Published in final edited form as:

Phys Rev Lett. 2005 September 23; 95(13): 138102.

Efficient Sampling of Protein Structures by Model Hopping

Wooseop Kwak^{*} and Ulrich H.E. Hansmann[†]

Department of Physics, Michigan Technological University, Houghton, MI 49931-1295, USA

Abstract

We introduce a novel simulation method, model hopping, that enhances sampling of low-energy configurations in complex systems. The approach is illustrated for a protein folding problem. Thermodynamic quantities of proteins with up to 46 residues are evaluated from all-atom simulations with this method.

The solution of many pharmacological and medical problems, such as rational drug design or the pathology of diseases associated with the mis-folding of proteins, requires a detailed understanding of the relation between chemical composition and structure of proteins. However, despite more than two decades of research, this so-called protein-folding problem has remained a hard computational task. This is because proteins in an all-atom representation are characterized by a rough energy landscape with a huge number of local minima separated by high energy barriers. Various numerical techniques exist that can alleviate this multiple minima problem. One popular example is parallel tempering (also known as replica exchange method) [1], a technique that was first introduced to protein studies in Ref. [2]. In its most common form, one considers an artificial system built up of N *non-interacting* copies of a molecule, each at a different temperature T_i . In addition to standard Monte Carlo or molecular dynamics moves that affect only one copy, parallel tempering allows also the exchange of conformations between two copies i and $j = i + 1$ with probability $w(\mathbf{C}^{old} \rightarrow \mathbf{C}^{new}) = \min(1, \exp(\Delta\beta\Delta E))$. The resulting random walk in temperature enables configurations to cross energy barriers and move out of local minima leading in this way to an enhanced sampling of low-energy structures. Variants of parallel tempering introducing non-canonical weights have been proposed [2,3].

Here, we introduce another variant of this idea, dubbed by us "model hopping" (MH), where as in the Hamilton Exchange method [4] or the Multi-Self-Overlap-Ensemble [5] the random walk in temperatures is replaced by one through an ensemble of models with slightly altered energy functions. For this we assume that the energy function can be separated in two terms: $E = E_A + aE_B$. As in parallel tempering, MH considers N non-interacting copies of the molecule, but copies are now exchanged according to

$$w(\mathbf{C}^{old} \rightarrow \mathbf{C}^{new}) = \min(1, \exp\{-\beta[E_A(\mathbf{C}_j) + a_j E_B(\mathbf{C}_j) + E_A(\mathbf{C}_i) + a_j E_B(\mathbf{C}_i) \quad (1)$$

$$- E_A(\mathbf{C}_i) - a_i E_B(\mathbf{C}_i) - E_A(\mathbf{C}_j) - a_j E_B(\mathbf{C}_j)]\}) \quad (2)$$

$$= \min(1, \exp\{\beta\Delta a\Delta E_B\}). \quad (3)$$

Here, $\Delta a = a_j - a_i$ and $\Delta E_B = E_B(\mathbf{C}_j) - E_B(\mathbf{C}_i)$. Due to this exchange move configurations perform a random walk on a ladder of models with $a_1 = 1 > a_2 > a_3 > \dots > a_N$ that differ by

^{*} E-mail: wkwak@mtu.edu.

[†] E-mail: hansmann@mtu.edu

the relative contributions of E_B to the total energy E of the molecule. For instance, barriers in the energy landscape of proteins often arise from van der Waals repulsion between atoms that come too close. Generalized-ensemble techniques circumvent the problem by performing a random walk in energy that allows crossing of these barriers. On the other hand, in MH the protein walks randomly up and down on a ladder of models with successively smaller contributions from the van der Waals energy. While the "physical" system is on one side of the ladder (at $a_1 = 1$), the (non-physical) model on the other end of the ladder (at $a_N \ll 1$) may allow atoms to share the same position in space. As the protein "tunnels" in this way through energy barriers, sampling of low-energy configurations will be enhanced in the "physical" model (at $a_1 = 1$).

In computational protein science, algorithms are often tested for simple molecules such as the pentapeptide Met-enkephalin. However, such test can be misleading as the physics (and therefore also the complexity of the problem) is different for larger molecules. For this reason, we have chosen already for our initial tests a much larger molecule, the 36-residue villin headpiece sub-domain HP-36, which has recently raised considerable interest in computational protein studies [6–9]. With its well-defined secondary and tertiary structure [10] it is sufficiently complex and with 596 atoms large enough that numerical simulations become indeed a challenge [11]. Using the computer program SMMP [12] we simulate eight copies of HP-36 at temperature $T = 300$ K, with energies given by:

$$E_{tot}^{(i)} = a_i E_{LJ} + E_C + E_{HB} + E_{Tors} + E_{Solv} = a_i E_{LJ} + E_{other}. \quad (4)$$

Here, the interactions between the atoms of the HP-36 molecule are described by the ECEPP/2 force field [13] given by the sum of an electrostatic term E_C , the Lennard-Jones energy E_{LJ} , hydrogen-bond term E_{HB} and torsion energy E_{tor} . The protein-water interactions are approximated by $E_{Solv} = \sum_i \sigma_i A_i$. Here, the sum goes over all atoms, the A_i are the solvent accessible surface areas of atoms i and the parameters σ_i are taken from Ref. [14].

Our implementation of MH attempts to enhance sampling of low-energy configurations by performing a random walk through a ladder of models with successively lowered contribution of the van-der-Waals energy E_{LJ} . Short preliminary runs indicated as suitable weights for HP-36 $a_i = 0.61, 0.64, 0.68, 0.74, 0.79, 0.86, 0.91, 1.00$. With these weights we performed then a simulation with 200,000 sweeps. After each sweep, an exchange move between two copies (a_i, a_{i+1}) is accepted or rejected according to Eq. 3. That the resulting random walk in the strength of the vdW-term leads to a faster convergence of the system can be seen in Fig. 1a where we draw the time series in energy (for $a_i = 1.0$) and contrast it with a regular canonical simulation at $T = 300$ K. We have plotted here the energies as a function of "effective" MC sweeps taking into account that one sweep in MH corresponds to $N = 8$ sweeps in regular Monte Carlo. Note that the energies in the MH run are on average 70 kcal/mol lower than that of the regular MC run that got trapped in a local minimum.

How compares the new technique with other generalized-ensemble methods? Fig. 1b relates the MH run to a parallel tempering simulation (PT) of 20 replicas where the temperature distribution is the one used by us in earlier work [8]. The energies are again plotted against "effective" MC-sweeps. For long MC times (more than 1,000,000 effective MC sweeps) both techniques lead to the same energy distribution centered around $E = -350$ kcal/mol. MH approaches this value in $\approx 250,000$ effective sweeps while the PT simulation requires $\approx 800,000$ sweeps for convergence. We conclude that in our example model hopping is by a factor three more efficient than regular parallel tempering. This improvement is the more remarkable as the temperature protocol of the PT run is highly optimized.

As noticed in earlier work [8], our simulations are limited by the accuracy of our energy function that overestimates helical structures. The vast majority of configurations at $T = 300$ K differ from the native structure by more than $7 \text{ \AA}$ [8]. It was argued in Ref. [8] that this shortcoming is due to the solvation term E_{solv} which underestimates the protein-water interactions. A possible explanation is that our implicit solvent model assumes formation of a hydrophobic core which the comparatively small HP-36 barely exhibits. It is interesting to see whether our energy function performs better for larger molecules with a more pronounced hydrophobic core. Hence, we extend our investigations to the 46-residue long 10–55 fragment of the B domain of staphylococcal protein A. This three-helix bundle [15] has been studied extensively with both coarse-grained [16,18,19] and all-atom models [17]. The experimentally observed structure (PDB code 1BDD) of the protein is displayed in Fig. 2 with the three helices marked I, II, and III, respectively. After minimization in our force field, energy of the protein is $E_{Tot} = -596.9 \text{ kcal/mol}$.

As in the case of HP-36 we have performed short preliminary runs to obtain a suitable set of weights a_i . We find $a_i = 0.60, 0.63, 0.68, 0.74, 0.80, 0.85, 0.93, 1.00$. With these weights we then performed a simulation with 150,000 sweeps. The resulting time series of energy is displayed in Fig. 3. The total energy E_{Tot} decreases almost linearly in the first 40,000 sweeps to stay afterward constant. This initial decrease results from the ECEPP energy E_{ECEPP} while the implicit solvent term E_{solv} increases. Hence, the interactions within the molecule are optimized over the course of the simulation while the interactions with the solvent become less favorable. Such behavior is expected for a helical protein where the intra-molecular hydrogen bonds compete with such formed with the surrounding water. After the initial 40,000 sweeps the N-terminal helix (helix I) is formed in more than 90% of all configurations and the C-terminal helix (helix III) in $\approx 55\%$ of configurations. The middle helix (II) is partially formed in more than 90% of configurations but is only in 2% completely formed. This is due to the GLY₂₁ residue which introduces a flexibility to the chain that tends to interrupt growth of this helix.

Fig. 4a displays the configuration with lowest energy obtained in our MH simulation. After minimization its energy of $E_{Tot} = -666.5 \text{ kcal/mol}$ is 70 Kcal/mol lower than that of the PDB-structure. The radius of gyration $r_{gy} = 10.7 \text{ \AA}$ and solvent accessible surface $A = 3804 \text{ \AA}^2$ are larger than the values found for the PDB-structure ($r_{gy} = 9.7 \text{ \AA}$ and $A = 3333 \text{ \AA}^2$). All three helices are formed. However, the middle helix II is broken up by the GLY₂₁ residue into helices II(a) and II(b), and this configuration is therefore build out of four helices. The total helicity, defined as the number of residues that are part of an α -helix, is with 85% higher than for the PDB-structure (74%). The root-mean square deviation (rmsd) between both configurations (calculated over backbone atoms) is $6.8 \text{ \AA}$.

As in the case of the smaller HP-36 the native structure of the B domain of protein A is *not* the global minimum configuration for our energy function. On the other hand, about 10% of our configurations are with rmsd values of $4\text{--}5 \text{ \AA}$ closer to the native structure than the best one ($5.7 \text{ \AA}$) found for HP-36. One example is shown in Fig. 4b. While this configuration has only a helicity of 63%, a solvent accesible area $A = 3923 \text{ \AA}^2$ and is with $R_{gy} = 10.7 \text{ \AA}$ less compact. The rmsd to the PDB structure is only $4.2 \text{ \AA}$ and comparable with results obtained in earlier simulations of less detailed models [16,18]. The rmsd results from differences in helix II that is not formed for residues 16 to 21 (the helix-breaking glycine). Instability of helix II in this region has been also observed in earlier work using a more coarse grained model [16]. The solvation energy $E_{solv} = -193.3 \text{ kcal/mol}$ of this configuration is lower than that of the lowest energy structure, but its intra-molecular energy is with $E_{ECEPP} = -463.8 \text{ kcal/mol}$ higher. As a consequence, its total energy $E_{Tot} = 657.1 \text{ kcal/mol}$ is 9 kcal/mol higher than that of the lowest energy structure.

Table 1 lists various partial energies for the the PDB-structure of Fig. 2 and the two low-energy configurations of Fig. 4 that were found in our MH simulation. The Lennard-Jones term E_{LJ} and the torsion term E_{Tors} favor the decoy structures over the native one. Hence, an accurate solvation term should compensate these intra-molecular contribution. However, the solvation energy E_{Solv} of the PDB-structure (Fig. 2) is more than 20kcal/mol higher than that of the incorrectly folded lowest-energy configuration of Fig. 4a. For a more detailed analysis we have divided E_{Solv} into two parts E_{np} and E_p that result from summation over only non-polar atoms and polar atoms, respectively. The differences in E_{np} (resulting from exposure of non-polar atoms to water) are small between the three configurations while the areas A_{np} differ notably. On the other hand, the contributions E_p that result from polar atoms in the pdb-structure is 20 kcal/mol higher than that of the two decoys where the solvent accessible area of polar atoms is almost 1000 Å² larger. Hence, our approximation of the solvation energy seems to favor excessively exposure of polar groups to water. This is consistent with previous observations that a solvent-accessible surface approach describes only poorly the behavior of polar groups [20].

Our results indicate that the inaccuracies of our implicit solvent model do not cancel out for larger proteins. They rather suggest to tune the solvation parameters σ_i so that exposure of polar groups is favored less. In a simulation of protein A of 100,000 sweeps with such a (not optimized) modification of the solvation energy we found indeed a lowest-energy configuration that differs from the native structure by only 3.2 Å. This configuration resembles by visual inspection the one of Fig. 4b but is with a radius of gyration $R_{Gy} = 10.0$ Å more compact and has a smaller solvent accessible surface area (3553Å²). Especially the polar atoms are substantially less exposed to water. Its area $A_{np} = 1864$ Å² is in between the values for the native structure and the ones of Fig. 4.

As the success of protein simulations seems to be less limited by the search algorithms than by the available energy functions the quest for advanced simulation techniques becomes even more important. This is because unveiling the shortcomings of energy functions requires many folding simulations of sufficiently large proteins. Such simulations are too time-consuming with traditional methods but have become possible with techniques such as model hopping. In our example, configurations move in an ensemble of models with successively reduced contributions of the van der Waals term. Other modifications of the energy may prove even more successful and are currently investigated by us. We are especially interested in application of MH to multi-scale modeling of biological molecules [21].

Acknowledgements

Support by a research grant from the National Institutes of Health (GM062838) is gratefully acknowledged. Part of this article was written while U.H. was visiting the Center of Excellence in Bioinformatics at Buffalo University. He thanks Jeff Skolnick and his group for kind hospitality.

References

1. Hukushima K, Nemoto K, Geyer GJ. J Phys Soc (Jpn) Stat Sci 1996;1992;657:1604. 437.
2. Hansmann UHE. Chem Phys Lett 1997;281:140.
3. Whitfield TW, Bu L, Straub JE. Physica A 2002;305:157.
4. Fukunishi H, Watanabe O, Takada S. J Chem Phys 2004;116:9058.
5. Chikenji G, Kikuchi M, Iba Y. Phys Rev Lett 1999;83:1886.
6. Hansmann UHE, Wille LT. Phys Rev Lett 2002;88:068105. [PubMed: 11863858]
7. Shen MY, Freed KF. Proteins 2002;49:439. [PubMed: 12402354]
8. Lin CY, Hu CK, Hansmann UHE. Proteins: Structure, Functions and Genetics 2003;52:436.
9. Ripoli DR, Vila JA, Scheraga HA. J Mol Biol 2004;339:915. [PubMed: 15165859]
10. McKnight CJ, Doebring DS, Matsudaria PT, Kim PS. J Mol Biol 1996;260:126. [PubMed: 8764395]

11. Duan Y, Kollman PA. Science 1998;282:740. [PubMed: 9784131]
12. Eisenmenger F, Hansmann UHE, Hayryan Sh, Hu CK. Comp Phys Comm 2001;138:192.
13. Némethy G, Gibson KD, Palmer KA, Yoon CN, Paterlini G, Zagari A, Rumsey S, Scheraga HA. J Phys Chem 1992;96:6472.
14. Ooi T, Obatake M, Nemethy G, Scheraga HA. Proc Natl Acad Sci USA 1987;8:3086. [PubMed: 3472198]
15. Gouda H, Torigoe H, Saito A, Sato M, Arata Y, Shimada I. Biochemistry 1992;31:9665. [PubMed: 1390743]
16. Favrin G, Irbäck A, Wallin S. Proteins 2002;47:99. [PubMed: 11933057]
17. EM, Boczko; ChL, Brooks, III. Science 1995;269:393. [PubMed: 7618103]
18. Lee J, Liwo A, Scheraga HA. Proc Nat Aca Sci (USA) 1999;96:2025.
19. Ghosh A, Elber R, Scheraga HA. Proc Nat Aca Sci (USA) 2002;99:10394.
20. Karplus PA. Protein Science 1997;6:1302. [PubMed: 9194190]
21. J. Meinke and U.H.E. Hansmann, in preparation.

Fig. 1.

Time series of total energy as obtained in a MH run (for $a = 1$) at $T = 300$ K and (a) in a regular canonical Monte Carlo simulation and (b) a parallel tempering (PT) run at the same temperature.

Fig. 2.
Experimental observed structure of the fragment 10–55 of the B domain of staphylococcal protein A (PDB-code 1BDD).

Fig. 3. Time series of intra-molecular energy E_{ECEPP} , solvation energy E_{Solv} and total energy $E_{Tot} = E_{ECEPP} + E_{Solv}$ as a function of Monte Carlo time.

Fig. 4.

- (a) Lowest energy configuration of the B domain as obtained in our MH simulation.
- (b) A competing low-energy structure that resembles the native structure of Fig. 2.

TABLE I

Various energy terms for the protein A configuration shown in Figs. 2 and 4. We also list the corresponding areas A_{np} and A_p . Energies are given in kcal/mol and areas in $\text{\AA}^2$.

	Fig. 2	Fig. 4a	Fig. 4b
E_C	-51.9	-57.1	-51.1
E_{LJ}	-351.0	-374.2	-364.0
E_{HB}	-69.3	-74.0	-64.9
E_{Tors}	34.2	16.0	16.2
E_{np}	23.7	24.5	23.9
E_p	-182.2	-203.4	-217.2
A_{np}	1463	1728	1721
A_p	1341	2121	2254