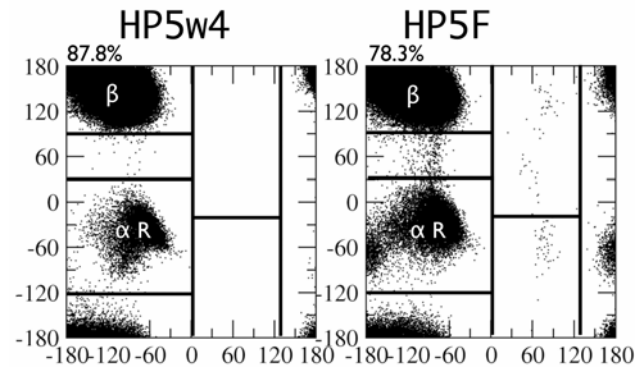
**Table 3.** Comparison of first pass *ab initio* folding times (ns) between united-atom and all-atom models. A total of 20 independent trajectories for each system in each model were simulated. The mean of the tenth and eleventh first pass times of the folding set is defined as the median folding times.

Molecule	1	2	3	4	5	6	7	8	9	10	11
HP5w4 UA	21.8	44.8	59.4	62.5	64.7	73.6	163.4	165.1	172.6	182.0	186.3
HP5w4 AA	25.2	38.9	60.5	94.7	125.7	171.6	181.3	185.1	190.0	365.7	518.9
HP5F UA	4.0	8.5	37.0	38.2	38.2	46.8	69.3	89.9	95.0	102.8	139.1
HP5F AA	8.7	9.0	39.3	44.4	58.2	79.4	143.5	146.6	244.0	431.3	531.9

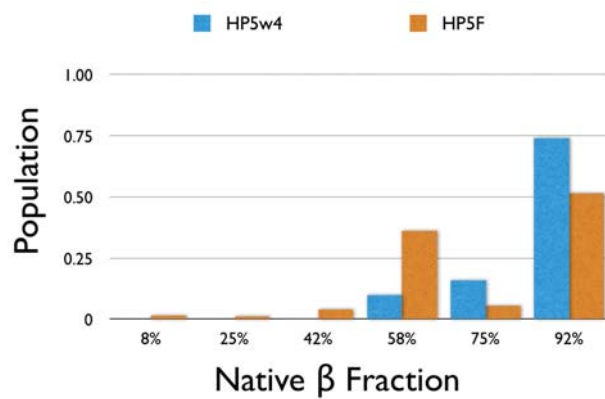
## Figures



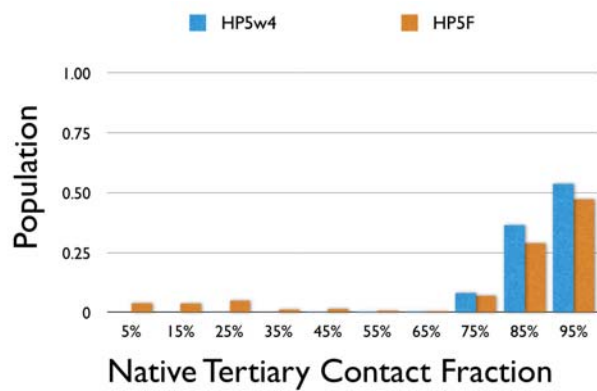
**Figure 1.** Time evolutions of CSD estimated in the 280K simulations. Gray line denotes the united-atom HP5w4 simulation, purple the all-atom HP5w4, orange the united-atom HP5F, and green the all-atom HP5F.



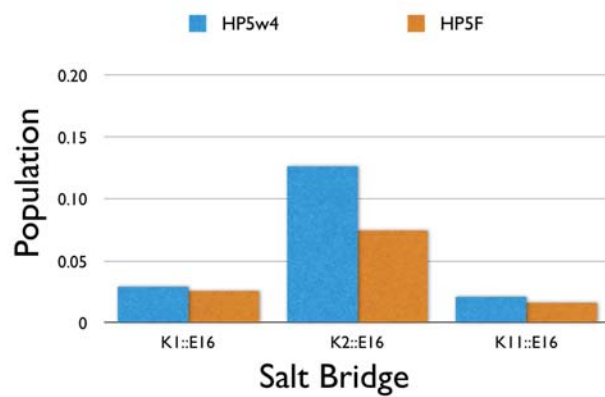
**Figure 2.**  $\phi/\psi$  distributions of  $\beta$  residues during the last 10 ns of native equilibrium simulated at 300K. The secondary structure definitions are from Hu et.al.<sup>84</sup>



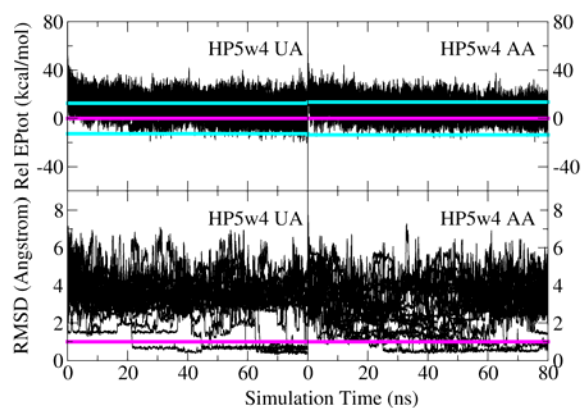
**Figure 3.** Distribution of native  $\beta$  fractions from the united-atom simulations.



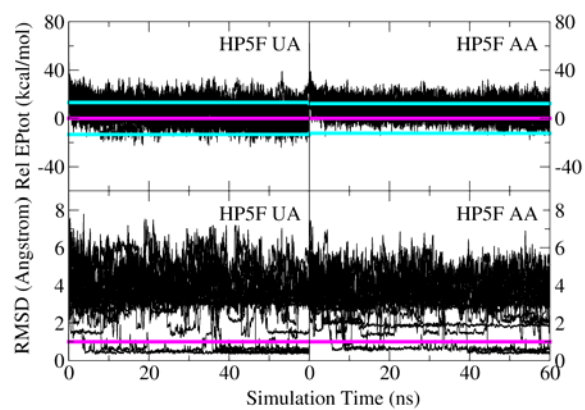
**Figure 4.** Distribution of native tertiary contact fractions from the united-atom simulations.



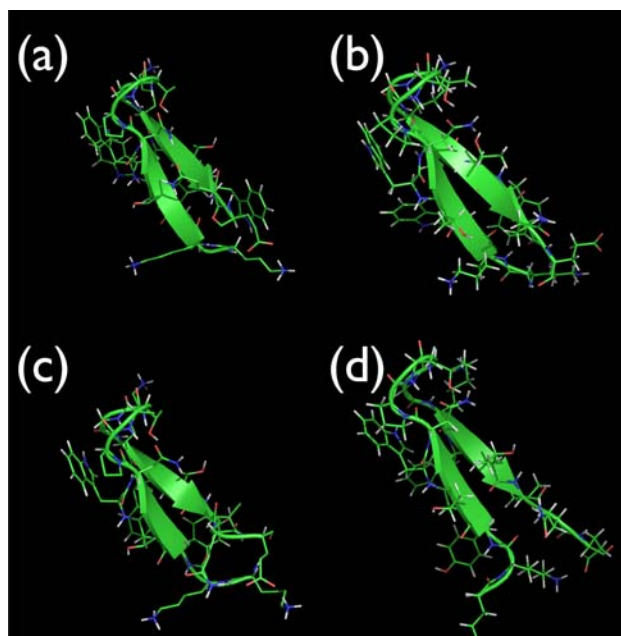
**Figure 5.** Distribution of salt bridges in the united-atom simulations.



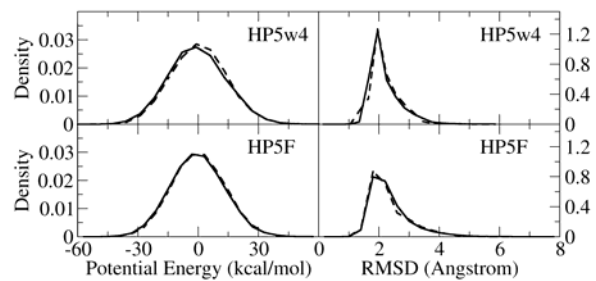
**Figure 6.** Potential energy and RMSD running averages versus simulation time for *ab initio* simulations of HP5w4 (in both united-atom and all-atom). In the potential energy plot, magenta lines indicate the averages of the reference simulations, and cyan lines indicate the standard deviations. Magenta lines in the RMSD plot represent the RMSD cutoff for folding events.



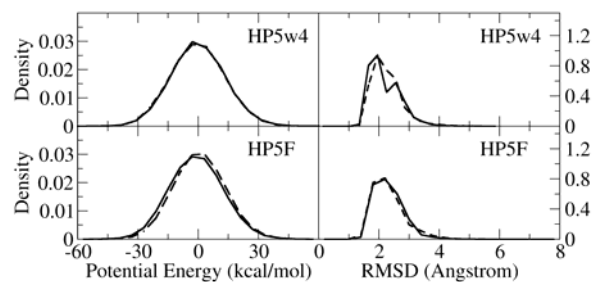
**Figure 7.** Same as Figure 6, but for HP5F.



**Figure 8.** Representative snapshots for the average folded conformations in *ab initio* folding simulations in united-atom (left) and all-atom (right): (a) and (b) are for HP5w4 and (c) and (d) are for HP5F. A representative snapshot is defined as the closest frame to the average folded conformation in the *ab initio* folding trajectories.



**Figure 9.** Potential energy distribution and RMSD distribution (binned over 20 windows) for the united-atom simulations. The solid curves are the distributions from the *ab initio* set of 20 simulations and the dashed curves are the distributions from the *convergence* set of 20 simulations.



**Figure 10.** Potential energy distribution and RMSD distribution (binned over 20 windows) for the all-atom simulations. The solid curves are the distributions from the *ab initio* set of 20 trajectories and the dashed curves are the distribution from the *convergence* set of 20 trajectories.

## Table of Contents Image

