Thermodynamics and Kinetics of Large-Time-Step Molecular Dynamics

Francesco Rao[a]* and Martin Spichty[b]*

Molecular dynamics (MD) simulations provide essential information about the thermodynamics and kinetics of proteins. Technological advances in both hardware and algorithms have seen this method accessing timescales that used to be unreachable only few years ago. The quest to simulate slow, biologically relevant macromolecular conformational changes, is still open. Here, we present an approximate approach to increase the speed of MD simulations by a factor of ~4.5. This is achieved by using a large integration time step of 7 fs, in combination with frozen covalent bonds and look-up tables for nonbonded interactions of the solvent. Extensive atomistic MD simulations for a flexible peptide in water show that the approach reproduces the peptide's equilibrium conformational changes, preserving the essential properties of both thermodynamics and kinetics. Comparison of this approximate method with state-of-the-art implicit solvation simulations indicates that the former provides a better description of the underlying free-energy surface. Finally, simulations of a 33-residue peptide show that these fast MD settings are readily applicable to investigate biologically relevant systems. © 2011 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.21990

Introduction

Elucidating the mechanisms responsible for cellular function at atomic resolution is necessary to understand living organisms at a fundamental level.[1] Functions such as signaling and molecular transport are achieved by cascades of molecular processes, for example, chemical reactions, formation of protein–ligand and protein–protein complexes, and (large-scale) conformational changes.[2] In principle, the investigation of those molecular processes is directly accessible from molecular dynamics (MD) computer simulations.[1,3] Two problems have mostly limited the applicability of computational approaches: (i) the timescale gap between the nanosecond sampling capabilities of atomistic molecular mechanics simulations and the micro-to-millisecond biological relevant transitions; (ii) the accuracy of current force-fields to describe the atomic interactions.[4] It turns out that these two limitations are usually tightly interlaced. For example, while quantum mechanical approaches would provide a first-principles description of atomic interactions, only short molecular trajectories of the order of pico-seconds can be produced. Conversely, coarse-grained models speed up dramatically the calculations, allowing for microsecond trajectories at the expense of accuracy and loss of information (e.g., atomic resolution).

The quest for a balance between statistical errors (due to limited sampling) and systematic errors (due to Hamiltonian approximations) have led to different approaches depending on the application: coarse-grained models to obtain qualitative descriptions of protein folding and association;[5–7] implicit representations of the solvent to study protein conformational changes;[8–10] classical all-atom representations to gain atomistic insights into protein function and dynamics;[11] mixed quantum classical (QM/MM) approaches to investigate enzymatic catalysis,[12,13] and combinations thereof in multiscale simulations.[14] Among them, classical MD simulations are the most widely used because they provide an acceptable compromise between performance and accuracy. Highly efficient implementations have become available that allow the use of multiple CPUs to speed up the simulation. With the advent of new massively parallel machines, first examples of microsecond-long all-atom simulations of proteins[11,15–18] have recently become available. Though encouraging, these performances are still out of reach for the largest fraction of the scientists in the field.

A simple way to speed up conventional MD simulations is to increase the time step for the numerical integration of the equation of motion.[19] The maximum step size that will yield a stable simulation (i.e., conservation of energy) is proportional to the period of the fastest fundamental modes of the system's motion. The fastest motions are usually the vibrations of the covalent hydrogen bonds, which limits the integration time step to 1 fs. Algorithms, that are now standard procedure, eliminate these degrees of freedom allowing an integration time step of

[a] *F. Rao*
Freiburg Institute for Advanced Studies, School of Soft Matter Research,
Albertstrasse 19, 79104 Freiburg im Breisgau, Germany
E-mail: francesco.rao@frias.uni-freiburg.de
Fax: +49 0761 203 97451

[b] *M. Spichty*
Équipe Différenciation et Cycle Cellulaire, Laboratoire de Biologie Moléculaire de la Cellule, UMR 5239 (CNRS/École Normale Supérieure de Lyon/Université Claude Bernard Lyon 1), 46 allée d'Italie, 69364 Lyon cedex 07, France
E-mail: martin.spichty@ens-lyon.fr
Fax: +33 0472 728 080

Contract/grant sponsor: Pôle Scientifique de Modélisation Numérique (PSMN); Contract/grant sponsor: Grand Eqipment National de Calcul Intensif; Contract/grant numbers: x2010076381 and x2011076381; Contract/grant sponsor: Excellence Initiative of the German Federal and State Governments (F.R.)

2 fs.[20–22] It is even possible to use larger time steps when freezing the geometry of flexible groups[17] and eventually, adding masses to the hydrogen atoms to slow down their bending motions.[23,24]

As a result of these changes to the Hamiltonian, the dynamics of the system is modified. Beside these enforced changes, there are also hidden ones that come with the use of large time steps: commonly used symplectic integrators such as the leap-frog integrator do not sample the correct Hamiltonian but rather a perturbed, shadow Hamiltonian due to the finite step size of the numerical integration[25]: the larger the time step, the larger the perturbation. Interestingly, though large-time-step MD simulations have increasingly been used in the last years, an open question is how the sum of these enforced and shadow changes influences the thermodynamics of biomolecular systems and their kinetics on a large microsecond-timescale.

One limitation in assessing this point is the lack of a common benchmark and the difficulty to have sufficient sampling with both conventional and more approximate setups. To this aim, an interesting class of systems is given by $(GlySer)_x$ peptides. They have been used in experiments for quite some time as flexible linkers.[26,27] Short stretches of this peptide represent an interesting ground for theoretical studies. Showing a dynamic interchange between compact and extended structures, they are characterized by a complex free-energy surface[28,29] and, at the same time, they can be extensively simulated due to the small number of atoms.

In this work, a straightforward approach to speed up (all-atom) MD simulations is presented and compared to conventional calculations of a four residue peptide, $(GlySer)_2$, solvated in explicit water. The approximate approach includes the use of a large 7-fs integration time step combined with frozen covalent bonds and 4-amu protein hydrogen masses as well as look-up tables for nonbonded interactions of the solvent.[30] The latter is a recent development which replaces the direct calculation of atomic forces for nonbonded interactions by the use of look-up tables, "trading floating point operations for memory use".[30] An illustrative example of a 33-residue protein (human Pin1 WW domain mutant)[4,31] simulated with the 7-fs setup, also demonstrates the applicability of the approach for studying the dynamics of biologically relevant systems.

Methods

Conventional MD

All MD simulations were performed with the program CHARMM[32]; the input file can be found as Supporting Information. Trajectories were analyzed with the program WORDOM.[33,34]

The peptide Gly-Ser-Gly-Ser with ionic N- and C-terminus, $(GlySer)_2$, was modeled with the all-atom topology and the parameter set 22 using CMAP correction terms[35] for the peptide backbone. It was solvated using a truncated octahedral box of 708 TIP3P water molecules. A cutoff of 12 Å was used for the nonbonded interactions (CTOFNB) with force shifting for electrostatic terms and potential shifting for VDW terms. The nonbonded list was updated heuristically, that is, updating occurred if any atom moved by more than (CUTNB − CTOFNB)/2 since the last list-update, where CUTNB is the cutoff for the nonbonded list generation. For this conventional setup, the speed of CHARMM's MD code is not very sensitive to the value of CUTNB, as long as it ranges between 14 and 16 Å (see Supporting Information, Fig. S1). For the production runs we used CUTNB = 14 Å. The NPT ensemble was sampled with CHARMM's extended Nosé–Hoover algorithm using a reference pressure and temperature of 1 atm and 300 K, respectively. The mass of the pressure and thermal pistons were set to the standard values as recommended in the software documentation (PMASS = 500 amu, TMASS = 1000 kcal mol^{-1} ps^2). The equation of motion was integrated with the CHARMM's leap-frog integrator and a time step of 2 fs. Covalent hydrogen bonds were constrained to their parameter values with SHAKE.[20] Five independent simulations (each of length 1.4 μs) for a total of 7 μs were performed. Atomic coordinates were saved every 280 fs for a total of 5×10^6 snapshots per run.

With CHARMM's implementation of table look-ups for nonbonded interactions,[30] the calculations could have been performed about 1.4 times faster (see Table 1). As the aim here was to obtain a reference MD simulation for conventional settings, we decided not to use table look-ups in this first instance. Note, however, that with table look-ups the relative error in the atomic forces is small (of the order of 10^{-3}, see Supporting Information, Fig. S2, and Ref. 30). For the $(GlySer)_2$ peptide, no influence on the time evolution of the radius of gyration could be noticed for less than 420 fs of MD (see Supporting Information, Fig. S3). This observation holds for both time steps, 2 fs and 7 fs. We therefore decided to use table look-ups for the following fast simulation setup to save CPU time.

Fast settings

Calculations with an increased integration time step of 7 fs were performed by setting the hydrogen masses of the peptide to 4 amu and constraining all covalent bonds with SHAKE[20] to their parameter values (i.e., the reference distance of the harmonic bond term) with a tolerance of 10^{-6} (the default value is 10^{-9}). Table look-ups were used for nonbonded interactions of the solvent with the parameter TABI set to 10.[30] With the large time step of 7 fs, the atomic movements in each MD step are larger than with 2 fs. For this reason, the CUTNB was changed from 14 to 20 Å. In terms of speed, enlarging the span between CTOFNB and CUTNB is beneficial for the used (heuristic) algorithm of updating the nonbonded list. Furthermore, table look-ups profit from a large CUTNB as well (see Supporting Information, Fig. S1).* All other settings are the same as for the conventional runs (described in the previous section); including the simulation length (5×1.4 μs) and saving frequency. Finally, we produced for the fast settings a series of MD simulations in the NVE ensemble to measure the energy drift as a function of the integration time

*The use of table look-ups and the increase of the CUTNB-value require more RAM. For the systems studied in this work, the additional amount of required memory with the program CHARMM is, however, below 0.8 GB with respect to the conventional settings.

Table 1. Summary for the explicit water simulations of the $(GlySer)_2$ peptide and the Pin1 WW domain mutant.

Setup name	Integration time step (fs)	Optimal CUTNB (Å)	Covalent nonhydrogen bonds	Hydrogen mass in protein (amu)	Table look-ups	Relative speed on eight cores
$(GlySer)_2$						
Conventional	2	15	Flexible	1.008	No	≡ 1.0
Conventional (look-ups)	2	18	Flexible	1.008	Yes	1.4 (1.4)
Fast (no look-ups)	7	17	Rigid	4.000	No	3.1 (3.1)
Fast	7	20	Rigid	4.000	Yes	4.6 (4.6)
Pin1 WW domain						
Conventional	2	16	Flexible	1.008	No	≡ 1.0
Conventional (look-ups)	2	19	Flexible	1.008	Yes	1.4 (1.4)
Fast (no look-ups)	7	18	Rigid	4.000	No	2.9 (2.9)
Fast	7	22	Rigid	4.000	Yes	4.4 (4.3)
Pin1 WW with Ewald						
Conventional+PME	2	16	Flexible	1.008	No	0.7 (0.7)
Fast+PME	7	21	Rigid	4.000	Yes	3.4 (3.4)

For the speed-up comparison, we performed short runs of 180 ps [$(GlySer)_2$] and 45 ps (Pin1 WW) using the optimal value of CUTNB for each setup. The speed-up is calculated in terms of CPU time consumed by two quad-core processors [2 × Intel(R) Xeon(R) E5472 (3.00 GHz) of a single SGI ALTIX ICE cluster node]. The values in parentheses are the speed-up factors for the wall-clock time, that is, the total elapsed time.

step. We first equilibrated the system for 0.5 ns at NVT, and then analyzed 1 ns of NVE-MD as described in Ref. [30].

Implicit solvation models

Equilibrium ensembles of the $(GlySer)_2$ peptide were also sampled with four different implicit solvation models of CHARMM applying their recommended standard settings. For EEF1[36] its specific polar-hydrogen topology 19 was used, for FACTS,[37] GBSW,[38] and GBMV[39] the all-atom topology 22 was used. GBSW (self-consistent) was performed with CMAP correction terms using the corresponding optimized parameters.[40] For GBMV (Type II) we performed two simulations with and without CMAP correction terms. Constant temperature was achieved through Langevin dynamics with a friction coefficient of 4 ps^{-1} (for all atoms). The small value of the friction allowed on one side to have faster convergence but, on the other side, precludes a direct comparison of the kinetics with explicit water simulations (this comparison would, in any case, not be very meaningful due to the large discrepancy in the free-energy landscape between implicit and explicit water simulations, see "Results"). SHAKE constraints were applied to the hydrogen bonds. The time step for the leap-frog integrator was 2 fs. From five independent simulations (each of length 224 ns), a total of 4×10^6 equilibrium structures were extracted (i.e., the saving frequency was 280 fs).

Pin1 WW domain mutant

The initial structure of the MD simulation corresponded to the native β-sheet conformation of the Pin1 WW domain mutant Fip35 (variant 23 of Ref. [31]); the coordinates were obtained directly from the authors of Ref. [4]. The protein was simulated for 27 ns in explicit water using periodic boundary conditions with truncated octahedral geometry. In addition to the protein, the simulation cell consisted of 5845 TIP3 water molecules, five chlorine anions, and two sodium cations (to neutralize the system with an ionic strength of 0.1 M). Simulations were performed with the conventional settings and the fast settings as described for the $(GlySer)_2$ peptide. The last 25 ns of the trajectory were analyzed (2 ns served as equilibration). In addition to the above finite-cutoff setup for electrostatic interactions, we tested a setup with the smooth particle-mesh Ewald (PME) method.[41] CHARMM's standard PME parameters were used with 64 grid points in each direction (i.e., the grid spacing is 0.95 Å). MD simulations were performed for the conventional and fast settings, the length was the same as for the finite-cutoff setup. Furthermore, we measured the energy drift using NVE simulations of length 0.5 ns (pre-equilibrated for 0.5 ns at NVT).

Results

In the following three sections, the influence of the fast simulation settings (i.e., 7-fs integration time step combined with frozen covalent bonds, increased hydrogen masses and table look-ups) for the $(GlySer)_2$ peptide on both thermodynamics and kinetics is presented. In the fourth section, we present a comparison with implicit solvation models. The last section of the "Results" is reserved to the application of the fast settings to a larger protein system (i.e., a 33-residue Pin1 WW domain mutant).

Stability and speed-up

In all of the five simulations with fast settings, the total energy as well as temperature and cell side (of the simulation box) for the $(GlySer)_2$ peptide show a stable behavior in the whole time range of 1.4 μs (see Supporting Information, Fig. S4). When comparing with the conventional settings both the potential energy and the cell size of the 7-fs runs show different values, being larger by 3 and 0.5%, respectively. Because of the enforced changes to the Hamiltonian (SHAKE, heavy-masses of hydrogen atoms) as well as the inaccurate integration of the equation of motion in the fast runs, a different Hamiltonian is sampled with respect to the conventional runs. A change in the potential energy and kinetic energy is therefore expected. Freezing of all covalent bonds, for example, decreases the potential energy (positive harmonic bond terms are removed) and lowers the kinetic energy (smaller amount of degrees of freedom), so that the total energy is lowered by about 15 kcal mol^{-1}. On the other side, the equation of motion is less accurately integrated with the large integration time step, so that unfavorable interactions (such as overstretching of harmonic angles, electrostatic repulsion, and steric clashes) occur with higher probability than with the conventional time step. Thus, the 7-fs time step samples a perturbed Hamiltonian with significant higher potential energy. This increase in potential energy is comparable to sampling at a slightly higher temperature which would also explain the larger volume (i.e., lower density) of the simulation cell in the fast runs with respect to the conventional runs. This virtual temperature increase is about about 18 K: simulating the $(GlySer)_2$ system at 318 K with the fast settings but a conventional integration time step of 2 fs yielded the same potential energy and volume as in the fast runs at 300 K (see Supporting Information, Table S1). In terms of total energy, the increase in potential energy due to the inaccurate integration largely overcompensates the effect of the bond freezing, so that the total energy is about 150 kcal mol^{-1} higher than in the conventional runs.

We performed for the fast settings a series of MD simulations with different integration time steps, sampling the NVE ensemble. There is a drift in total energy which increases with increasing time step; the measured values are (K ns^{-1}): 1.49 (2.5 fs), 3.57 (4 fs), 6.02 (5 fs), 9.53 (6 fs), 35.72 (6.5 fs), and 169.12 (7 fs). The drift for the conventional settings (2 fs) is 0.08 K ns^{-1}. We note, that for the fast settings, a dramatic energy drift sets in for integration time steps close to 7 fs. This is not surprising because the inaccurate integration of the equation of motion with such large time steps leads to rapid heat accumulation (e.g., due to steric clashes). The use of a thermostat is therefore mandatory to dissipate the heat. The benefit of thermostats "to avoid steady energy drifts caused by the accumulation of numerical errors during MD" is well known.[42] The value of the energy drift (in the NVE ensemble) is often used as a measure for the quality of a MD simulation code. It is therefore an interesting question how a drastic drift (such as the one for the 7-fs-time-step simulation) influences the thermodynamic and kinetic properties of the system (in the NVT ensemble).

Further increase of the time step to 8 fs or more, leads to several numerical instabilities, that is, SHAKE errors. This suggests that for the molecular systems studied in this work, 7 fs is the largest time step achievable with the applied Hamiltonian modifications (i.e., frozen covalent bonds, heavy hydrogens) and with the used leap-frog integrator. Specialized integration schemes could be used to further extend the maximal integration time step of the simulation system (examples include Refs. [19, 43]) or of subparts of the Hamiltonian (e.g., long range electrostatic interactions) as it is done in multiple-time-step dynamics.[44]

With the fast settings presented herein, the MD is 4.6 times faster in terms of wall-clock (elapsed) time than with the conventional settings. Without table look-ups for the nonbonded interactions of the solvent, this factor would still be 3.1, which is only slightly smaller than the ratio of the integration time steps, that is, 7 fs/2 fs.

Structural properties

The structural differences between the conventional and fast runs are investigated by analyzing the radius of gyration (R_g). Given the large flexibility of the peptide, R_g is a good observable for a global structural analysis.[26, 28] As shown in Figure 1, both conventional and fast simulations show very similar R_g probability distributions. The peptide is characterized by three distinct peaks, indicating the presence of compact ($R_g \sim 3.1$ Å), extended ($R_g \sim 4.2$ Å), and semiextended ($R_g \sim 3.7$ Å) conformations. The fast runs show a lower population of the compact conformations than the conventional runs. It should be noted, however, that the error in determining the population of the compact structures relative to the extended structures is high because the passage time between those basins is in the 10-ns-time range (see below), that is, there are "only" about 100 transitions from $R_g < 3.1$ Å to $R_g > 4.2$ Å (and back) in a simulation of length 1.4 μs. As a result, the error in the estimation of the relative population is high. Using the relative population, we can calculate the free-energy difference between the (semi-)extended structures ($R_g \geq 3.37$ Å) and the compact structures ($R_g < 3.37$ Å). Taking the mean value of the five independent MD-simulations, we find $G_{\text{compact}} - G_{(\text{semi}-)\text{extended}} = \Delta G = 0.47 \pm 0.05$ kcal mol^{-1} and $\Delta G = 0.54 \pm 0.03$ kcal mol^{-1} for the 2-fs and 7-fs runs, respectively. The error corresponds to the standard error of the mean. Thus, the mean values are indeed different but the error bars are overlapping. When performing Welch's t-test[45] with these numbers (see Supporting Information, section S7), a p-value of 0.31 is obtained which indicates a low statistical significance for the difference in the means: the probability that the free-energy difference for the fast and conventional runs is actually identical is 31%. In t-statistics, a probability of less than 5% (p-value of 0.05) is usually considered as statistically significant. Substantially more sampling, especially for the conventional runs, would be necessary to increase the certainty of having (or not having) a systematic deviation between the fast and conventional setup.

In addition to these marginal population differences, the R_g distribution for the fast runs shows a weak but systematic shift of the peaks toward smaller values of the order parameter R_g. This shift is due to the SHAKE constraints used for the nonhydrogen bonds in the fast runs. All bonds were frozen at their reference distance r_0 of the harmonic bond term $E_{\text{vib}} = k_b(r - r_0)^2$. For the CHARMM force-field, the value of r_0 is chosen in a way, that in combination with all other terms (e.g., nonbonded interactions), a value close to the experiment is reproduced.[46] For almost all nonhydrogen bonds of the $(GlySer)_2$ peptide, r_0 is about 1.01 times smaller than the experimental distance (or the value that is obtained in MD simulations without SHAKE constraints, see Supporting Information, Table S2). Thus, the fast

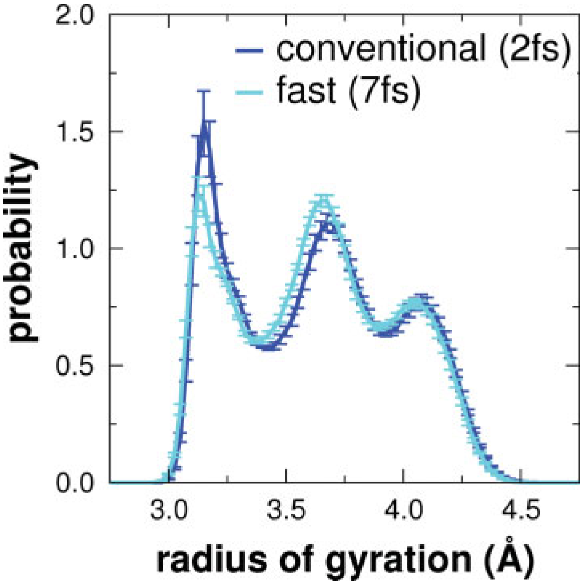


Figure 1. Radius of gyration probability distributions. Data for the conventional and the fast runs are shown in blue (2 fs) and light blue (7 fs), respectively.

runs (that constrain the nonhydrogen bonds to r_0) tend towards a slightly more compact protein than the conventional runs. For molecular processes where the molecule compactness is critical (e.g., diffusion processes), the fast simulation setup needs to be changed. The covalent bonds could be constrained, for example, to the bond lengths of a minimized reference structure (instead of the r_0-distances).

Protein kinetics

Peptide kinetics are studied in terms of first-passage-time (fpt) distributions.[47,48] This is the distribution of times to reach a given target structure from any sampled conformation different from the target. Arrival times depend on the definition of the target; in this case it is the compact peptide structure of Figure 2A. Hence, when a peptide conformation has an all-atom root-mean-square deviation smaller than 0.4 Å an arrival time is recorded.[†] In Figure 2B, fpt distributions of the conventional and fast runs are compared. The distributions reveal two-exponential kinetics indicating the presence of two relaxation processes. The fast exponential decay is less evident due to the superposition with the intrabasin fast relaxations to the target structure (as also observed in Ref. [48]). The two curves are almost in quantitative agreement but the 7-fs runs show the overall faster kinetics. This is a somewhat surprising result because the higher inertia of the protein hydrogen atoms could have potentially led to slower kinetics. We interpret this result as a consequence of the higher average energy of the system, comparable to running at a slightly higher temperature (see above). Fitting the slowest relaxation with an exponential function, we obtain a timescale of 18.6 and 11.1 ns for the conventional and the fast runs, respectively.

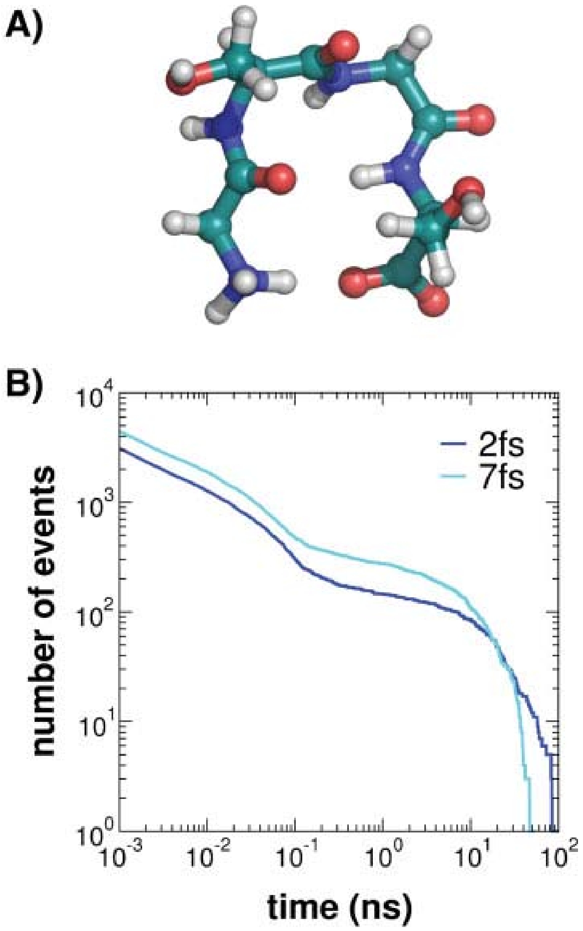


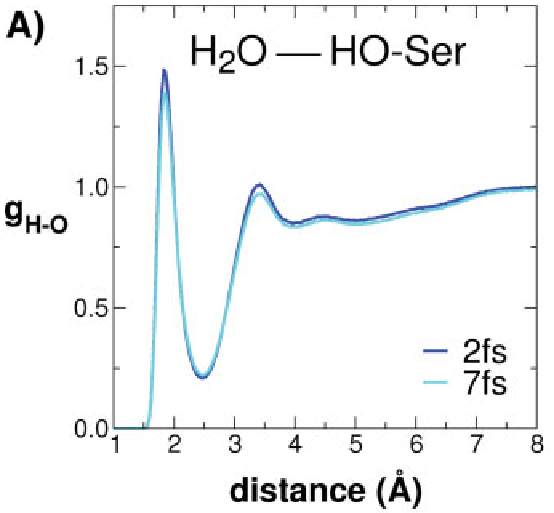
Figure 2. Analysis of the kinetic properties of the $(GlySer)_2$-peptide. (A) The target structure for the kinetic analysis. (B) First-passage-time distributions for the conventional (2 fs) and fast runs (7 fs). For the sampled structures of the trajectories, we recorded the time span the system required to make its first passage to the target structure. The figure shows the number of recorded events as a function of the time span.

Protein-solvent interactions

We now turn to the properties of the water molecules that directly interact with characteristic sites of the $(GlySer)_2$ peptide. The distribution of distances between the water oxygen atoms and the hydrogen atoms of the serine hydroxy groups shows a marked peak around 1.85 Å in both setups (Fig. 3A), which corresponds to the peptide's first solvation shell. In general, the shape of the radial distribution functions (RDF) is almost identical for the two simulation setups. Note that the molecular density is slightly smaller in the fast runs than in the conventional runs so that the RDF of the former is shifted to slightly lower probabilities with respect to the latter.

From a dynamic point of view, hydrogen-bond kinetics at the positively charged N-terminus, for example, show an average

[†] In principle, it is possible to perform this analysis also with a R_g-based criterion; but, in this case the definition of the target structure would include microstates from several different conformational basins, and the fpt distributions would be less well defined. This is not the case with the tight all-atom rmsd-based criterion.



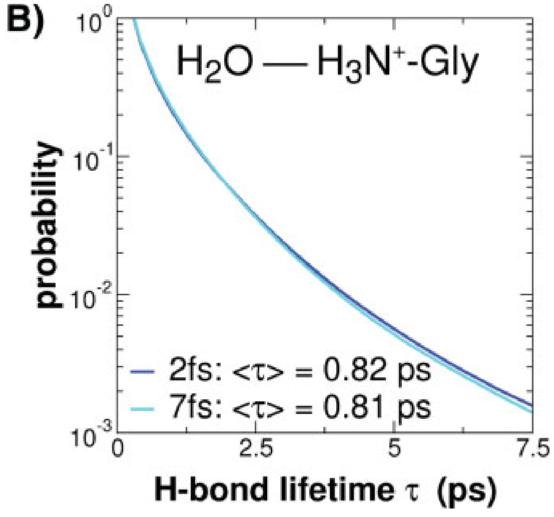


Figure 3. Characterization of protein–solvent interactions. (A) RDF between the polar hydrogen atoms of the serine side-chains and the oxygen atoms of the water molecules. (B) Life-time distribution of hydrogen bonds between the positively charge N-terminus and the water molecules.

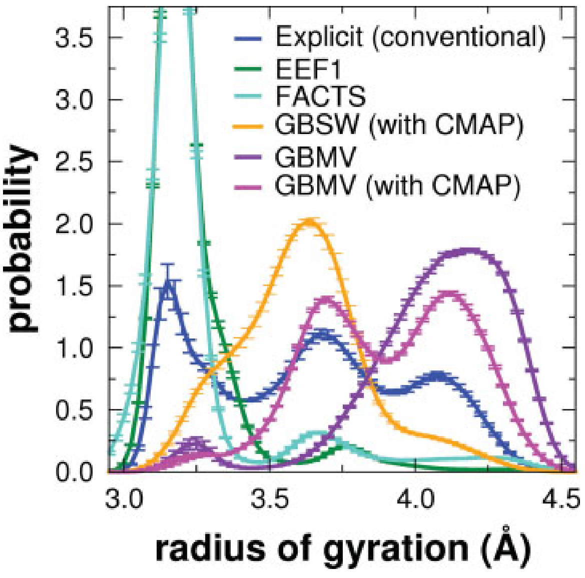


Figure 4. Comparison of R_g probability distributions for MD simulations with explicit water and implicit solvation models.

lifetime of 0.81 and 0.82 ps for the conventional and fast runs, respectively (Fig. 3B). These results indicate that protein–solvent interactions are not affected by the fast simulation settings.

Comparison with implicit solvation models

In this section, our previously obtained R_g distributions are compared to the ones coming from implicit solvent models. These simplified solvent representations are computationally more efficient at the expense of atomic resolution for the solvent. Results for four, among the most widely used, implicit solvation models implemented in CHARMM (EEF1,[36] FACTS,[37] GBSW[38] and GBMV[39]) are shown in Figure 4. The implicit models deviate significantly from the explicit water simulations. The computationally less expensive ones (EEF1 and FACTS) over-populate the compact conformations dramatically. It should be noted that those models do not include any CMAP correction. The more sophisticated models (GBSW, GBMV) with CMAP corrections, on the other hand, over-populate the extended conformations.

This comparison indicates that (i) reproducing the population balance between compact and extended structures of the peptide is not trivial and (ii) a description of the solvent at the particle level seems to be needed to recover the explicit-water free-energy landscape for this specific system. For these reasons, we further conclude that the $(GlySer)_2$ peptide can be regarded as an interesting test case for implicit solvation models being both non-trivial and computationally accessible.

Larger protein systems

Given the relatively small size of the $(GlySer)_2$ peptide, we wanted to test the fast simulation setup on a more complex system. We analyzed a MD run of 25 ns for the 33-residue Pin1 WW domain mutant Fip35.[4,31] In Figure 5A, the evolution of the C^α root-mean-square (rms) deviation (with respect to the starting structure) is shown for the fast and conventional simulation setup. The variance in the RMSD values is comparable, and both runs reach a plateau value of 1.5 Å.

The rms fluctuations of the C^α-atoms (Fig. 5B) show that the residues of the structural motifs (i.e., β-strands) exhibit lower fluctuations than the loops or the tails, as it is expected. Strikingly, there is almost no difference between the two setups, which indicates that the fast simulation setup fully conserves the protein's flexibility pattern of the conventional setup.

The fast settings can also be applied to a setup with smooth PME[41] treatment for long range electrostatic interactions. Again, the profile of the C^α rms fluctuations for the fast settings matches the ones obtained by conventional runs (the fluctuations with PME are marginally smaller than without PME). Both treatments

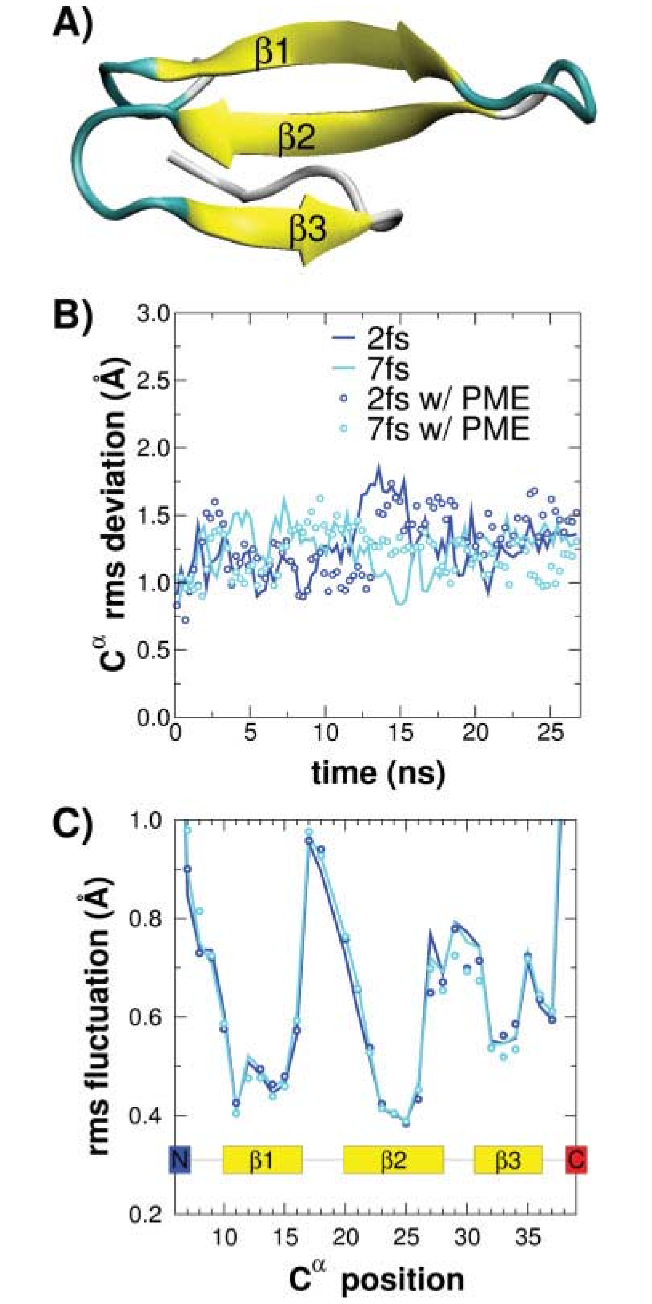


Figure 5. Results for the Pin1 WW domain mutant. (A) Illustration of the protein fold. (B) Running averages (with a time span of 280 ps) are shown for the rms deviation of the C^α atoms. (C) The rms fluctuations of the C^α atoms are shown as a function of the amino acid position. Note that, the numbering of this 33-residue protein runs from 6 to 39; there is a deletion at position 19. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

of electrostatics show similar energy drifts in NVE simulations, that is, 153 and 162 K ns^{-1} for finite-cutoff electrostatics and PME, respectively, using the fast settings. Thus, the drift for the Pin1 WW domain is similar to the one obtained for the $(GlySer)_2$ system (169 K ns^{-1}).

When comparing the required CPU time for the fast and conventional settings, we note, that the speed-up factor with PME treatment is slightly larger (4.8) than with finite-cutoff electrostatics (4.4; Table 1), which is a result of different computational costs for the calculation of the atomic forces.[‡]

Discussion

In spite of a less accurate integration and modified Hamiltonian, the fast simulation setup yielded, for the $(GlySer)_2$ peptide, comparable results to conventional simulations both in structural and dynamic terms. The gain in speed is 4.6 when table look-ups are used in the fast runs; without table look-ups the factor reduces to 3.1. Test runs on a larger system, a 33-residue protein, demonstrated that this protocol can be readily applied on complex molecular systems too.

Speed-up comes at the price of larger approximations. For this reason, it would be useful to estimate the extent of the systematic error (introduced by approximations) relative to the error provided by limited statistics, assessing the range of applicability for one method or the other. Formulated differently, it is important to know when it is pertinent to sample more but with a less accurate setup (e.g., the fast settings presented herein), or to sample less but with fewer approximations (e.g., the conventional setup). To address this question, we focused on the free-energy difference, ΔG, between the compact ($R_g < 3.37$ Å) and the (semi-)extended ($R_g \geq 3.37$ Å) structures. As a matter of fact, ΔG is a challenging quantity to estimate from MD trajectories, depending on both the accuracy of the model and sampling length. In Figure 6, ΔG is represented as a function of CPU time. We note that for short runs (less than 70 ns, $\approx$ 1% of the total CPU time) there is a significant bias in the estimate; that is, the error bars of the 2-fs estimates with less than 1% of CPU time do not overlap with the value of the full data (solid black line). Given the fact that, all simulations started from an extended conformation, this bias is not surprising. The deviation is also seen in the 7-fs setup but it disappears earlier (i.e., less CPU time is needed). For more than 1% of CPU time (i.e., where there is no obvious bias), the statistical error of the 7-fs estimate is less than half the error of the 2-fs estimate. The reason for the more rapid convergence of the 7-fs setup is twofold. First, given the faster kinetics of the 7-fs runs, equilibrium is reached in a smaller amount of time. Second, for a given amount of CPU time, the 7-fs runs can sample more.

The estimate from the 7-fs runs approaches a slightly different value of ΔG for the total amount of CPU time than the 2-fs runs (0.54 ± 0.03 vs. 0.47 ± 0.05 kcal mol^{-1}). This could represent a systematic error of the 7-fs approach (with respect to the 2-fs

[‡] In MD simulations, there are CPU costs that decrease linearly with increasing integration time step (such as the costs for the calculation of atomic forces). Conversely, there are costs that decrease less than linearly (or not at all) with increasing time step (e.g., the cost for the heuristic nonbonded list updating). The larger the fraction of "linearly-decreasing" costs, the larger the benefit in CPU time with increasing integration time step. PME requires more CPU time for the calculation of atomic forces than finite-cutoff electrostatics, and thus, the former features a larger fraction of linearly decreasing costs than the latter. As a result, PME profits more from the fast settings than finite-cutoff electrostatics in terms of CPU time reduction.

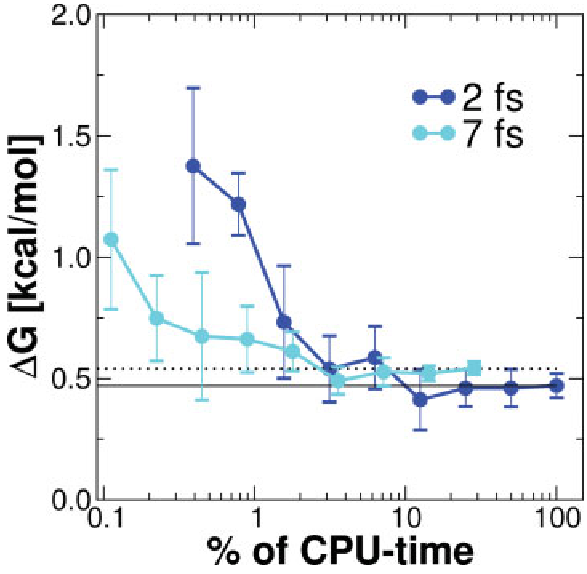


Figure 6. The estimate of a thermodynamic quantity as a function of invested CPU time. The CPU time for 100% corresponds to 7 μs of MD with a time step of 2 fs. For the comparison with the 7-fs runs, we do not take into account any speed-up that results from the use of table look-ups because they can be applied to both, 2-fs and 7-fs runs. ΔG is the free-energy difference between the compact and (semi-)extended structures of the $(GlySer)_2$ peptide. For a given point in the plot, we analyzed the corresponding percentage of sampling from the beginning of five independent MD simulations (that all started from the same extended structure). For example, for the 2-fs point at 50%, we took 700 ns from the beginning of each of the five conventional simulations which totals to 3.5 μs (=50% of 7 μs). The value of ΔG was determined from the mean of the five simulations, and the error bars were obtained from the standard error of the mean. Free-energy differences obtained from the entire data are marked as solid and dashed horizontal lines for the 2-fs and 7-fs runs, respectively. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

setup), which amounts to 0.07 ± 0.06 kcal mol^{-1}. Our observations suggest that the systematic error is usually significantly smaller than the statistical error due to limited sampling, at least in the case studied. For example, when performing 100 ns of 2-fs sampling (about 1.4% of the total CPU time), we find a statistical error in the estimate of ΔG of 0.23 kcal mol^{-1}, which is more than three times larger than the systematic error of 0.07 kcal mol^{-1}. Only when extensive sampling is performed, that is, for more than 1 μs of MD in the case of the $(GlySer)_2$ peptide, the statistical error becomes comparable to the systematic error. We conclude that the statistical error is the largest source of error. The small size of the systematic error for the large-time-step approach seems somewhat surprising. The use of too large integration time steps can introduce large systematic errors in the estimate of free energies (for example, an error of 2.5 kcal mol^{-1} for the confinement of a small protein[10]). However, these effects cancel out when conformational free-energy differences are considered.[10]

Conclusions

In this article, we investigated the applicability of a faster simulation protocol to sample the conformational space of peptides and proteins with MD simulations. The approach represents a perturbation of the correct Hamiltonian, combining a large integration time step of 7 fs with heavy protein hydrogens and frozen covalent-bonds. Together with table look-ups for nonbonded solvent–solvent interactions, we obtained a remarkable speed-up factor of ~4.5 with respect to conventional simulations. From our study, three main points have emerged.

First, the conformations sampled by the fast 7-fs runs are in quantitative agreement with conventional calculations in terms of structural distributions. This is not a trivial achievement. Implicit solvent representations showed significant deviations. Such models come with a 20–30 lower cost in CPU time than the fast explicit water setup but at the price of larger disagreements. These disagreements are system dependent; the $(GlySer)_2$ seems to be a very difficult test case. For the explicit water simulations, the most distinct discrepancy between conventional and fast runs is the slightly smaller peptide geometry observed for the latter. This is a result of the freezing of the covalent bonds in the fast runs, and not due to the large integration time step.

Second, despite the presence of higher masses on the hydrogen atoms, system kinetics are very similar in the two setups. For example, the double exponential kinetics of the peptide is conserved. However, fast runs are characterized by faster conformational relaxations. Because of integration inaccuracies with the large 7-fs time step, the average potential energy is increased, similar to an apparent increase in temperature. This suggests that, the system can overcome local energy barriers more easily. If more accurate kinetics is required, the temperature of the simulation thermostat can be lowered to correctly reproduce the value of the potential energy (at the temperature of interest). Test calculations are in preparation to address this point.

Third, comparison between systematic and statistical errors indicates that the latter is the major source of error when it comes to the estimation of structural properties from MD simulations. Consequently, a faster but more approximate simulation setup can be more accurate than a conventional setup when the latter is not able to extensively sample the conformational space, for example, more than 100 forward and backward transitions between the states of interest.

Large-time-step simulation setups, similar to the one presented herein, are applicable to many biomolecular simulation packages, including CHARMM,[32] GROMACS,[17] NAMD[49] and ACEMD.[24] At a small expense of accuracy, the presented approach speeds up the simulation by a factor of ~4.5. An illustrative 25-ns simulation of a WW domain protein shows that the approach is stable for biologically relevant systems. The introduction of even more approximate integration schemes to increase the time step will represent a viable method to study sampling-limited problems.

Acknowledgments

The authors thank Marco Cecchini and Brian B. Rudkin for helpful discussions; Peter Freddolino and Klaus Schulten for making available to us the input structure of the Pin1 WW domain mutant.

Keywords: biomolecular computer simulations • force field • all-atom model • speed-up • free energy • first-passage-time distribution

How to cite this article: F. Rao, M. Spichty, *J. Comput. Chem.* **2012**, *33*, 475–483. DOI:10.1002/jcc.21990

Additional Supporting Information may be found in the online version of this article.

[1] M. Karplus, J. Kuriyan, *Proc. Natl. Acad. Sci.* USA **2005**, *102*, 6679.

[2] K. Okazaki, S. Takada, *Proc. Natl. Acad. Sci.* USA **2008**, *105*, 11182.

[3] M. Karplus, J. A. McCammon, *Nat. Struct. Mol. Biol.* **2002**, *9*, 646.

[4] P. L. Freddolino, S. Park, B. Roux, K. Schulten, *Biophys. J.* **2009**, *96*, 3772.

[5] V. Tozzini, *Curr. Opin. Struct. Biol.* **2005**, *15*, 144.

[6] C. Clementi, *Curr. Opin. Struct. Biol. 2008*, *18*, 10.

[7] L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, S. J. Marrink, *J. Chem. Theory. Comput.* **2008**, *4*, 819.

[8] A. van der Vaart, J. P. Ma, M. Karplus, *Biophys. J.* **2004**, *87*, 562.

[9] M. Spichty, M. Cecchini, M. Karplus, *J. Phys. Chem. Lett.* **2010**, *1*, 1922.

[10] M. Cecchini, S. V. Krivov, M. Spichty, M. Karplus, *J. Phys. Chem. B* **2009**, *113*, 9728.

[11] D. M. Rosenbaum, C. Zhang, J. A. Lyons, R. Holl, D. Aragao, D. H. Arlow, S. G. F. Rasmussen, H. J. Choi, B. T. DeVree, R. K. Sunahara, P. Seok Chae, S. H. Gellman, R. O. Dror, D. E. Shaw, W. I. Weis, M. Caffrey, P. Gmeiner, B. K. Kobilka, Nature **2011**, *469*, 236.

[12] J. Gao, D. G. Truhlar, *Ann. Rev. Phys. Chem.* **2002**, *53*, 467.

[13] H. Senn, W. Thiel, *Top. Curr. Chem.* **2007**, *268*, 173.

[14] G. S. Ayton, W. G. Noid, G. A. Voth, *Curr. Opin. Struct. Biol.* **2007**, *17*, 192.

[15] K. Y. Sanbonmatsu, C. S. Tung, *J. Struct. Biol.* **2007**, *157*, 470.

[16] P. L. Freddolino, F. Liu, M. Gruebele, K. Schulten, *Biophys. J.* **2008**, *94*, L75.

[17] B. Hess, C. Kutzner, D. van der Spoel, E. Lindahl, *J. Chem. Theory Comput.* **2008**, *4*, 435.

[18] R. O. Dror, D. H. Arlow, D. W. Borhani, M. O. Jensen, S. Piana, D. E. Shaw, *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 4689.

[19] T. Schlick, E. Barth, M. Mandziuk, *Ann. Rev. Biophys. Biomol. Struct.* **1997**, *26*, 181.

[20] J. P. Ryckaert, G. Ciccotti, H. J. C. Berendsen, *J. Comput. Phys.* **1977**, *23*, 327.

[21] H. C. Andersen, *J. Comput. Phys.* **1983**, *52*, 24.

[22] S. Miyamoto, P. A. Kollman, *J. Comput. Chem.* **1992**, *13*, 952.

[23] J. Franklin, S. Doniach, *J. Chem. Phys.* **2005**, *123*, 124909.

[24] M. J. Harvey, G. Giupponi, G. De Fabritiis, *J. Chem. Theory Comput.* **2009**, *5*, 1632.

[25] R. D. Engle, R. D. Skeel, M. Drees, *J. Comput. Phys.* **2005**, *206*, 432.

[26] O. Bieri, J. Wirz, B. Hellrung, M. Schutkowski, M. Drewello, T. Kiefhaber, *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 9597.

[27] A. Möglich, K. Joder, T. Kiefhaber, *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 12394.

[28] F. Rao, *J. Phys. Chem. Lett.* **2010**, *1*, 1580.

[29] F. Rao, *J. Comput. Chem.* **2011**, *32*, 1113.

[30] L. Nilsson, *J. Comput. Chem.* **2009**, *30*, 1490.

[31] F. Liu, D. Du, A. A. Fuller, J. E. Davoren, P. Wipf, J. W. Kelly, M. Gruebele, *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 2369.

[32] B. R. Brooks, C. L. Brooks, III, A. D. Mackerell, Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, M. Karplus, *J. Comput. Chem.* **2009**, *30*, 1545.

[33] M. Seeber, M. Cecchini, F. Rao, G. Settanni, A. Caflisch, *Bioinformatics* **2007**, *23*, 2625.

[34] M. Seeber, A. Felline, F. Raimondi, S. Muff, R. Friedman, F. Rao, A. Caflisch, F. Fanelli, *J. Comput. Chem.* **2011**, *32*, 1183.

[35] A. D. Mackerell, Jr., M. Feig, C. L. Brooks, III, *J. Comput. Chem.* **2004**, *25*, 1400.

[36] T. Lazaridis, M. Karplus, *Protein Struct. Funct. Genet.* **1999**, *35*, 133.

[37] U. Haberthür, A. Caflisch, *J. Comput. Chem.* **2008**, *29*, 701.

[38] W. Im, M. S. Lee, C. L. Brooks, III, *J. Comput. Chem. 2003*, *24*, 1691.

[39] M. S. Lee, F. R. Salsbury, Jr., C. L. Brooks, III, *J. Chem. Phys.* **2002**, *116*, 10606.

[40] J. Chen, W. Im, C. L. Brooks, III, *J. Am. Chem. Soc.* **2006**, *128*, 3728.

[41] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, L. G. Pedersen, *J. Chem. Phys.* **1995**, *103*, 8577.

[42] P. H. Hünenberger, *Adv. Polym. Sci.* **2005**, *173*, 105.

[43] J. A. Izaguirre, S. Reich, R. D. Skeel, *J. Chem. Phys.* **1999**, *110*, 9853.

[44] M. E. Tuckerman, B. J. Berne, G. J. Martyna, *J. Chem. Phys.* **1991**, *94*, 6811.

[45] B. L. Welch, *Biometrika* **1947**, *34*, 28.

[46] L. Nilsson, The CHARMM forum: Why are bonds larger than their equilibrium value?, **2011**, Available at: http://www.charmm.org, Post 26576.

[47] S. F. Chekmarev, S. V. Krivov, M. Karplus, *J. Phys. Chem. B* **2005**, *109*, 5312.

[48] F. Rao, M. Karplus, *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 9152.

[49] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, K. Schulten, *J. Comput. Chem.* **2005**, *26*, 1781.

Received: 14 June 2011
Revised: 22 September 2011
Accepted: 8 October 2011
Published online on 19 December 2011