

Linearly scalable hybrid Monte Carlo method for conformational sampling of large biomolecules*

Scott Hampton[†], Jesús A. Izaguirre[‡]

September 21, 2002

Abstract

We present a variation on the hybrid Monte Carlo (HMC) algorithm that improves the sampling of phase space. This new algorithm, called Shadow Hybrid Monte Carlo (SHMC), achieves a nearly linear scalability with system size rather than $O(N^{5/4})$ of HMC. We tested both methods using vacuum and solvated biological molecules. Our results show that, for example, when sampling the conformational space of a solvated Melittin with 5123 atoms, SHMC can use time steps of 0.8 fs using leapfrog, whereas HMC requires time steps of 0.05 fs to produce comparable acceptance rates (around 99%). This represents an asymptotic sixteen-fold speedup in the sampling rate. The benefits could be greater for even larger systems. If

*This work was supported by NSF Grant BIOCOMPLEXITY-IBN-0083653, NSF CAREER award ACI-0135195.

SH had an Arthur J. Schmitt fellowship from the University of Notre Dame.

[†]shampton@cse.nd.edu

[‡]izaguirr@cse.nd.edu, Department of Computer Science and Engineering, University of Notre Dame,

Notre Dame, IN 46556-0309, USA

combined with other sampling enhancement techniques, this method might serve to sample the conformational space of proteins in folding studies or similarly challenging problems. SHMC is based on computing a cheap modified Hamiltonian for leapfrog at arbitrarily high accuracy at the expense of extra storage, using the shadow Hamiltonian method of Skeel and Hardy [SIAM J. on Sci. Comp., 23(4):1172, 2001]. This increased accuracy gives substantial gains in the acceptance rates of SHMC. Finally, we present important theoretical and practical issues we still need to address to make the method widely useful.

1 Introduction

The problem of sampling the conformational space of large biological macromolecules (biomolecules) has elicited considerable research on improving sampling methods. Applications such as protein folding require an understanding of the conformational space of these macromolecules. Biomolecules with the common all-atom force fields have rugged energy landscapes and have multiple time scales that both limit the small time step of molecular dynamics (MD) and require the integration of many time steps to sufficiently sample the slow molecular modes. This is exacerbated by trap-pings in local minima. To try to overcome these problems, techniques such as hybrid Monte Carlo (HMC) combine relatively short MD trajectories that are accepted or rejected using Monte Carlo (MC).

HMC is an attractive method but it suffers poor scalability with system size. The acceptance rate of HMC decays exponentially with the product of N and the square of the discretization error of the MD trajectory. MD integrators of higher accuracy can help the scalability problem, but in simulations of biomolecules the cost of evaluating the forces is dominant and more than offsets

any gains on acceptance rate.

We present in this paper an HMC algorithm that uses the shadow Hamiltonian, an approximation of arbitrary accuracy to the modified Hamiltonian computed exactly by the symplectic MD integrator. This calculation is fast and consists of dot products of previous values of velocities and positions. The expense is thus in terms of storage. Using this new method we achieve significant speedups over regular HMC.

2 Sampling techniques

Here we discuss MD, MC, and HMC, the sampling techniques most relevant to this paper. Berne and Straub [2] discuss several novel sampling techniques such as J-walking and potential-smoothing methods. Some of these techniques use HMC as one of their components, and thus our new method could be combined with them to boost the performance of either method separately.

2.1 Molecular Dynamics and Monte Carlo

Molecular dynamics is a natural tool to explore the conformational space of molecules, particularly proteins and other biomolecules [12, p. 434]. A conformation is a particular geometry of the molecule. Stable geometries are believed to correspond to an energy minimum [7]. For example, MD simulations permit ligands and receptors to explore their configurations in space to determine the efficiency of molecular docking. MD is also commonly used to study thermodynamic properties of systems [1, p. 46].

The basic MD algorithm involves a time integration of the system of molecules. Some initial configuration is chosen for time t_0 which includes the positions X and the momenta P for each

molecule. The system is then integrated to time t_1 based on the positions and momenta at time t_0 . This process is repeated using the values from time t_1 to compute the values for time t_2 and so forth. Many algorithms exist for performing this type of integration. One of the more popular algorithms is called Verlet/Leapfrog [17].

The algorithm starts with a half-kick which updates the momenta at $t + \frac{\Delta t}{2}$. The updated momenta depend on the forces from the previous time step. These momenta are then used to compute the new positions at time $t + \Delta t$. Once the new positions are calculated, the forces can be recomputed at time $t + \Delta t$. Finally, the momenta are calculated for time $t + \Delta t$ based on the most recent forces. Once the algorithm has been started, the two half-kicks may be combined to form a single kick. In this way, the calculations of the momenta and positions *leapfrog* each other.

The MD algorithm has many positive attributes. It is easy to understand, simple to program and does a good job of moving between states in phase space. However, the algorithm is limited by instabilities [10]. These instabilities require the use of a small timestep in relation to the processes of interest. MD is handicapped by the amount of time necessary to complete a calculation.

For a simulation to equilibrate quickly it is extremely important, if not sufficient, to use methods that explore a representative selection of configurations as quickly as possible. The power of an MD algorithm is that it takes relatively large steps in phase space¹ and allows the positions and velocities of all the degrees of freedom of the system to be updated simultaneously. However, one disadvantage is that the calculated trajectories are not necessarily ergodic², particularly when the system contains harmonic degrees of freedom. Monte Carlo (MC) simulations do not have this

¹Phase space is a multidimensional space defined by the positions and velocities of all particles in the system.

²The ergodic hypothesis states that the time average of a thermodynamic property provided by molecular dynamics should be the same as the ensemble average provided by Monte Carlo, cf. [9, pp.15,16].

problem because they can improve ergodicity by making random changes to the configurations. The drawback to traditional MC simulations is the slow rate at which they explore phase space, particularly for dense systems or those with long-range interactions [3, 12].

2.2 Hybrid Monte Carlo Method

Given that the two techniques complement each other in their ability to explore the phase space, a variety of hybrid methods have been devised, in which the simulation algorithm alternates between MD and MC [3, 4, 8]. These hybrid Monte Carlo (HMC) algorithms combine the large steps taken in phase space by MD with the ability of MC to change direction of the trajectory randomly. Thus the MC part of the simulation ensures ergodicity and eliminates inaccuracies in the energy while the MD part speeds up the simulation by allowing large steps to be taken in phase space. The great difficulty of HMC is that it does not scale well with system size, and thus it becomes less attractive for large biomolecules.

3 The Shadow Hybrid Monte Carlo algorithm

We propose a novel method based on HMC that uses a shadow Hamiltonian to determine acceptance. Because the shadow Hamiltonian can be computed to a high accuracy in our method, we have a much higher acceptance rate than traditional HMC. This allows us to make more moves in phase space and therefore improve sampling of the phase space.

Note that the acceptance rate, $\min(1, e^{-\langle \delta H \rangle})$, will be unity if the molecular dynamics conserve Hamiltonian exactly, that is, $\langle \delta H \rangle = 0$. It is impossible to integrate Hamilton's equation exactly for the forces used in MD: the discretization error increases with time step and system size.

However, the acceptance ratio will be improved if the discrete integrator conserves the Hamiltonian with higher accuracy. The cost of using higher accuracy integrators for MD offsets potential gains, except at very large problem sizes [5].

Symplectic integrators integrate a modified Hamiltonian system that is close to the Hamiltonian of interest [15]. A cheap approximation to the modified Hamiltonian has been introduced recently. It is called the shadow Hamiltonian method [18] and it can achieve arbitrary accuracy in calculating the modified Hamiltonian using k past values of velocities and positions, with an accuracy $p = 2k$, that is $\delta H = O(\delta t^{2k})$.

We followed the analysis of the scalability of HMC with system size found in [5, 11] and [14, p. 84]. The result of these studies shows that the computational effort of HMC using leapfrog moving from one configuration to an approximately independent configuration is proportional to $N^{5/4}$. To maintain a fixed acceptance rate, the step size must decrease as $N^{-1/4}$ as the system size N increases. In general, for a symplectic integrator it is shown that the acceptance rate is approximately $\exp(-N \langle \delta H^2 \rangle)$, where $\langle \delta H \rangle$ is the discretization error made by the integrator. Thus, the computational cost becomes $N^{1+\frac{1}{2p}}$ where p is the order of accuracy of the integrator ($p = 2$ for leapfrog).

Using a highly accurate shadow Hamiltonian for the acceptance rule in HMC, the asymptotic behavior can be improved towards linear in a systematic way. Indeed, for a shadow Hamiltonian of order p one would expect a speedup over leapfrog in the order of $N^{5/4}/N^{\frac{2p+1}{2p}} = N^{\frac{p-2}{4p}}$. As p grows this method will gain a speedup of $N^{1/4}$. Because the shadow Hamiltonian method uses k values of forces and energies which are already available during the integration of equation of motion, the additional computational cost is low, only a few dot products and vector additions. There is an

additional memory requirement to form the shadow Hamiltonian.

For $p = 8$, we would expect a speedup over plain HMC of $N^{3/16}$. For $N = 5,000$ the speedup would be ≈ 9 . For $N = 30,000$ it would be ≈ 14 . More importantly, the speedup will grow with N , which makes this method a scalable algorithm. Our numerical results show that for a system of 5123 atoms we get a speedup of 16 by using SHMC rather than HMC, although correlations in the initial and final trajectories are not taken into account; see Section 4. When one takes into account that bigger system sizes require slightly longer runs, the speedups will be closer to our theoretical estimate.

3.1 Implementation of the Shadow Hybrid Monte Carlo algorithm

The SHMC is shown in Algorithm 1. The main difference with HMC is that we are using the shadow Hamiltonian instead of the Hamiltonian in the acceptance step. Even though a rigorous justification is still lacking, there is evidence suggesting this may be valid: First, it can be proven that for a trajectory y_n calculated by a symplectic integrator such as leapfrog, there exists a nearby modified Hamiltonian $\hat{H}(y)$ whose analytical trajectory $\hat{y}(t)$ satisfies $\hat{y}(nh) - y_n = O(\exp(-c/\delta t))$ for time $nh \leq c/\delta t$. Second, it has been reported in the literature that “in spite of the short predictability time t_p on a single trajectory [of MD] we found that the statistical properties are not sensitive to small changes in the evolution law ... This feature holds also for correlation functions at a delay larger than t_p .” [6]. This implies that the statistics are robust to small perturbations in the Hamiltonian.

The additional expense of the SHMC is in terms of storage since it is necessary to keep k copies of the positions, momenta, and a beta term to increase the accuracy to $O(\delta t^{2k})$. In our numerical

Generate new random velocities

Compute shadow energy \mathcal{SH}_0

Run MD algorithm for *cyclelength* steps

Compute shadow energy \mathcal{SH}_1

Compute change in shadow energy $\delta\mathcal{SH} = \mathcal{SH}_1 - \mathcal{SH}_0$

Choose a uniform random number, r , between $[0, 1]$

Accept new positions if $r < \exp^{-\frac{\delta\mathcal{SH}}{TK_b}}$. Otherwise, reject new positions and restore old positions.

Algorithm 1: Shadow Hybrid Monte Carlo Algorithm. T is the temperature and K_b is Boltzmann’s constant.

experiments we get very good speedups with $k = 4$ for an $O(\delta t^8)$ accurate shadow Hamiltonian.

HMC and SHMC were implemented inside of PROTOMOL, a framework designed for testing molecular simulation algorithms [13]. PROTOMOL is implemented using an object oriented and generic design and contains a hierarchical integrator structure. The integrator hierarchy is based on a general `Integrator` class. This class contains methods that are general to all types of integrators. Inheriting from `Integrator` is `StandardIntegrator` which is in turn a super class to both `MTSIntegrator` and `STSIntegrator`. This arrangement lets us distinguish between multiple and single time stepping integrators respectively. `LeapfrogIntegrator` is a subclass of `STSIntegrator` while `HMCIntegrator` is a subclass of `MTSIntegrator`. The code for both HMC and SHMC is located in class `HMCIntegrator`. A simplified version of the hierarchy is located in Figure 1. A boolean flag set when PROTOMOL is initialized determines which method is used. The shadow Hamiltonian is an approximation to the true Hamiltonian:

$$\mathcal{H}_{[2k]} = \mathcal{H} + O(h^{2k}).$$

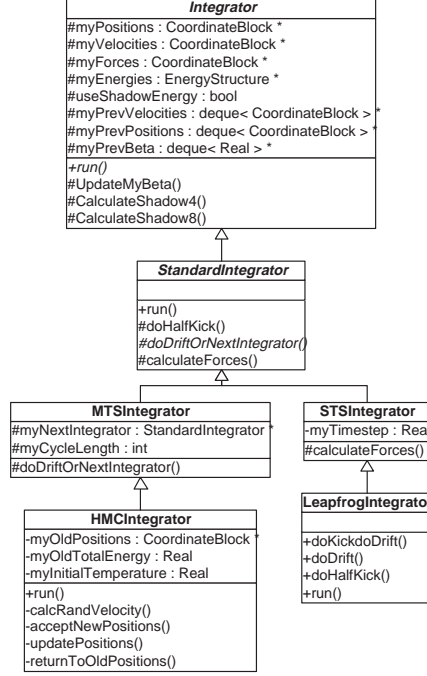


Figure 1: A simplified version of the integrator hierarchy found in PROTOMOL .

We found the 8^{th} order shadow to be a good balance between accuracy and storage cost. The formula for the 8^{th} order shadow ($k = 4$) [18] is as follows:

$$\mathcal{SH}_8 = A_{10} + \frac{5}{42} * A_{30} + \frac{13}{105} * A_{32} - \frac{2}{7} * A_{12} - \frac{19}{210} * A_{14} - \frac{1}{140} * A_{34} \quad (1)$$

The formula for A_{ij} is somewhat complicated to reproduce here. Those interested in the details may find them in [18]. Fortunately, the problem is essentially one of dot products and is therefore relatively easy to implement. For example,

$$\begin{aligned} A_{10} &= X_1 \cdot V_0 - V_1 \cdot X_0 - \frac{1}{2}(\beta^{n+1} - \beta^{n-1}) \\ &= \left(\frac{1}{2M}(X^{n+1} - X^{n-1}) \cdot V^n\right) - \left(\frac{1}{2M}(V^{n+1} - V^{n-1}) \cdot X^n\right) - \frac{1}{2}(\beta^{n+1} - \beta^{n-1}) \end{aligned}$$

Here, X^i and V^j represent the positions and velocities at time i and j respectively. Similarly

for β , which is a variable that is propagated along with the momenta by the integrator.

4 Numerical experiments

The objective of our experiments was to test the scalability of HMC and SHMC as system size and time step are increased. Our experiments were performed on four different molecules of increasing size. They are Alanin with 66 atoms, a TIP3P water with 423 atoms, Melittin³ with 868 atoms and Melittin plus TIP3P water with 5143 atoms.

The experiments involved testing the acceptance rate of HMC versus SHMC for each of the molecules. The same PROTOMOL configurations were used for each molecule. The temperature was set to be 300 K. We used periodic boundary-conditions. The forces computed include Lennard Jones using non-bonded cutoff of 6.5Å with a C^2 switching function on the energy. Also, Coulomb using Particle Mesh Ewald with a 20^3 grid size. All simulations were run for 100 ps except Melittin+water which was run for 10 ps. Leapfrog was used for the MD integration. The only difference between the runs using HMC and those using SHMC was the acceptance criteria. HMC uses the Hamiltonian, whereas SHMC uses the 8th order shadow.

For each molecule, we ran several experiments and varied the timestep of Leapfrog and the cyclelength of HMC/SHMC. We chose cyclelengths of 20, 40 and 80. We found that the cyclelength does not have a significant effect on the acceptance rate except in special cases: there are artifacts due to resonances of periodic forces that prohibit simulations at many timesteps. This agrees with observations in [14]. We chose timesteps that were stable with respect to Leapfrog. Once a

³Melittin [19], is a toxic substance found in the venom of honey bees. It can be retrieved from the Protein Data Bank using ID 2mlt

timestep was reached for which HMC produced a zero acceptance rate, we went no further. The timesteps vary between different molecules. The reason being that as system size increases, the discretization error increases and thus the timestep must be lowered in order to get any acceptance at all. This is especially true for water and Melittin.

Our experiments show that SHMC always out performed regular HMC as the system size was increased. Based on acceptance rates of 97% or higher, SHMC can use timesteps from 4 to 16 times larger than those of HMC for the same system. This is evident in Figure 2. For our four systems, we plotted the largest timestep for which HMC and SHMC returned an acceptance rate

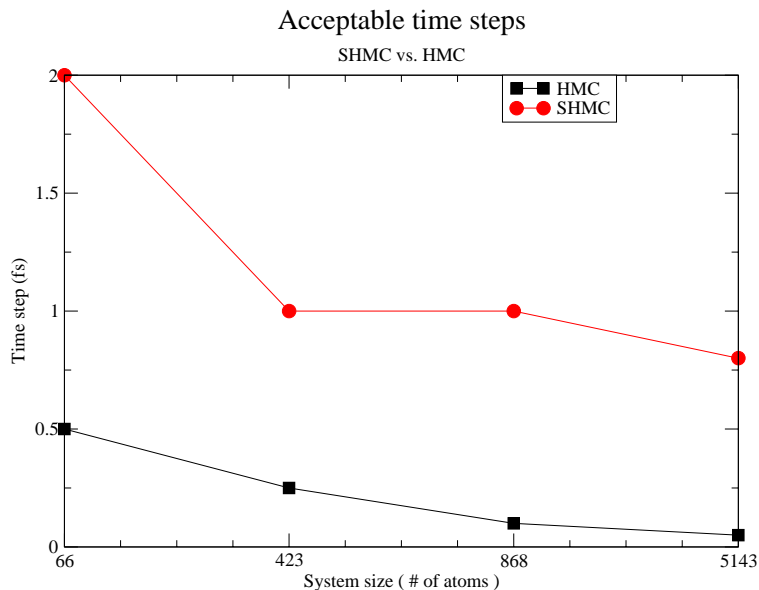


Figure 2: This graph contains the longest timestep for which each algorithm produced an acceptable timestep. A timestep was acceptable if the median acceptance rate was greater than 97.5%.

higher than 97%. For Alanin, SHMC does four times better than HMC with timesteps of 2 fs and 0.5 fs respectively. For Melittin+water SHMC's 0.8 fs timestep is 16 times that of HMC's 0.05 fs timestep. Based on the analysis of the previous section, we believe that as the size of the system increases, so will the difference in acceptance rates of SHMC and HMC.

In addition, our experiments showed that SHMC maintained higher acceptance rates versus HMC as the timestep for a particular molecule was increased. Figure 4 shows the results for Alanin, TIP3P water, Melittin, and Melittin+water.

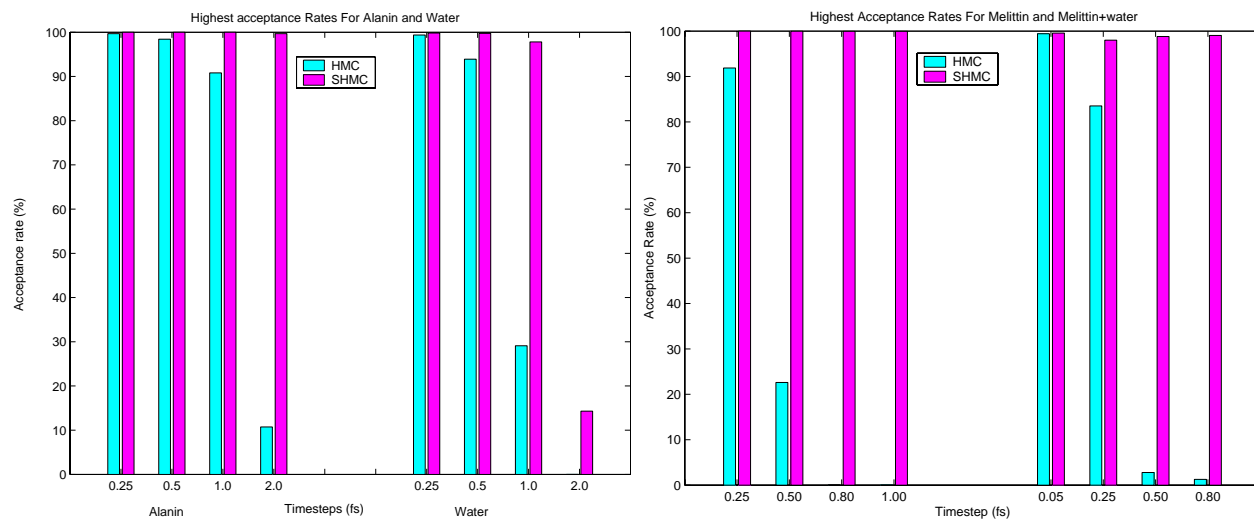


Figure 3: The bar graph on the left shows the best acceptance rates of Alanin (66 atoms), TIP3P (423 atoms), Melittin (868 atoms), and Melittin+water (5123 atoms).

While Alanin is a very small molecule, only 66 atoms, the results are telling. HMC does well for timesteps up to 1 fs, but decreases rapidly at 2 fs. SHMC manages near perfect acceptance rates all the way through. Melittin contains 868 atoms, a full order of magnitude greater than that of Alanin. However, the results for SHMC are nearly identical for both molecules. The largest system, Melittin+water is where we see the biggest difference. At 0.5 fs, HMC barely manages a 5% acceptance rate. SHMC gets 97% or higher all the way through 0.8 fs and even has a respectable 52% acceptance for 1 fs.

5 Future Work

We want to test our methods for the sampling of the β hairpin folding in explicit water, similarly to [20]. This will be a real test of the sampling properties of the method. Possible improvements to the method are using some tempering technique, cf. [16], or combining it with J-Walking. The MD simulation protocol can be improved by using multiple time stepping and exploiting fast electrostatics methods. This would require an extension to the shadow Hamiltonian, currently only defined for single time stepping integrators. Finally, the major question remains whether the perturbation to the Hamiltonian implied by using the shadow Hamiltonian matters in practice for sampling the canonical system. A rigorous theoretical answer is beyond reach at the moment, and we plan to explore it in simpler models and gathering more experimental evidence.

References

- [1] M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids*. Clarendon Press, Oxford, New York, 1987. Reprinted in paperback in 1989 with corrections.
- [2] B. J. Berne and J. E. Straub. Novel methods of sampling phase space in the simulation of biological systems. *Curr. Topics in Struct. Biol.*, 7:181–189, 1997.
- [3] A. Brass, B. J. Pendleton, Y. Chen, and B. Robson. Hybrid Monte Carlo simulations theory and initial comparison with molecular dynamics. *Biopolymers*, 33:1307–1315, 1993.
- [4] M. E. Clamp, P. G. Baker, C. J. Stirling, and A. Brass. Hybrid Monte Carlo: An efficient algorithm for condensed matter simulation. *J. Comp. Chem.*, 15(8):838–846, 1994.

- [5] M. Creutz and A. Gocksch. Higher-order hybrid monte carlo algorithms. *Physical Review Letters*, 63(1):9–12, 1989.
- [6] A. Crisanti, M. Falcioni, and A. Vulpiani. On the effects of an uncertainty on the evolution law in dynamical systems. *Physica A*, 160:482–502, 1989.
- [7] K. A. Dill and H. S. Chan. From levinthal to pathways to funnels. *Nature Struct. Biol.*, 4:10–19, 1997.
- [8] A. Fischer, F. Cordes, and C. Schütte. Hybrid Monte Carlo with adaptive temperature in a mixed-canonical ensemble: Efficient conformational analysis of rna. Technical Report SC 97-67, Konrad-Zuse-Zentrum für Informationstechnik Berlin, Dec. 1997. Available via <http://www.zib.de/bib/pub/pw/>.
- [9] J. M. Haile. *Molecular Dynamics Simulation*. John Wiley and Sons, 1992.
- [10] J. A. Izaguirre, Q. Ma, T. Matthey, J. Willcock, T. Slabach, B. Moore, and G. Viamontes. Overcoming instabilities in Verlet-I/r-RESPA with the mollified impulse method. In T. Schlick and H. H. Gan, editors, *Proceedings of 3rd International Workshop on Methods for Macromolecular Modeling*, volume 24 of *Lecture Notes in Computational Science and Engineering*. Springer-Verlag, Berlin, New York, 2002.
- [11] A. D. Kennedy and B. Pendleton. Acceptances and autocorrelations in hybrid monte carlo. *Nuclear Physics B (Proc. Suppl.)*, 20:118–121, 1991.
- [12] A. R. Leach. *Molecular Modelling: Principles and Applications*. Addison-Wesley Longman, Reading, Massachusetts, July 1996.

- [13] T. Matthey, A. Ko, and J. A. Izaguirre. ProtoMol, an object-oriented framework for prototyping novel applications of molecular dynamics. Submitted to the 2003 International Conference on Software Engineering, 2002.
- [14] R. M. Neal. Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, university of Toronto, 1993. papers available from <http://www.cs.toronto.edu/~radford/papers-online.html>.
- [15] J. M. Sanz-Serna and M. P. Calvo. *Numerical Hamiltonian Problems*. Chapman and Hall, London, 1994.
- [16] C. Schutte, A. Fischer, W. Huisinga, and P. Deuffhard. A direct approach to conformational dynamics based on hybrid Monte Carlo. *J. Comput. Phys*, 151(1):146–168, May 1, 1999.
- [17] R. D. Skeel. Integration schemes for molecular dynamics and related applications. In M. Ainsworth, J. Levesley, and M. Marletta, editors, *The Graduate Student's Guide to Numerical Analysis*, SSCM, pages 119–176. Springer-Verlag, Berlin, 1999.
- [18] R. D. Skeel and D. J. Hardy. Practical construction of modified Hamiltonians. *SIAM J. on Sci. Comput.*, 23(4):1172–1188, Nov. 2001.
- [19] T. C. Terwilliger and D. Eisenberg. The structure of melittin. *J. Biol. Chem.*, 257:6010, 1982.
- [20] R. Zhou, B. J. Berne, and R. Germain. The free energy landscape for β hairpin folding in explicit water. *Proc. Natl. Acad. Sci. USA*, 98:14931–14936, 2001.