Conserving Energy During Molecular Dynamics Simulations of Proteins in Water

Douglas B. Kitchen, Fumio Hirata† and Ronald M. Levy*
Department of Chemistry
Rutgers University
New Brunswick, New Jersey 08903

Abstract

Molecular dynamics simulations have been carried out for a fully solvated protein (PTI in water) using Andersen's velocity version of the Verlet algorithm with internal constraints (the RATTLE algorithm). The accuracy with which the equations of motion are integrated has been analyzed for several different simulation conditions. The effects of various nonbonded interaction truncation schemes on the conservation of energy have been examined, including the use of atomic cutoffs, and residue (neutral group) cutoffs. The use of a smoothing function to eliminate the discontinuities in the potential at the cutoff leads to a significant improvement in the accuracy with which the equations of motion are integrated.

Introduction

Molecular dynamics simulations of proteins provide the most detailed theoretical method available for studying their structural, dynamical, and thermodynamic properties. With this method, Newton's equations of motion are solved step by step for all the atoms of the system. Due to the large number of atoms in a protein these calculations are computationally intensive, and for this reason the earliest molecular dynamics simulations of proteins were carried out in vacuo. With the vastly increased speed of supercomputers, it has become possible to include the solvent environment explicitly in the simulations. The current generation of simulations can handle proteins of molecular weight up to \( \sim 20,000 \) solvated by several thousand water molecules. The technology has progressed to the point where there are now commercial software packages available to perform these calculations. However, there are still fundamental questions regarding the accuracy that can be achieved under diverse simulation conditions. Many comparisons between the results of simulations and experimental probes of protein structure and dynamics have been reported (1-3). Differences between calculated and experimental results are most often attributed to the approximate nature of the potential functions, and there are ongoing major

†Current address: Department of Chemistry, Kyoto University, Kyoto Japan.

*Author to whom correspondence should be addressed.
efforts in many laboratories to improve the empirical potentials. An additional source of error in simulations of these very large systems can arise from inaccurate integration of the equations of motion. Such errors can result from the integration algorithm employed, the time step used or, very commonly, from the way in which long range interactions are treated. The analysis of these effects has most frequently been reported in the context of liquid state simulations of uncharged homogeneous systems (4,5). The evaluation of the relative accuracies of different integration schemes as applied to proteins in vacuo has also been reported (6-8). Calculating trajectories accurately for solvated proteins presents additional problems beyond those for pure liquids or proteins in vacuo. This analysis is beginning to receive attention in the literature (9). In this paper we review the problem of the accurate integration of the equations of motion for a solvated protein, taking as an example the simulation of the dynamics of PTI in water.

Errors in the integration of the equations of motion for these complex systems can be manifest in several ways: for example large fluctuations in the total energy and/or large temperature increases over the course of the trajectory. If, in the microcanonical ensemble, the equations of motion were integrated with no error, the total energy would be conserved and no fluctuation of the total energy would be observed. In practice, the total energy does fluctuate during simulations within the \((N,V,E)\) ensemble. One standard measure of the accuracy of a simulation is given by the ratio of the fluctuations in the total energy, \(\langle \Delta E^2 \rangle^{1/2}\) to the average of the total energy \(\langle E \rangle\) (5). For liquid state simulations, a rule of thumb often employed is that \(\langle \Delta E^2 \rangle^{1/2}/\langle E \rangle \) should be less than 0.01% (5). Another criteria sometimes applied to assess the accuracy of the integration is given by the ratio of \(\langle \Delta E^2 \rangle^{1/2}/ \langle \Delta KE^2 \rangle^{1/2}\) where the denominator is the square root of the fluctuations in the kinetic energy of the system (6). It should be noted that the two criteria will scale differently with the size of the simulated system as the total energy is proportional to \(N\), the total number of atoms in the system, while the square root of the fluctuations in the kinetic energy increases as \(N^{1/2}\) (10). The fluctuation in the kinetic energy (and therefore the temperature) is proportional to the heat capacity of the system. Therefore when the error criteria is expressed as \(\langle \Delta E^2 \rangle^{1/2}/ \langle \Delta KE^2 \rangle^{1/2}\), the ratio ties the error in the integration to an experimental quantity since this value is directly related to the error in the estimate of the heat capacity of the system (10). Different authors have suggested acceptable values for this quantity for protein simulations ranging from 1-10% (6-8).

When simulating systems containing thousands of atoms, it is necessary to introduce a cutoff in the interaction pair potential at some distance \(R_c\), typical choices of \(R_c\) vary between 7 Å and 12 Å. The simplest way to implement such a cutoff is to abruptly truncate the potential at the cutoff distance, however this introduces impulsive forces into the simulation as the potential is discontinuous at the boundary. The total energy is not conserved when simulations are performed with a truncated potential and the net effect is a steady increase in the temperature of the system over time. To compensate for this effect in simulations which employ an abrupt spherical cutoff, the temperature of the system must be rescaled during the simulation. Several different schemes for rescaling the temperature (kinetic energy) have been proposed,
Energy Conservation & Molecular Dynamics

including those designed to approximate the simulation of a canonical ensemble (11,12). Two drawbacks of temperature scaling are: (1) it can mask errors resulting from inaccurate integration of the equations of motion and (2) for heterogeneous solute in solvent systems, uniform temperature scaling of the entire system leads to a disparity between solute and solvent temperatures (with the solvent hotter than the solute). With respect to masking errors, a prudent practice is to carry out trial simulations of the system under study without temperature scaling in order to assess the relative accuracy of the trajectory under the particular simulation conditions to be used for production runs and then to turn on temperature scaling. The problem of disparate temperatures between solute and solvent is not completely solved at present; a common practice is to scale the temperature of the solvent and solute separately.

An alternative approach for dealing with truncation effects is to alter the form of the potential function in a zone at the cutoff boundary so that the potential and its first derivative (the force) go to zero smoothly at the cutoff. The application of a smoothing function leads to much more stable trajectories in simulations of liquids (4,13). The smoothing procedure does not affect the short range liquid structure and has a very small effect on the internal energy but it does have a large effect on the long range orientational order. As the long range orientational order is related to the dielectric response of the liquid, the use of a smoothing function will, in general, alter the dielectric properties of the liquid (14). For pure liquid simulations, formulas have been derived which account for the effect of smoothing on the dielectric response (15), but analogous treatments are not available for heterogeneous systems such as simulations of proteins in water. Thus it must be kept in mind that the use of smoothing functions in simulations of solvated proteins may affect the dielectric properties in these simulations in ways that are not well understood at present.

In this paper we analyze the effect of various simulation conditions on the accuracy with which the equations of motion are integrated. We have studied the effect of such parameters as step size, frequency with which the nonbonded list is updated, and methods used to truncate the long range interactions, on the conservation of energy during the simulation. Several properties of the truncation scheme have been evaluated including the cutoff distance, the use of a smoothing function, and the use of atomic versus neutral group truncation.

Methods

All of the simulation results reported in this paper were carried out using the simulation package IMPACT (Integrated Modeling Program/Applied Chemical Theory) developed in our laboratory over several years (16-18). IMPACT is a program for simulations of condensed phase systems, including macromolecular solutes in solution. Integration of the equations of motion is accomplished within IMPACT using the velocity version of the Verlet algorithm with constraints (the RATTLE algorithm) of Andersen (19):

$$r(t + h) = r(t) + hv(t) + \frac{h^2 [r(r(t)) + \epsilon_{RR}(t)]}{2m}$$  \[I\]
\[ v(t + h) = v(t) + \frac{h [f(r(t)) + g_{RR}(t) + f(r(t + h)) + g_{RV}(t)]}{2m} \]

where \( r(t), v(t) \) are the atomic positions and velocities at time \( t \), \( f(r(t)) \) is the force on an atom at time \( t \), \( m \) is the atomic mass and \( h \) is the time step. The constraint forces due to the rattle algorithm are contained in \( g_{RR} \) and \( g_{RV} \), respectively (19). This form of the Verlet algorithm has two advantages over the conventional Verlet algorithm. First, the velocities are calculated explicitly at each step which is necessary for simulations in which temperature scaling is used. Temperature scaling can then be used to simulate the canonical ensemble while at the same time controlling the drift in the total energy during the course of a simulation. Also the velocity version of the Verlet algorithm has higher numerical precision associated with the way terms in powers of \( h \) are combined (19). To our knowledge, the RATTLE algorithm has not previously been applied to protein simulations. As reported in the following section, our experience has been that the algorithm performs well for simulations of solvated macromolecules.

The protein was placed in a box of water with dimensions 43.2 Å × 46.9 Å × 47.3 Å by removing all waters whose distance to any protein atom was within the sum of the overlap distances of the water oxygen and the protein atom. All the side chains which are ionized at neutral pH, were fixed in their ionic form. Chloride ions were used to replace sequentially the six water molecules with the greatest electrostatic potential due to the protein. The potential parameters used were the SPC values (20) for water; for the protein simulations the molecular mechanics force field of Kollman and coworkers was employed (21). In this force field, the electrostatic properties of sulfur atoms on e.g. methionine and cysteine residues are modeled using pseudo atomic groups with unit mass to represent the lone pair electrons, placed at a distance of 0.67 Å from the sulfur nucleus. We found that the very short bond length and light mass of this pseudo atomic group leads to unstable behavior in the integration of the equations of motion, and constraints were therefore always applied to bonds involving these groups during the molecular dynamics simulations. Additionally, in the simulations reported in the following section, constraints were applied to all the bonds of the protein. Prior to carrying out simulations at constant total energy for analysis, the system was first equilibrated during simulations lasting 5 ps at a temperature close to 300 K using temperature scaling.

**Treatment of Nonbonded Interactions**

There are several methods available for constructing nonbonded interaction lists. The simplest is to construct the list based on a strict interatomic cutoff distance criteria. A major difficulty with this method is that electrostatic interactions are treated in an inconsistent way when dipolar units located on pairs or groups of atoms are split during the generation of the list. In order to circumvent this problem, list generation can be based on the evaluation of the interactions between neutral groups rather than individual atoms. For proteins, a logical choice for the grouping of atoms into "neutral" groups is the amino acid residue; for charged residues, the group has a unit charge. In the following section, we refer to the neutral group
method of generating the nonbonded list as the residue cutoff method. When using a neutral group cutoff criteria, a decision must be made regarding how to calculate a cutoff distance between the groups. For pure water simulations, the choice of a cutoff distance between a pair of water molecules can be based either on the oxygen-oxygen distance as is usually the case, or on the distance between the center of mass of the two molecules, a more general method useful for simulating noneaqueous polyatomic solvents. For proteins, a criteria based on the distance between the centers of mass of the two residues is not a good choice because the spatial extent of the residues may lead to large interatomic interactions between two residues whose centers of mass separation is greater than a cutoff distance. In the program IMPACT, we have implemented a residue cutoff criteria based on a minimum distance between any two atoms of the residue pair; that is if any atom i on residue α is less than \(R_{\text{cut}}\) away from any atom j on residue β, then the interactions between all the atoms in residue α and all the atoms in residue β are included in the nonbonded interaction list. There is a very significant computational advantage to the use of a residue based nonbonded list as compared with an atomic nonbonded list associated with storage requirements. For simulations of proteins in water, the number of nonbonded interactions that must included with a cutoff of \(\sim 8 - 10\) Å is very large. Consequently, when using an atom based method for constructing nonbonded pair lists, the (two) arrays which store the partner information must be very large; whereas when the list is constructed using residue criteria, the partner lists are much smaller. For example the size of an atom-based list in the PTI/water simulation using a 9 Å cutoff is on the order of 1.5 million interactions, all of which must be stored explicitly. A residue based list would store only 200,000 pairs representing roughly the same number of atomic interactions. There is an additional large advantage to the use of a residue based list associated with the number of pairwise interactions that must be calculated when generating the list. Thus for water, looping over molecular (centers of mass) pairs rather than atom pairs saves a factor of nine in the generation of the nonbonded list. In IMPACT, the cutoff criteria used for the solvent (water) is always a molecular (residue) criteria, while for the protein various options are permitted.

As shown in the following section, the multiplication of the nonbonded interactions close to the cutoff by a smoothing function, leads to more stable trajectories. The smoothing function in IMPACT used for simulations is given by the 5th order polynomial (4)

\[
S(R) = 1 - (C_3 \delta^3 + C_4 \delta^4 + C_5 \delta^5)
\]

\[
C_3 = \frac{10}{(R_H^2 - R_L^2)^3}, \quad C_4 = \frac{-15}{(R_H^2 - R_L^2)^4}, \quad C_5 = \frac{6}{(R_H^2 - R_L^2)^5}, \quad \delta = (R^2 - R_L^2)
\]

This function and its first and second derivatives are continuous. The smoothing function tapers continuously from the value 1 at \(R_L\) to the value 0 at \(R_U\); the derivative of the smoothing function is zero at the lower and upper cutoff. The distance, \(R\), can be chosen within the IMPACT program to be either the distance between any two atoms, the distance between any two (neutral) groups of atoms specified by the centers of mass of the groups, or the distance between an atom (on the protein) and a neutral
group (a solvent water molecule). We have evaluated the effects of a smoothing function on the conservation of energy in simulations of protein in water (using a neutral group cutoff criteria for the water, and an atomic criteria for the protein). The non-bonded interaction energy between atom or group i and atom or group j is given by:

$$E_{ij} = S_{ij}E_{ij}$$  \[4\]

A shifting function may be obtained from this smoothing function by setting the lower cutoff ($R_{ij}$) to zero and this function is shown in Equation [5].

$$S(R) = 1 - \left( \frac{10R^3}{(R_{ij})^3} - \frac{15R^4}{(R_{ij})^4} + \frac{6R^5}{(R_{ij})^5} \right)$$  \[5\]

The fact that $S_{ij}$ may be a function of a molecule's center of mass complicates the derivatives of Equation [4]. Equation [6] shows the first derivative of [4] with respect to an arbitrary atom-atom distance ($R_{ij}$) when center of mass coordinates are used to calculate the smoothing function value.

$$\frac{\partial E_{ij}}{\partial R_{ij}} = S_{ij} \frac{\partial e_{ij}}{\partial R_{ij}} + \frac{\partial S_{ij}}{\partial R_{ij}} E_{ij}$$  \[6a\]

$$\frac{\partial R_{ij}}{\partial x_i} = \frac{m_r}{m_T} (2) \left( \frac{\Delta X_{ij}}{R_{ij}} \right)$$  \[6b\]

$$\frac{\partial S_{ij}}{\partial R_{ij}} = \frac{\partial S_{ij}}{\partial R_{ij}} \frac{\partial R_{ij}}{\partial R_{ij}}$$  \[6c\]

In Equation [6], $R_{ij}$ is the distance between the center of mass of group I and group J and $\Delta X_{ij}$ is the difference between the x coordinates of atom i' in group I and atom j' in group J. Center of mass coordinates are often used to calculate the smoothing function distance and this means that the second term in [6a] must also be used in the calculation of the derivatives. This additional term complicates the programming of the derivative since $E_{ij}$ is the total group to group potential energy and therefore the nonbonded list must be looped over twice for atoms in the smoothing region, in order to calculate the $E_{ij}$ first, and then to calculate the forces using the $E_{ij}$.

The relative timings of the various cutoff approximations differ depending on the system which is being calculated. However, for simulations of proteins in water, the relative timings are approximately equal for the three methods: atomic cutoff/no smoothing, atomic cutoff/smoothing and residue cutoff. This is due to the fact the most time consuming part of the calculation comes from the water-water interactions. For vacuum protein simulations the differences between methods are much more substantial. Residue cutoffs include more interactions and therefore take much
Energy Conservation & Molecular Dynamics

Table I
Simulations of PTI in Water

<table>
<thead>
<tr>
<th>Step</th>
<th>List Update</th>
<th>Cutoff Radius (Å)</th>
<th>Truncation Scheme</th>
<th>Shake/Rattle</th>
<th>Equilibration Conditions</th>
<th>(&lt;\Delta E^2&gt;^\text{a})/&lt;E&gt; (%)</th>
<th>(&lt;\Delta E^2&gt;^\text{b})/&lt;\Delta E^2&gt; (%)</th>
<th>Temperature (°K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTI (in water)</td>
<td>1</td>
<td>5</td>
<td>8.0</td>
<td>no smoothing</td>
<td>y</td>
<td>a</td>
<td>0.811</td>
<td>447.0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>5</td>
<td>8.0</td>
<td>no smoothing</td>
<td>y</td>
<td>b</td>
<td>0.603</td>
<td>367.0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>5</td>
<td>9.0</td>
<td>no smoothing</td>
<td>y</td>
<td>d</td>
<td>0.2987</td>
<td>244.0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>5</td>
<td>8.0</td>
<td>smoothing</td>
<td>y</td>
<td>c</td>
<td>0.055</td>
<td>51.1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>5</td>
<td>9.0</td>
<td>smoothing</td>
<td>y</td>
<td>c</td>
<td>0.0018</td>
<td>1.24</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>5</td>
<td>9.0</td>
<td>smoothing</td>
<td>y</td>
<td>c</td>
<td>0.0325</td>
<td>17.27</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>5</td>
<td>10.0</td>
<td>smoothing</td>
<td>y</td>
<td>c</td>
<td>0.00068</td>
<td>0.439</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>5</td>
<td>8.0</td>
<td>residue/nosm</td>
<td>y</td>
<td>d</td>
<td>0.0393</td>
<td>107.00</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>5</td>
<td>9.0</td>
<td>residue/nosm</td>
<td>y</td>
<td>d</td>
<td>0.034</td>
<td>105.21</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>5</td>
<td>10.0</td>
<td>residue/nosm</td>
<td>y</td>
<td>d</td>
<td>0.029</td>
<td>31.01</td>
</tr>
</tbody>
</table>

Equilibration process: i) Minimization - 500 steps of Conjugate Gradient (CG).
ii) Molecular Dynamics initial temperature 10 K, time step = 0.5 fs, 1 ps with temperature scaling to 298 K, 0.02 fs relaxation time.
iii) Equilibration with 1 fs time step, temperature scaling to 298 K with a 0.01 ps relaxation time.
iv) Collection of statistics for 1 ps unless otherwise indicated.
Total equilibration times for stage ii a-c.
a) 4 ps equilibration (1 fs time step); b) 6 ps equilibration (1 fs time step); c) 8 ps equilibration (1 fs time step); d) 1 ps equilibration beginning with b (no smoothing) statistics over 200 fs in stage iv.

more time than atomic cutoffs. Smoothing increases the calculation time because the smoothing function must be recalculated for each nonbonded interaction. With an atomic cutoff, the smoothing function must be calculated for each atom pair in the boundary zone; with a residue cutoff (i.e., for water-water interactions), the smoothing function must be recalculated for each residue pair (i.e., approximately 1/9th as often for water-water interactions).

Results and Discussion

The results of the PTI in water simulations are presented in Table I. In these simulations the net (+6) charge on PTI is neutralized by the addition of 6 Cl⁻ counterions and the dielectric constant is set to 1 in the electrostatic potential as water is included explicitly. In the solvated protein simulations, it was necessary to equilibrate the system with temperature scaling for times up to 8 ps before initiating the constant energy portion of the simulations. With insufficient equilibration, the temperature of the system rose by as much as 100 K when temperature scaling was turned off.

When an atom based cutoff criteria is used without smoothing, the ratio of the fluctuations in the total energy to the fluctuations in the kinetic energy is unacceptably large. For example, with an 8 Å cutoff, a stepsize of 1 fs and a nonbonded list update every 5 steps, the ratio of the root mean square fluctuations in the total energy to kinetic energy is greater than 400%. The very large fluctuations in the total energy are likely due to the effects of truncating the coulomb interactions at relatively short range in this ionic system. It is clearly not possible to calculate thermodynamic
quantities from statistical mechanical fluctuation formulas with any accuracy in these simulations. Despite this, we note that the total energy is still conserved to better than 1% change in the total energy over the course of the simulation. It is instructive to decompose the total energy and other energy quantities in the solvated PTI simulations into contributions from the water-water, the protein-protein, and the protein-water interactions. These simulations included 1 protein molecule (912 atoms, including all hydrogens), and 2,943 water molecules (8,829 atoms). The decomposition of the total energy into components for one such simulation (with smoothing) is displayed in Table II. The average total energy is $-30,870$ Kcal/mol in this simulation, while the root mean square fluctuation in the total energy over 1,000 steps is $4.4$ Kcal/mol. Partitioning the total energy into components, $\sim 80\%$ of the total energy arises from the water-water interactions, $\sim 18\%$ of the total comes from the intramolecular energy of the protein, while the interaction between the protein and solvent accounts for $\sim 2\%$ of the total energy in the system.

A substantial improvement in the accuracy with which the equations of motion are integrated is obtained using either an atomic cutoff with smoothing function or a residue cutoff. The results obtained with these boundary conditions are listed in Table I. While the residue based cutoff scheme conserves energy better than the atom based method without smoothing, the fluctuations in the total energy are still much greater than those obtained with a smoothing function and atom cutoff. With a 9 Å cutoff, the fluctuation in the total energy is 0.01% relative to the total energy and 1.24% relative to the fluctuation in the kinetic energy in the simulation which
Energy Conservation & Molecular Dynamics

included a smoothing function and atom based nonbonded list. The fluctuation in the total energy is substantially reduced using a 9 Å cutoff relative to the values obtained with an 8 Å or shorter cutoff. Using a 10 Å cutoff with smoothing, these values are reduced by two thirds with respect to the 9 Å smoothing results. It is unclear why there is such a large decrease in the fluctuations in the total energy when the cutoff is increased from 8 Å to 9 Å, and whether there are specific ionic interactions at ~8 Å in the solvated PTI simulation which account for this. Further work on this problem is in progress.

We have examined the effect of increasing the timestep on the fluctuations in total energy; the results are listed in Table 1. With a 9 Å cutoff, the relative fluctuation in the total energy increases from 0.002% to 0.03% and the fluctuation in the total energy divided by the fluctuation in the kinetic energy increases from 1% to 17% when the stepsize is doubled from 1 fs to 2 fs. These values are still considerably smaller than the corresponding values for simulations with a 1 fs timestep and an 8 Å cutoff. The increase in computation time in going from an 8 Å cutoff to a 9 Å cutoff is ~20%, therefore, our results suggest that to achieve a prescribed level of accuracy most efficiently, it is better to choose a longer time step coupled to a longer range cutoff of the nonbonded interactions.

In summary, we find the most favorable simulation parameters to choose for simulations of solvated proteins, from the standpoint of minimizing the error with which the equations of motion are integrated, involves the use of an atomic cutoff at relatively large distance (9 Å or greater) with the use of a smoothing function. The residue based cutoff scheme leads to a much larger ratio of the fluctuations in total energy to fluctuations in kinetic energy; although the relative error in the total energy is still small. The advantages of a residue based cutoff are two fold: (1) there is a saving in computer storage requirements associated with the much smaller size of a residue (as opposed to atomic) pair list and (2) the interaction potential is not distorted by the application of an artificial (smoothing) function. The residue based cutoff has the disadvantage that the cutoff boundary around any given atom is irregular. With respect to the cpu time required per cycle, both cutoff schemes are roughly comparable. Unfortunately, there has been little analysis done on the effects of different nonbonded cutoff methods on the average structural or dynamic properties of proteins in molecular dynamics simulations. Brooks has compared the effects of varying the long range cutoff on the structures and fluctuations in a series of molecular dynamics simulations of myoglobin in vacuum (22). He found that the use of a shifted potential lead to large deviations from the x-ray structure. We are currently analyzing these effects in a series of solvated PTI simulations. The goal is to delineate conditions for simulating proteins in solution which maximize the accuracy of the simulations while minimizing the computational cost.

Acknowledgments

This work has been supported in part by the National Institutes of Health grant GM-30580. DBK and FH have been supported in part by Supercomputer Postdoctoral Fellowships from the New Jersey Commission on Science and Technology. DBK is
the recipient of a National Institutes of Health NRSA fellowship. We also wish to acknowledge a grant of supercomputer time from the John Von Neumann Center, a NSF national supercomputer facility. We thank David Kofke for many helpful discussions and John Westbrook for help with the development and implementation of IMPACT code for molecular dynamics simulations.

References and Footnotes