Long Time Molecular Dynamics for Enhanced Conformational Sampling in Biomolecular Systems

P. Minary,[1] M. E. Tuckerman,[1,*] and G. J. Martyna[2]

[1] *Department of Chemistry, New York University, New York, New York 10003, USA*

[2] *IBM Physical Sciences Division, PO Box 218, Yorktown Heights, New York 10598, USA*

(Received 20 April 2004; published 8 October 2004)

Although molecular dynamics methods are commonly used to drive biomolecular simulations, the technique provides insufficient sampling to impact studies of the 200-300 residue proteins of greatest interest. One severe limitation of molecular dynamics is that the integrators are restricted by resonance phenomena to small time steps ($\Delta t < 8$ fs) much slower then the time scales of important structural and solvent rearrangements. Here, a novel set of equations of motion and a reversible, resonance-free, integrator are designed which permit step sizes on the order of 100 fs to be used.

DOI: 10.1103/PhysRevLett.93.150201 PACS numbers: 02.70.Ns, 87.15.Aa

Developing the capability to determine rapidly the conformational equilibria of models of proteins in aqueous solvent by simulation is a major goal of computational biophysics. If achieved, this would permit theoretical studies to impact the current understanding of the structure, function, and design of proteins.

Modern molecular dynamics (MD) and Monte Carlo methods are currently limited by two bottlenecks: the high barriers which arise during the folding process and the multiple time scales inherent in the dynamics. Enhancing the rate of barrier crossing events without introducing bias, or introducing bias that can be rigorously removed *a posteriori*, is at the heart of many new approaches [1,2]. Although promising, these methods do not directly address the limitation imposed by the multiple time scales inherent in the problem and, hence, suffer from an *a priori* loss of efficiency. For example, the period of a C-H stretch in the backbone of a peptide is on the order of femtoseconds while large scale domain motion occurs on time scales of microseconds for the larger more complex proteins of real interest. In addition, the methods of Refs. [1,2] often concentrate on the solute and do not attempt to speed equilibration/sampling of the solvent which then becomes the bottleneck in the computation.

Multiple time step (MTS) integrators, such as the reference system propagator algorithm (RESPA) [3], were introduced to alleviate time scale separations in a fairly straightforward manner, but efficiency gains are fundamentally limited by resonance phenomena which restrict the time steps used by these solvers to $\Delta t < 8$ fs in biophysical systems [4]. The basic difficulty is that MTS integrators, even those that possess the symplectic property of Hamilton's equations, are derived using perturbation theories and, therefore, suffer from the same resonance problems that plague all perturbative techniques [4]. Although clever methods have been developed to help overcome resonance phenomena [5], current approaches involve implicit equations that require both nonreversible solvers and the introduction of a stochastic bath. The latter actually reduces phase sampling by overdamping the motions that are often of key biological importance [6]. The use of stochastic baths, also, obviates the symplectic property and destroys time reversal symmetry, both of which are key to employing current techniques as part of hybrid Monte Carlo (HMC) schemes [7].

In this Letter, a novel set of resonance-free equations of motion is developed for use in MD simulations via the non-Hamiltonian theoretical statistical framework of Ref. [8]. The new equations of motion are shown to generate, rigorously, the canonical distribution in the physical configuration space. Integrators for this set of equations are derived using symmetric operator splitting techniques that preserve the resonance resistant character of the equations. These solvers thus permit very large time steps to be employed and ergodic sampling to be obtained without recourse to stochastic baths and/or implicit methods. While MD is very efficient on novel large scale parallel computers like IBM's BlueGene [9] and, hence, is used here, the new numerical integrator could also be used to drive HMC sampling schemes based on a new non-Hamiltonian HMC formalism [10]. By applying the method to simple examples as well as a model of liquid water and a protein *in vacuo*, it is possible to evaluate, separately, the enhancement in time step for a solvent and a typical biomolecule. It is found that time steps of $\Delta t \approx 100$ fs can now be employed to treat both solutes and solvents without losing numerical stability or affecting equilibrium properties, an increase of a full order of magnitude in step size compared to multiple time standard methods.

The new equations were designed to have the following properties: (1) The equations should be simple and sample the desired phase space efficiently. (2) They eliminate resonance by incorporating a set of constraints that formally prevents energy from building up in any one mode. Since any mode in the system can become resonant, a constraint should be placed on each degree of freedom.

(3) They employ a deterministic heat bath coupled to each degree of freedom in order to ensure ergodicity.

Consider and N -particle system with positions $\mathbf{r}_1, \ldots, \mathbf{r}_N \equiv \mathbf{R}$ and momenta $\mathbf{p}_1, \ldots, \mathbf{p}_N$ interacting via potential, $\phi(\mathbf{r}_1, \ldots, \mathbf{r}_N)$. Let x, v denote one Cartesian position and velocity pair, respectively. For each pair, the following non-Hamiltonian equations are proposed:

$$\dot{x} = v; \qquad \dot{v} = \frac{F}{m} - \lambda v$$

$$\dot{\bar{\eta}} = -\sum_{j=1}^{L}\left[\frac{Q v_{\eta_{2,j}} v_{\eta_{1,j}}^2}{k_B T} - \sum_{i=2}^{M} v_{\eta_{i,j}}\right]$$

$$\dot{v}_{\eta_{1,j}} = -v_{\eta_{1,j}} v_{\eta_{2,j}} - \lambda v_{\eta_{1,j}} \qquad j = 1, L$$

$$\dot{v}_{\eta_{i,j}} = \frac{G_{i,j}}{Q} - v_{\eta_{i,j}} v_{\eta_{i+1,j}} \qquad j = 1, L; \quad i = 2, M-1$$

$$\dot{v}_{\eta_{M,j}} = \frac{G_{M,j}}{Q} \qquad j = 1, L \tag{1}$$

where $F = -\partial\phi(\mathbf{R})/\partial x$ and $G_{i,j} = Q v_{\eta_{i-1,j}}^2 - k_B T$. Here, $Q = k_B T \tau^2$, k_B is Boltzmann's constant, T is the temperature, and τ is the time scale associated with the bath. The Lagrange multiplier, λ, is determined such that the equations of motion satisfy the constraint, $2K(v, v_\eta) = L k_B T$ with

$$2K(v, v_\eta) = \left\{ m v^2 + \left(\frac{L}{L+1}\right)\sum_{j=1}^{L} Q v_{\eta_{1,j}}^2 \right\}, \tag{2}$$

which yields $\lambda = [vF - L/(L+1)\sum_{j=1}^{L} Q v_{\eta_{1,j}}^2 v_{\eta_{2,j}}]/2K(v, v_\eta)$. Equation (2) ensures that the maximum kinetic energy that can accumulate in any one mode is $L k_B T$, which effectively eliminates resonance artifacts in associated numerical integrators. The equations of motion constitute a nontrivial combination of the massive Nosé-Hoover chain [11] method and isokinetic dynamics [11,12]. While Eqs. (2) can be used to define a HMC algorithm, they are not, themselves, equivalent to Metropolis, Kawasaki, or Glauber Monte Carlo schemes.

In order to show that the equations of motion, Eqs. (1), produce the canonical configurational distribution, assuming ergodicity, the statistical theory of non-Hamiltonian systems is employed [8]. The generalized ensemble sampled by a set of equation of motion, $\dot{\Gamma} = \xi(\Gamma)$ with n_c conservation laws, $f_k(\Gamma) = C_k$, can be written as [8]

$$\Omega = \int d\Gamma \sqrt{g(\Gamma)} \prod_{k=1}^{n_c} \delta(f_k(\Gamma) - C_k) \tag{3}$$

where the metric determinant factor, $\sqrt{g(\Gamma)}$, is given by [8] $d\log\sqrt{g}/dt = -\nabla_\Gamma \cdot \xi(\Gamma)$. Here, $\Gamma = \{\mathbf{r}, \mathbf{v}, \bar{\eta}^{(k)}, v_{\eta ij}^{(k)}, i = 1, \ldots, M, j = 1, \ldots, L, k = 1, \ldots, 3N\}$. Computing the divergence $\nabla_\Gamma \cdot \xi(\Gamma)$ for Eqs. (1) yields $k_B T(d\log\sqrt{g}/dt) = -(d/dt)[\phi(\mathbf{R}) + \sum_k \bar{\eta}^{(k)}]$ and the equations give rise to the isokinetic distribution

$$\Omega = \int d^N\mathbf{r} d^N\mathbf{v} d^{3NLM} v_\eta \prod_{k=1}^{3N} \delta\left(K[v^{(k)}, v_\eta^{(k)}] - \frac{L k_B T}{2}\right) \times e^{-\beta[\phi(\mathbf{R}) + \sum_k H_\eta(v_\eta^{(k)})]}$$

$$\propto \int d^N\mathbf{r} e^{-\beta\phi(\mathbf{R})}. \tag{4}$$

Here, $H_\eta(v_\eta) = (1/2)\sum_{j=1}^{L}\sum_{i=2}^{M} Q v_{\eta_{i,j}}^2 \beta = 1/k_B T$, and trivial conservation laws of the form $H_\eta(v_\eta) + k_B T\bar{\eta} = C$ for each degree of freedom have been integrated out. The distribution is, indeed, canonical in the physical configuration space, $\{\mathbf{r}_1, \ldots, \mathbf{r}_N\}$, allowing Eqs. (1) to be employed rigorously as sampling tool.

Eqs. (1) can be numerically integrated using operator splitting techniques following Refs. [11,12]. In the Liouville operator formalism, the phase space evolves in time according to $\Gamma(t) = \exp(iLt)\Gamma(0)$, and a single time step of evolution is given by $\Gamma(\Delta t) = \exp(iL\Delta t)\Gamma(0)$ with $\Delta t = t/P$. The Liouville operator, $iL = \dot{x}(\partial/\partial x) + \dot{v}(\partial/\partial v) + \sum_{ij}\dot{v}_{\eta_{ij}}(\partial/\partial v_{\eta_{ij}}) + \dot{\bar{\eta}}(\partial/\partial\bar{\eta})$, is decomposed as

$$iL = iL_x + \sum_{p=1}^{N_d} iL_{v,p} + iL_{\mathrm{NHC}}; \qquad iL_x = v\frac{\partial}{\partial x}$$

$$iL_{v,p} = (F_p - \lambda_p v)\frac{\partial}{\partial v} - \sum_j \lambda_p v_{\eta_{1,j}}\frac{\partial}{\partial v_{\eta_{1,j}}}$$

$$iL_{\mathrm{NHC}} = \sum_{i=2}^{M}\sum_{j=1}^{L}\frac{G_{i,j}}{Q}\frac{\partial}{\partial v_{\eta_{i,j}}} - \sum_{i=1}^{M-1}\sum_{j=1}^{L} v_{\eta_{i,j}} v_{\eta_{i+1,j}}\frac{\partial}{\partial v_{\eta_{i,j}}} - \sum_{j=1}^{L}\lambda_{\mathrm{NHC}} v_{\eta_{1,j}}\frac{\partial}{\partial v_{\eta_{1,j}}} + \dot{\bar{\eta}}\frac{\partial}{\partial\bar{\eta}} \tag{5}$$

where the force has been split into N_d parts, $\sum_{p=1}^{N_d} F_p = F$, whose strength/time scale is assumed to decrease/increase with p. Using MTS parameters, $\delta t = \Delta t/N_{\mathrm{MTS}}$, $N_{\mathrm{MTS}} = \prod_{p=1}^{N_d} n_p$, $n_{N_d} = 1$, $w_p = \prod_{k=1}^{p-1} n_k$, $w_1 = 1$, and a Trotter decomposition of the classical propagator, $\exp(iL\Delta t)$, an accurate approximation to the true evolution can be written as

$$\Gamma(\Delta t) \approx \left\{ e^{i\tilde{L}_{N_d}^{(t)}\delta t} \ldots \left(e^{i\tilde{L}_2^{(t)}\delta t}\left[e^{i\tilde{L}_1\delta t}\right]^{n_1-2} \times e^{i\tilde{L}_2\delta t}\right)^{n_2-2} \ldots e^{i\tilde{L}_{N_d}\delta t}\right\}\Gamma(0) e^{i\tilde{L}_k\delta t}$$

$$= e^{iL_{\mathrm{NHC}}\frac{\delta t}{2}} e^{iL_{v,1}\frac{\delta t}{2}} e^{iL_x\delta t} e^{\sum_{p=1}^{k} iL_{v,p} w_p \frac{\delta t}{2}} e^{iL_{\mathrm{NHC}}\frac{\delta t}{2}} \tag{6}$$

where $\exp(i\tilde{L}_k^{(t)}\delta t)$is the transpose of $\exp(i\tilde{L}_k\delta t)$. In this way, the weaker forces which are assigned larger p, are evaluated less frequently but with larger weight w_p which

equalizes them to the strong forces [e.g., $w_{p+1}/w_p = n_p$ is the ratio of the strength of the p^{th} and the $(p+1)^{\text{st}}$ force]. The number of evaluations of the p^{th} force is N_{MTS}/w_p where, again, N_{MTS} is the total number of small steps in the multiple time step procedure [3]. The error in the scheme is $O(\Delta t^3)$ for one full step and $O(t\Delta t^2)$ for the trajectory. Analytical solutions for each of the factorized parts of the evolution operators, $\exp(i\tilde{L}_k \delta t)$, can be obtained easily provided iL_{NHC} is further decomposed following Ref. [11]. The decomposed multipliers, $\{\lambda_p, \lambda_{\text{NHC}}\}$ which sum to the total multiplier, $\sum_p \lambda_p + \lambda_{\text{NHC}} = \lambda$, ensure the constraints, Eq. (2), are exactly satisfied by each individual exponential operator, thereby maintaining the resonance-free character of the exact equations of motion. Judicious choices of the force decomposition into strong intramolecular vibrations, weak short range forces and weaker long range forces have been discussed in detail elsewhere [3]. The multiple time step approach improves efficiency because for chemical systems the strongest forces are the least computationally intensive to calculate. The new method will subsequently be referred to as the isokinetic Nosé-Hoover chain RESPA or the Iso-NHC-RESPA (INR) technique.

In order to demonstrate the efficacy of INR, three problems with very different separations of time scale were selected for study. First, a quartic oscillator is chosen to demonstrate unequivocally that the most basic resonance phenomenon has been eliminated. In Fig. 1, the convergence of the probability distribution function of the quartic oscillator integrated at the resonant time step, $\Delta t = \pi/\omega$, is given under both the new INR ($L = 1$, $M = 3, N_d = 2$), and the standard NHC-RESPA (NR) ($M = 3, N_d = 2$) [11]. In detail, an inner or small time step of $\delta t = \pi/(100\omega)$ is used to integrate the harmonic part, while an outer or large time step of $\Delta t = \pi/\omega$ is used to integrate the quartic term. The runs are of length $(5 \times 10^7)\Delta t$. While NR fails to sample the phase space properly, INR yields the correct distribution. The error in the distribution as a function of simulation time is estimated in Fig. 1 by the quantity $\zeta(t) = (\alpha/N_P) \times \sum_{i=1}^{P} |P(r_i; t) - P(r_i; \infty)|$, where N_P is the number of discrete points at which the distribution is computed, α is an appropriately chosen normalization, and $P(r_i; \infty)$ is the exact distribution.

Second, a very challenging problem with extremely large separations of time scales is a flexible water model [13] at standard conditions simulated using particle mesh Ewald [14] to compute the long range electrostatic interactions. Because of resonance, a single time step method cannot be employed using small time steps larger that $\delta t \approx 2$ fs while MTS methods require large time steps less than $\Delta t \approx 8$ fs. However, INR ($L = 1, M = 3, N_d = 3$) generates the radial distribution function accurately using a $\Delta t = 100$ fs time step as shown in Fig. 2(a). The intramolecular forces were evaluated at a time step of $\delta t = 0.5$ fs , intermediate range forces at a time step of 3 fs, and long range forces at the large time step, Δt, specified in the legend of the figure for a given run. All runs were of length 400 ps. The mean square displacement of the oxygen atoms is independent of Δt, hence, the new method is free from the drag that would be caused by the introduction of an overdamped bath. Since the overhead of the technique is only 15% , threefold larger speedups are obtained than previously reported for MTS methods [3,15]. The ideal tenfold increase in computational efficiency over *standard MTS methods* is not fully realized (i.e., the time step is, now, 10 times larger) due to the overhead in computing the intermediate and short range forces. It is expected that a sixfold increase in efficiency over standard MTS methods can be obtained based on data given in Ref. [15] which would yield an overall factor of thirtyfold increase in efficiency compared to a single time step method. Efforts are currently under way to implement the improved force decomposition of Ref. [15] in the PINY-MD software package [16] which was employed to generate the results presented here. *Although it is possible to use still larger time steps* under INR, 100 fs is the correlation time for the observables given here, and using larger steps does not speed convergence as measured by $\zeta(t)$ [see Fig. 2(b)].

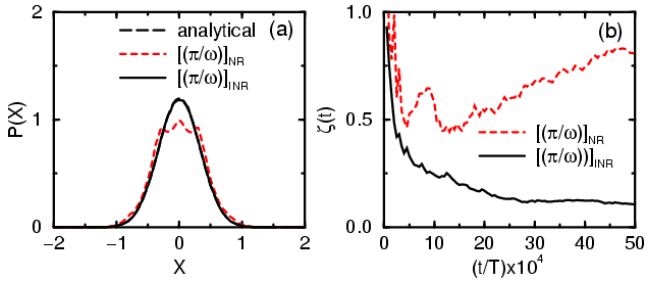

FIG. 1 (color online). (a) Distribution function of the quartic oscillator, $\phi(x) = (9/2)x^2 + (0.1/4)x^4$ for the NR and the INR methods using $\Delta = \pi/\omega$, the resonant time step and $\delta t = \pi/(100\omega)$. Here, $T = 2\pi/\omega$. (b) The error in the distribution function as a function of time.

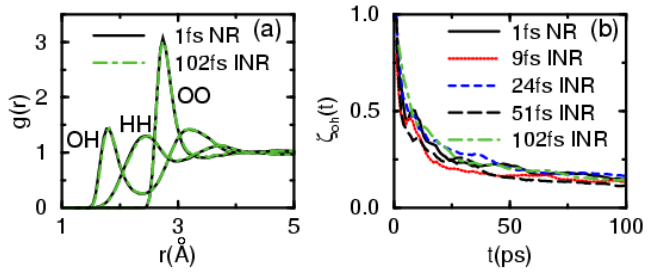

FIG. 2 (color online). (a) The radial distribution function of flexible liquid water calculated using NR and the INR methods. (b) The error in the radial distribution function as a function of time. The "exact" distribution is generated from a long run with a small time step. The legend reports the large or outer time step, Δt.

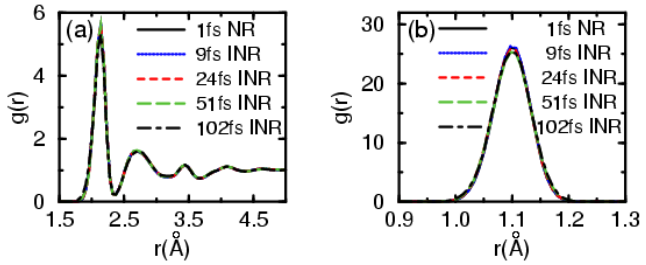

FIG. 3 (color online). (a) The C-H radial distribution function of HIV protease computed using NR and the INR methods. (b) The intramolecular part of the distribution. All the carbon and hydrogen atoms in the protein were considered in generating the distribution function. The legend reports the large or outer time step, Δt.

Third, a protein (the wild-type HIV-1 protease) is studied in *vacuo* in order to demonstrate that the INR technique can handle large molecules without masking inefficiencies by including solvent. The protein is treated using the all-atom CHARMM22 force field [17]. Again, a large step time of $\Delta t = 100$ fs was capable of providing excellent results as exemplified by both short and long range parts of the carbon-hydrogen radial distribution function depicted in Fig. 3. (All carbon and hydrogen atoms in the protein are used to construct the distribution function.) The $N_d = 3$ part force decomposition used for water was also employed here.

In the effort to develop methodology capable of driving biomolecules to sample correctly their conformational equilibrium, it is clear that combining a suite of techniques, each designed to tackle one aspect of the problem, is the path to a solution. Future work will focus on pursuing this course of action.

This research was supported by NSF-0229959 and IBM Research (G.J.M.) and NSF CHE-0121375, NSF CHE-0310107, the Camille and Henry Dreyfus Foundation, Inc. (TC-02-012), and New York University's Research Challenge Fund (M.E.T.).

*Also at the Courant Institute of Mathematical Sciences, New York University, New York, New York 10003, USA

[1] J. I. Siepmann and D. Frenkel, Mol. Phys. **75**, 59 (1992); P. Amara and J. E. Straub, J. phys. chem. **99**, 14840 (1995); R. Olender and R. Elber, J. Chem. Phys. **105**, 9299 (1996); N. Nakajima, H. Nakamura and A. Kidera, J. Phys. Chem. B **103**, 399 (1999); Y. H. Lee and B. J. Berne, J. Phys. Chem. A **104**, 86 (2000); D. Passerone and M. Parrinello, Phys. Rev. Lett. **87**, 108302 (2001); P. Eastman and N. Gronbech-Jensen, J. Chem. Phys. **114**, 3823 (2001).

[2] D. D. Frantz, D. L. Freeman and J. D. Doll, J. Chem. Phys. **93**, 2769 (1990); E. Marinari and G. Parisi, Europhys. Lett. **19**, 451 (1992); B. A. Berg and T. Neuhaus, Phys. Rev. Lett. **68**, 9 (1992); Forest and Suter, Mol. Phys. **82**, 393 (1994); A. Monge *et al.*, J. Mol. Biol. **247**, 995 (1995); H. Jönsson, Surf. Sci. **324**, 305 (1995); J. E. Shea *et al.*, J. Chem. Phys. **109**, 2895 (1998); J. Pillardy, A. Liwo and H. A. Scheraga, J. Phys. Chem. A **103**, 9370 (1999); H. Lu and K. Schulten, Proteins: Struct., Funct., Genet. **35**, 453 (1999); E. Shakhnovich and L. Mirny, Biophys. J. **78**, 761 Symp (2000); M. R. Sorenson and A. F. Voter, J. Chem. Phys. **112**, 9599 (2000);A. Laio and M. Parrinello, Proc. Natl. Acad. Sci. U.S.A. **99**, 12562 (2002); Z. Zhu *et al.*, Phys. Rev. Lett. **88**, 100201 (2002).

[3] M. E. Tuckerman, B. J. Berne, and G. J. Martyna, J. Chem. Phys. **97**, 1990 (1992); M. E. Tuckerman and G. J. Martyna, J. Phys. Chem. B **104**, 159 (2000).

[4] T. Schlick *et al.*, J. Comput. Phys. **140**, 1 (1998); Q. Ma, J. A. Izaguirre, and R. D. Skeel, SIAM J. Sci. Comput. **24**, 1951 (2003).

[5] E. Barth and T. Schlick, J. Chem. Phys. **109**, 1617 (1998); A. Sandu and T. Schlick, J. Comput. Phys. **151**, R45 (2001); S. Chin, J. Chem. Phys. **120**, 8 (2004).

[6] S. Hammes-Schiffer, Biochemistry **41**, 13335 (2002).

[7] S. Duane *et al.*, Phys. Lett. B **195**, 216 (1987).

[8] M. E. Tuckerman, C. J. Mundy, and M. L. Klein, Phys. Rev. Lett. **78**, 2042 (1997); M. E. Tuckerman, C. J. Mundy and G. J. Martyna, Europhys. Lett. **45**, 149 (1999); M. E. Tuckerman *et al.*, J. Chem. Phys. **115**, 1678 (2001).

[9] F. Allen *et al.*, IBM Systems Journal **40**, 310 (2001).

[10] M. Eleftheriou and G. J. Martyna, J. Chem. Phys. (to be published).

[11] G. J. Martyna, M. E. Tuckerman and M. L. Klein, J. Chem. Phys. **97**, 2635 (1992); G. J. Martyna *et al.*, Mol. Phys. **87**, 1117 (1996); M. E. Tuckerman *et al.*, J. Chem. Phys. **99**, 2796 (1993); D. J. Evans and G. P. Morriss, Chem. Phys. **77**, 63 (1983).

[12] P. Minary, G. J. Martyna, and M. E. Tucherman, J. Chem. Phys. **118**, 2510 (2003).

[13] J. Izaguirre, S. Reich and R. D. Skeel, J. Chem. Phys. **110**, 9853 (1999).

[14] U. Essmann, et al., J. Chem. Phys. **103**, 8577 (1995).

[15] R. Zhou *et al.*, J. Chem. Phys. **115**, 2348 (2001); X. Qian and T. Schlick, J. Chem. Phys. **116**, 5971 (2002).

[16] M. E. Tuckerman, *et al.*, Comput. Phys. Commun. **128**, 333 (2000).

[17] A. MacKerell, Jr., *et al.*, J. Phys. Chem. B **102**, 3586 (1998).