Grand canonical Monte Carlo simulations of water in protein environments

Hyung-June Woo[a)]
Department of Biochemistry, Weill Medical College of Cornell University, New York, New York 10021

Aaron R. Dinner[b)]
Department of Chemistry, The University of Chicago, Chicago, Illinois 60637

Benoît Roux[c)]
Department of Biochemistry, Weill Medical College of Cornell University, New York, New York 10021

(Received 12 April 2004; accepted 25 June 2004)

The grand canonical simulation algorithm is considered as a general methodology to sample the configuration of water molecules confined within protein environments. First, the probability distribution of the number of water molecules and their configuration in a region of interest for biochemical simulations, such as the active site of a protein, is derived by considering a finite subvolume in open equilibrium with a large system serving as a bulk reservoir. It is shown that the influence of the bulk reservoir can be represented as a many-body potential of mean force acting on the atoms located inside the subvolume. The grand canonical Monte Carlo (GCMC) algorithm, augmented by a number of technical advances to increase the acceptance of insertion attempts, is implemented, and tested for simple systems. In particular, the method is illustrated in the case of a pure water box with periodic boundary conditions. In addition, finite spherical systems of pure water and containing a dialanine peptide, are simulated with GCMC while the influence of the surrounding infinite bulk is incorporated using the generalized solvent boundary potential [W. Im, S. Bernèche, and B. Roux, J. Chem. Phys. **114**, 2924 (2001)]. As a last illustration of water confined in the interior of a protein, the hydration of the central cavity of the KcsA potassium channel is simulated. [DOI: 10.1063/1.1784436]

I. INTRODUCTION

A proper treatment of solvent effects often plays a crucial role in simulations of biomolecules.[1] This is particularly important in the case of confined environments created by proteins, as in the interior of deep crevices, pockets, or other poorly accessible regions. Prominent examples include the substrate binding sites of enzymes[2,3] and the narrow pore of membrane channels.[4–6] The average properties of solvent molecules are expected to differ substantially from those in bulk water in such regions. Computational approaches at different levels of complexity and sophistication can be used to describe the influence of solvent on biomolecular systems.[7] Those range from molecular dynamics (MD) simulations based on all-atom models in which the solvent is treated explicitly,[1,8] to Poisson-Boltzmann (PB) continuum electrostatic models in which the influence of the solvent is incorporated implicitly.[7,9] There are also a number of intermediate ''solvent boundary potential'' approaches which combine some aspects of both implicit and explicit solvent treatments. Such approaches consist in simulating a small number of explicit solvent molecules in the vicinity of a region of interest, while representing the influence of the surrounding solvent with an effective boundary potential.[10–19] The general concept is illustrated schematically in Fig. 1.

One recently developed method, called the generalized solvent boundary potential (GSBP),[11] includes both the solvent-shielded static field from the distant atoms of the macromolecule and the reaction field from the dielectric solvent acting on the atoms of the simulation region. GSBP is a generalization of spherical solvent boundary potential, which was designed to simulate a solute in bulk water.[10] In GSBP, all atoms in a small subvolume (macromolecule or solvent) can undergo explicit dynamics whereas the influence of the macromolecular and solvent atoms outside this region are incorporated implicitly via a potential of mean force (PMF). The influence of the surrounding outer region of the macromolecular system is represented in terms of a solvent-shielded static field and a solvent-induced reaction field. The reaction field due to changes in charge distribution in the dynamic inner region is expressed in terms of a basis set expansion of the inner simulation region charge density, the basis set coefficients corresponding to generalized electrostatic multipoles. The solvent-shielded static field from outer macromolecular atoms and the reaction field matrix, representing the couplings between the generalized multipoles, are both invariant with respect to the configuration of the explicit atoms in the inner simulation region. They are calculated only once for dielectric boundaries of arbitrary geometry using the finite-difference PB equation, leading to an accurate and computationally efficient hybrid MD/continuum method for simulating a small region of a large biological macromolecular system.

The number of water molecules that should be included

a)Curent address: Department of Cemistry/216, University of Nevada, Reno, NV 89557.
b)Fax: 773-834-5250. Electronic mail: dinner@uchicago.edu
c)Fax: 212-746-4843. Electronic mail: benoit.roux@med.cornell.edu

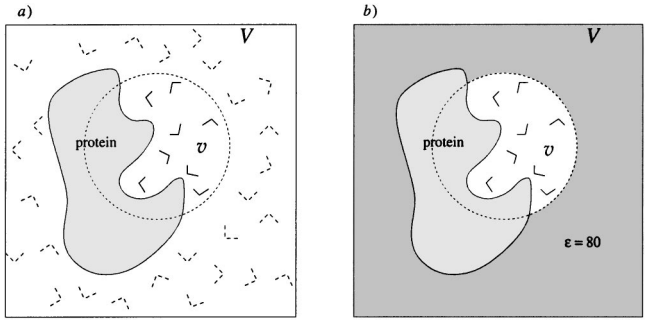

FIG. 1. Schematic representation of considering a small finite subvolume v part of a large biomolecule system comprising a macromolecular solute surrounded by a large reservoir of solvent molecules. (a) The entire system, divided into two regions, is shown with explicit solvent molecules. The small finite subvolume is in open equilibrium with a large volume V at temperature T containing N water molecules. The number of water n inside v fluctuates as they cross the system boundary. (b) A coarse-grained description for dynamical simulations, where n water molecules are explicitly simulated in the subvolume while the influence of the surrounding solvent is incorporated via the generalized solvent boundary potential (GSBP).[11]

in such finite simulation systems is problematical they should, in principle, be in open thermodynamic equilibrium with the bulk phase. The standard ''overlay'' method used for constructing an initial configuration of a solvated protein consists in inserting the protein structure into an equilibrated pure water box and removing the overlapping water molecules.[1] In typical cases, the initial configuration is sufficiently representative of a solvated protein that it can reach equilibrium during some reasonable simulation time. In more complicated cases, it is hoped that water molecules would diffuse in and out of confined protein regions such as binding pockets, and rapidly reach equilibrium with the bulk phase. However, achieving complete equilibration via diffusion may require unreasonably long simulation times when a region of space is poorly accessible. These difficulties are further compounded when a finite region of a much larger macromolecule is simulated with a fixed number of water molecules in the presence of effective mean-field solvent as illustrated in Fig. 1. Rigorously, the number of water molecules to be included in such simulation should be allowed to fluctuate since the system is assumed to be in open equilibrium with a bulk water reservoir. The correct way to simulate such systems is to treat the subvolume as a grand canonical ensemble in contact with a bulk water reservoir with a fixed chemical potential, and maintain equilibrium of the number of water molecules inside the simulation volume of interest. Such grand canonical Monte Carlo (GCMC) simulations have been implemented and used by many authors.[20–27] In large part, previous applications have been focused on examining the phase behavior and thermodynamic properties of bulk fluids. GCMC algorithms have also been used in biomolecular simulations to characterize lipid conformations in a bilayer,[28] and water positions in the crystal hydrates of small molecules.[29,30] The GCMC method has also been used in conjunction with MD in the nonequilibrium dual control volume simulations of fluid flows.[31–33] Finally, Im, Seefeld, and Roux[34] have utilized GCMC simulations to impose nonequilibrium concentrations of ions to calculate ion fluxes in Brownian dynamics simulation of membrane channels.

In this paper, we have adapted and implemented the GCMC algorithm to the problem of a finite simulation region in exchange with an infinite bulk reservoir of bulk water at fixed excess chemical potential. The present perspective differs slightly from that in the usual GCMC simulation of bulk liquids. Instead of evaluating the properties of the bulk liquid water at a given condition, one is typically interested in the solvation of a finite region of interest in contact with bulk water, with the thermodynamic properties of the bulk phase regarded as known input parameters.

In the following section, we cast the grand canonical ensemble to a form most natural for our purpose. In addition, particular techniques designed to increase the efficiency of the application of the GCMC algorithm to dense liquids, are considered. Direct insertions and deletions of solvent molecules attempted at random into dense liquid systems are mostly rejected and do not allow for adequate sampling. In Sec. II b, we implement a ''dynamical-grid algorithm'' and compare it to the ''random insertion'' method and to ''cavity-bias algorithm'' of Mezei.[21] The implementation of the GCMC algorithm is tested on simple systems in Sec. III: a bulk water box in periodic boundary conditions, a pure water sphere with GSBP,[11] a water sphere solvating a dialanine with GSBP, and the interior of the central cavity of an isolated KcsA potassium channel.

II. THEORETICAL DEVELOPMENTS

A. Grand canonical ensemble for finite systems in open equilibrium

We consider a macroscopic volume V (''supersystem'') containing a solute (e.g., a protein) in some fixed configuration denoted by $\mathbf{X}$ and solvated by a large number N of water molecules. The configuration of the water molecule is represented as $\{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$, each $\mathbf{x}_i$ denoting the coordinates of three atoms within a given molecule. The total potential energy of the system is U. The macroscopic volume is in contact with a heat reservoir at temperature T, and is sufficiently large such that, in regions relatively far away from the solute at the center of the volume, the local liquid water properties can be considered essentially identical to those of bulk water at ambient conditions. The probability density of the supersystem configurations is given by the canonical distribution,

$$P_N(\mathbf{X},\mathbf{x}_1,\ldots,\mathbf{x}_N)=\frac{e^{-\beta U(\mathbf{X},\mathbf{x}_1,\ldots,\mathbf{x}_N)}}{\int_V d\mathbf{x}_1\cdots d\mathbf{x}_N e^{-\beta U(\mathbf{X},\mathbf{x}_1,\ldots,\mathbf{x}_N)}}$$
$$=V^{-N}e^{\beta A_{\mathrm{ex}}^{(N)}}e^{-\beta U(\mathbf{X},\mathbf{x}_1,\ldots,\mathbf{x}_N)}, \quad (1)$$

where $1/\beta=k_BT$ is the Boltzmann constant times temperature and $A_{\mathrm{ex}}^{(N)}$ is the excess Helmholtz free energy

$$e^{-\beta A_{\mathrm{ex}}^{(N)}}=\frac{1}{V^N}\int_V d\mathbf{x}_1\cdots d\mathbf{x}_N e^{-\beta U(\mathbf{X},\mathbf{x}_1,\ldots,\mathbf{x}_N)}. \quad (2)$$

We are interested in the details of the system in a subvolume v, which is assumed to be much smaller than the supersystem serving as a thermodynamic bulk reservoir. The

probability $P_n^{(N)}$ of finding a subset of n solvent molecules with coordinates $\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$, while the rest of the molecules are confined outside the volume v is

$$P_n^{(N)}(\mathbf{X},\mathbf{x}_1,\ldots,\mathbf{x}_n) = \frac{N!}{n!(N-n)!}\int_{\text{out}} d\mathbf{x}_{n+1}\cdots d\mathbf{x}_N P_N(\mathbf{X},\mathbf{x}_1,\ldots,\mathbf{x}_N), \tag{3}$$

where $i=1,\ldots,n$ were chosen for the set of n molecules, and the prefactor accounts for the indistinguishability of molecular labels. The subscript "out" implies that the configurational integrals of the remaining molecules are restricted to the exterior of the subvolume v. Such restricted spatial integrals can be reexpressed as unrestricted integrals by introducing the step-functions $H_{\text{out}}(\mathbf{x}_i)$, equal to 0 when a molecule is located within the subvolume v, and 1 otherwise. From Eqs. (1) and (3), we have

$$P_n^{(N)}(\mathbf{X},\mathbf{x}_1,\ldots,\mathbf{x}_n) = \frac{N!}{n!(N-n)!V^N} e^{\beta A_{\text{ex}}^{(N)}} \times \int_V d\mathbf{x}_{n+1}\cdots d\mathbf{x}_N H_{\text{out}}(\mathbf{x}_{n+1})\cdots \times H_{\text{out}}(\mathbf{x}_N) e^{-\beta U(\mathbf{X},\mathbf{x}_1,\ldots,\mathbf{x}_N)}. \tag{4}$$

Effectively, the spatial restriction enforced by the step function can be regarded as a hard sphere potential acting only on the $i=n+1,\ldots,N$ water molecules in the bulk solvent. It is useful to consider a special case of Eq. (4)

$$P_0^{(N-n)}(\mathbf{X}) = \frac{e^{\beta A_{\text{ex}}^{(N-n)}}}{V^{N-n}}\int_V d\mathbf{x}_1\cdots d\mathbf{x}_{N-n} H_{\text{out}}(\mathbf{x}_1)\cdots \times H_{\text{out}}(\mathbf{x}_{N-n}) e^{-\beta U(\mathbf{X},\mathbf{x}_1,\ldots,\mathbf{x}_{N-n})}, \tag{5}$$

the probability density of having no solvent molecule inside the subvolume for a supersystem with total number of solvent $N-n$, and calculate the factor

$$G_n^{(N)}(\mathbf{X},\mathbf{x}_1,\ldots,\mathbf{x}_n) \equiv \frac{P_n^{(N)}(\mathbf{X},\mathbf{x}_1,\ldots,\mathbf{x}_n)}{P_0^{(N-n)}(\mathbf{X})}, \tag{6}$$

which from Eqs. (4) and (5) equals

$$G_n^{(N)}(\mathbf{X},\mathbf{x}_1,\ldots,\mathbf{x}_n) = \frac{N!V^{N-n}}{n!(N-n)!V^N} e^{\beta[A_{\text{ex}}^{(N)}-A_{\text{ex}}^{(N-n)}]} \frac{\int_V d\mathbf{x}_{n+1}\cdots d\mathbf{x}_N H_{\text{out}}(\mathbf{x}_{n+1})\cdots H_{\text{out}}(\mathbf{x}_N) e^{-\beta U(\mathbf{X},\mathbf{x}_1,\ldots,\mathbf{x}_N)}}{\int_V d\mathbf{x}_{n+1}\cdots d\mathbf{x}_N H_{\text{out}}(\mathbf{x}_{n+1})\cdots H_{\text{out}}(\mathbf{x}_N) e^{-\beta U^*(\mathbf{X},\mathbf{x}_1,\ldots,\mathbf{x}_N)}} \equiv \frac{N!}{n!(N-n)!V^n} e^{\beta[A_{\text{ex}}^{(N)}-A_{\text{ex}}^{(N-n)}]} e^{-\beta W(\mathbf{X},\mathbf{x}_1,\ldots,\mathbf{x}_n)}, \tag{7}$$

where in the first line, the molecular indices $i=1,\ldots,N-n$ in Eq. (5) have been replaced by $i=n+1,\ldots,N$, and the potential energy $U^*(\mathbf{X},\mathbf{x}_1,\ldots,\mathbf{x}_N)$ is that of a reference system in which all interactions between the surrounding $N-n$ water molecules with the water molecules $i=1,\ldots,n$ in the subvolume have been turned off. Equation (7) defines the PMF W corresponding to the reversible work required to assemble the collection of n solvent molecules in the configuration $\{\mathbf{X},\mathbf{x}_1,\ldots,\mathbf{x}_n\}$ from a reference noninteracting state in the presence of the solute. Suitable approximations to such PMF have been previously developed on the basis of continuum electrostatic models.[11] The factor G_n is analogous to the "binding factor" introduced in Ref. 35. Taking the thermodynamic limit in which $N \gg n$, we have

$$G_n(\mathbf{X},\mathbf{x}_1,\ldots,\mathbf{x}_n) \simeq e^{n\beta\mu_{\text{ex}}}\frac{(\overline{\rho})^n}{n!} e^{-\beta W(\mathbf{X},\mathbf{x}_1,\ldots,\mathbf{x}_n)}, \tag{8}$$

where $\mu_{\text{ex}} = \partial A_{\text{ex}}^{(N)}/\partial N$ and $\overline{\rho}$ are the excess chemical potential and number density of the unperturbed *bulk* solvent, respectively. The superscript in $G_n^{(N)}$ has now been dropped in Eq. (8) since the right-hand side is independent of N. The usefulness of the binding factor G_n introduced in Eq. (7) rests on the fact that it is a function only of the property W of the subsystem containing n water molecules, and the intensive variables of the bulk solvent μ_{ex} and $\overline{\rho}$. As is evident from Eq. (6), G_n is proportional to the probability density $P_n^{(N)}$, and the latter can be recovered from G_n given by Eq. (8) via

$$P_n(\mathbf{X},\mathbf{x}_1,\ldots,\mathbf{x}_n) = \frac{G_n(\mathbf{X},\mathbf{x}_1,\ldots,\mathbf{x}_n)}{\Xi}, \tag{9}$$

where Ξ is an effective grand canonical partition function for the finite subvolume,

$$\Xi \equiv \sum_{n=0}^{\infty}\frac{(\overline{\rho})^n}{n!}\int d\mathbf{x}_1\cdots d\mathbf{x}_n H_{\text{in}}(\mathbf{x}_1)\cdots H_{\text{in}}(\mathbf{x}_n) \times e^{-\beta[W(\mathbf{X},\mathbf{x}_1,\ldots,\mathbf{x}_n)-n\mu_{\text{ex}}]} = \sum_{n=0}^{\infty}(\overline{\rho})^n K_n, \tag{10}$$

where the K_n play the role of effective equilibrium binding constants,

$$K_n = \frac{1}{n!}\int d\mathbf{x}_1 \cdots d\mathbf{x}_n H_{\mathrm{in}}(\mathbf{x}_1)\cdots H_{\mathrm{in}}(\mathbf{x}_n) \times e^{-\beta[W(\mathbf{X},\mathbf{x}_1,\ldots,\mathbf{x}_n)-n\mu_{\mathrm{ex}}]}. \quad (11)$$

In Eqs. (10) and (11), $H_{\mathrm{in}}=(1-H_{\mathrm{out}})$ is equal to 1 when a molecule is located within the subvolume v and 0 otherwise. We can also write the probability distribution P_n in the form familiar in the conventional formulation of the grand canonical ensemble:[20]

$$P_n(\mathbf{X},\mathbf{x}_1,\ldots,\mathbf{x}_n) \propto \frac{e^{nB}}{n!} e^{-\beta W(\mathbf{X},\mathbf{x}_1,\ldots,\mathbf{x}_n)}, \quad (12)$$

where $B \equiv \beta\mu_{\mathrm{ex}} + \ln \overline{n}$, and $\overline{n} = \overline{\rho} v$ is the expected number of water molecules inside a subvolume v in the *bulk phase*. The two parameters μ_{ex} and $\overline{\rho}$ are assumed to be known properties of the unperturbed bulk solvent. At room temperature and atmospheric pressure, $\overline{\rho} = 0.0334\ \text{Å}^{-3}$. For the excess chemical potential of water, we utilized the most recent estimate for the TIP3P water model,[36] $\mu_{\mathrm{ex}} \simeq -5.8$ kcal/mol,[37] although values ranging from -6.5 to -5.4 kcal/mol have been previously reported.[10,38–40] In contrast to bulk simulations, where μ_{ex} is estimated from the average number of molecules $\langle n \rangle$ from the simulation using a chosen value of B, in the current formalism, Eq. (12) applies to the inhomogeneous system which is assumed to be in equilibrium with the *unperturbed* bath. Notably, the average number of molecules from the actual simulations $\langle n \rangle$ is expected to differ from $\overline{n}$ because of the influence of the perturbing potential.

It is of interest to relate the current formalism to the quasichemical theory of solvation,[41–43] where the solvation free energy of a solute ΔG is expressed in terms of hierarchy of solute-solvent association constants. Let us introduce the thermodynamic coupling parameter $0 \leqslant \lambda \leqslant 1$ rescaling the interaction potential between the solute and all of the solvent molecules ($\lambda=1$ corresponds to the full interaction, whereas $\lambda=0$ turns off the interaction). The solvation free energy of the solute in a fixed configuration $\mathbf{X}$ is

$$\begin{aligned} e^{-\beta\Delta G} &= \langle e^{-\beta[U(\lambda=1)-U(0)]}\rangle_{(\lambda=0)} \\ &= \sum_n p_n(\lambda=0)\langle e^{-\beta[U(\lambda=1)-U(0)]}\rangle_{(\lambda=0,n)} \\ &= p_0(\lambda=0)\sum_n (\overline{\rho})^n K_n(\lambda=0) \\ &\quad \times \langle e^{-\beta[U(\lambda=1)-U(0)]}\rangle_{(\lambda=0,n)}, \end{aligned} \quad (13)$$

where $p_n(\lambda=0)$ represents the probability of having n solvent molecules in the subvolume v and $K_n(\lambda=0)$ are the binding constant defined from Eq. (11) in the absence of the solute. The solvation free energy of the solute can be expressed as,

$$e^{-\beta\Delta G} = p_0(\lambda=0) \times \sum_n (\overline{\rho})^n \frac{1}{n!}\int d\mathbf{x}_1 \cdots d\mathbf{x}_n H_{\mathrm{in}}(\mathbf{x}_1)\cdots H_{\mathrm{in}}(\mathbf{x}_n) \times e^{-\beta[W^{\mathrm{tot}}(\mathbf{X},\mathbf{x}_1,\ldots,\mathbf{x}_n)-n\mu_{\mathrm{ex}}]}, \quad (14)$$

where W^{tot} is an effective PMF representing the total free energy for turning on all the interactions involving the entire content of the subvolume (solute and solvent molecules). The configuration integrals with the W^{tot} can be redefined in terms of effective equilibrium constant of the association reaction of the solvent and solute molecules K_n^{tot} similar to Eq. (11). It may be noted that the probability $p_0(\lambda=0)$ corresponds to the spontaneous occurrence of configurations with zero solvent molecule in the subvolume in the absence of the solute. This is directly related to $\Delta G_{\mathrm{hs}} \equiv -k_B T \ln[p_0(\lambda=0)]$, the reversible thermodynamic work for inserting a hard-sphere overlapping exactly with the subvolume into the bulk solvent. The solvation free energy of the solute thus becomes

$$\Delta G = \Delta G_{\mathrm{hs}} - k_B T \ln\left[\sum_n (\overline{\rho})^n K_n^{\mathrm{tot}}\right]. \quad (15)$$

Equation (15) has the form of the basic expressions obtained from the quasichemical theory of solvation.[42,43] This formulation provides a framework to treat the m nearest solvent molecules located in the first coordination shell explicitly (perhaps using high level *ab initio* methods) and treat the influence of the remaining bulk solvent at a more approximate level (perhaps with continuum dielectric theory or molecular mechanical force fields),

$$\Delta G \approx \Delta G_{\mathrm{hs}} - k_B T \ln[(\overline{\rho})^m K_m^{\mathrm{tot}}]. \quad (16)$$

B. Monte Carlo algorithms

The acceptance probabilities for the GCMC simulation including the molecule insertion and deletion attempts can now be obtained from Eq. (12). Monte Carlo algorithms in general contain various kinds of move types (e.g., translation, rotation, isomerization), each designed to sample relevant regions of the configuration space in the most efficient manner. If a move corresponding to the transition $i \to j$ from a state i to state j is selected among all possible sets with the *selection* probability g_{ij}, and accepted with the *acceptance* probability A_{ij}, the overall *transition* probability P_{ij} is required to satisfy the detailed balance condition

$$\frac{P_{ij}}{P_{ji}} = \frac{g_{ij}A_{ij}}{g_{ji}A_{ji}} = \frac{p_j}{p_i}, \quad (17)$$

where p_i is the equilibrium probability of the state i. A rearrangement of Eq. (17) gives

$$\frac{A_{ij}}{A_{ji}} = \frac{g_{ji}p_j}{g_{ij}p_i}, \quad (18)$$

and the Metropolis scheme amounts to the choice

$$A_{ij} = \min\left\{1, \frac{g_{ji}p_j}{g_{ij}p_i}\right\}. \quad (19)$$

For the translational and rotational moves, the selection probability g_{ij} is constant for all possible moves. Adapting the Metropolis algorithm with Eqs. (12) and (19) ($\lambda=1$ assumed in the following), the acceptance probability becomes

$$A(\mathbf{x}_i \to \mathbf{x}_i') = \min\{1, e^{-\beta[W(\mathbf{X},\mathbf{x}_i')-W(\mathbf{X},\mathbf{x}_i)]}\}. \quad (20)$$

For the GCMC attempts, we first assume that the orientation of a newly inserted molecule is generated randomly, and consider the acceptance probabilities for the GCMC attempts. An insertion attempt $n \to n+1$, where the number of water molecules inside the system is increased by one, can be regarded as a transfer of a molecule from the bulk reservoir to a particular position inside the GCMC volume. The forward transition involves a choice of the insertion position, and denoting the volume of the subspace in which the insertion is attempted as $v' \leqslant v$, we can write $g_{n,n+1} = a/v'$, where a is a microscopic unit of volume discretizing the space. In the backward transition, a molecule in the system is transfered to an arbitrary position in the homogeneous reservoir within a volume v, and $g_{n+1,n} = a/v$. An analogous reasoning applies to the corresponding probability for the deletion attempt $n \to n-1$, and from Eqs. (12) and (19) we obtain the insertion and deletion acceptance probabilities:

$$A_{n,n+1} = \min\left\{1, \frac{f_n}{n+1} e^{B - \beta W(\mathbf{X}, \mathbf{x}_i^{n+1}) + \beta W(\mathbf{X}, \mathbf{x}_i^n)}\right\}, \quad (21a)$$

$$A_{n,n-1} = \min\left\{1, \frac{n}{f_{n-1}} e^{-B - \beta W(\mathbf{X}, \mathbf{x}_i^{n-1}) + \beta W(\mathbf{X}, \mathbf{x}_i^n)}\right\}, \quad (21b)$$

where $W(\mathbf{X}, \mathbf{x}_i^n)$ denotes the energy and the potential of mean force of the given configuration with n molecules in the system, and $f_n = v'/v$ is the fractional volume of the subspace in which the insertion attempts are made.

The simplest choice for the selection probability, $f_n = 1$ in Eq. (21) corresponds to the "random insertion" algorithm, where $v' = v$ and insertion attempts are made uniformly throughout the GCMC volume by generating a random position and orientation of the new water molecule. The random insertion algorithm is generally rather inefficient in achieving equilibrations since most attempts to insert a molecule into a dense system are rejected. A number of modified algorithms to facilitate the insertion attempts in the bulk GCMC simulation have been developed by various authors.[21–26] In the "cavity-bias algorithm" of Mezei,[21] a set of candidate insertion points are generated within the volume v before an insertion attempt, and the points not within a preset distance σ_{ex} from any of the existing molecular centers (e.g., water oxygens or other heavy atoms of the protein backbones) are identified as the cavities. The insertion is then attempted only for one of the cavity sites. The fraction f_n of the cavity sites to the total number of sites generated gives an approximation to the fractional volume of the free available space in which a sphere of radius σ_{ex} would not overlap with any other molecular centers. The cavity probability f_n is calculated concurrently during the simulation by collecting statistics. In cases where no cavity sites exist, a random insertion is attempted, which requires that the deletion attempts should be accepted with the biased probability (21b), or with the simple unbiased probability with $f_n = 1$, the latter chosen with f_n^{nc}, the probability that no cavity sites exist in a configuration with the total number of solvent molecules n. The statistics for f_n^{nc} is also collected during the simulation.

The cavity-bias algorithm, where random generations of cavity candidate points at each insertion attempts are followed by the identification of the cavities through overlap checks with all present molecules, becomes time consuming at high densities. Under such conditions, it becomes more advantageous to generate and maintain a dynamical lattice ("grid") variable $\{s_{\mathbf{x}}\}$, defined on each discretized lattice sites spanning the volume v.[22–24] The grid variable at each step of the simulation is defined as the total number of nearby molecules whose centers are within the distance σ_{ex} from the lattice site. Its values are thus non-negative integers, and the sites denoting available space are those with $s_{\mathbf{x}} = 0$. The grid-insertion algorithm thus utilizes the same acceptance probability (21), but with f_n as a dynamical variable given by the ratio of the number of cavity lattice sites to the total number of sites. The "no-cavity" probability f_n^{nc} is calculated as before by collecting statistics during the course of the simulation. Each insertion is attempted on a random position within the cubic volume corresponding to one of the cavity lattice sites picked from the dynamically updated list of cavities, and the repeated overlap-checking becomes unnecessary. The process of updating the grid variable is illustrated schematically in Fig. 2.

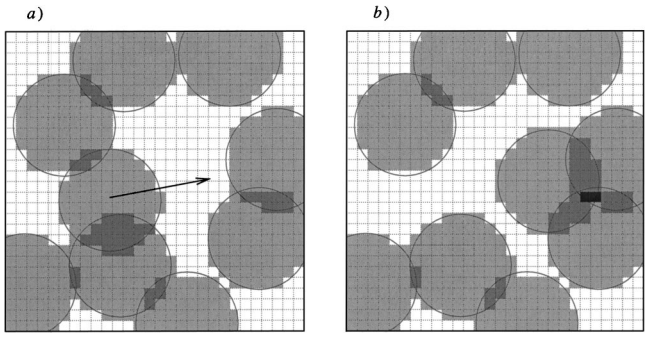

FIG. 2. Schematic illustrations of the grid variable update upon an attempt in the simulation of a monatonic fluid. The lattice sites on which $s_i > 0$ are shown as shaded, with increasing darkness levels representing higher values of s_i. On an insertion, a cavity site (shown as white) is selected randomly for the molecular center, and if the attempt is accepted, s_i is increased by 1 for sites whose centers lie within distance σ_{ex}. Upon a deletion attempt, one of the existing molecular centers is selected randomly, and s_i is decreased by 1 within the surrounding sphere. A rejected deletion attempt is followed by the restoration of s_i. A translational move entails (a) a deletion update followed by (b) an insertion update.

The grid variable update during the grid-insertion simulation is done as follows: the grid variables are initialized by setting them as 0 and looping over the molecular centers $\{\mathbf{x}_i\}$ present while increasing $s_{\mathbf{x}}$ by 1 for sites for which $|\mathbf{x}_i - \mathbf{x}| < \sigma_{ex}$ (Fig. 2). The molecular centers also include any non-GCMC atoms such as protein chains. An additional dynamical array is created, which contains the lattice coordinates of the cavity sites where $s_{\mathbf{x}} = 0$. For an insertion, a cavity site is chosen randomly from the cavity list, the attempt is accepted with the probability (21a) with the current value of f_n, and $s_{\mathbf{x}}$ is increased by 1 for those sites within the distance σ_{ex} from the insertion site. For a deletion, a molecular center is chosen randomly, $s_{\mathbf{x}}$ is decreased by 1 temporarily within the σ_{ex}-sphere, f_{n-1} is calculated, and the attempt is accepted with the probability (21b). If the attempt is rejected, the grid variable is restored to its original state. For non-GCMC translational moves, the grid is updated as if the molecule

had been deleted from its old position and reinserted at a new position.

For a molecular solvent such as water on which we focus our attention, the configurational space to be chosen for an insertion attempt consists not only of the translational but also of orientational degrees of freedom. Even with an optimal choice for the insertion site, randomly chosen orientations would most likely result in rejections in dense conditions. The orientational-bias algorithm[25,26] can be adapted as a non-Boltzmann sampling in the orientational space to further increase the efficiency of the simulation. For each insertion attempt, after the insertion center coordinate is chosen for example by the grid-insertion method, a set of m candidate orientations for the new molecule is generated. One orientation is selected with probability $p_l = e^{-\beta u_l}/\sum_{k=1}^{m} e^{-\beta u_k}$ where u_k is the orientational contribution to the energy change upon insertion. The forward selection probability is $g_{n,n+1} = a p_l / v'$, whereas for the backward transition $n+1 \rightarrow n$, $g_{n+1,n} = a/mv$ since the reservoir is homogeneous and m possible orientations for the transferred water molecule are equally likely. For a deletion attempt, $m-1$ orientations are generated in addition to that of the chosen molecule, and the relative probability of the current orientation p_l is calculated. It can be easily verified from an analogous reasoning for the selection probabilities that the overall acceptance probabilities of the insertion and deletion attempts with the orientational bias can be written as Eqs. (21) with $f_n = v'/vmp_l$. The GCMC simulation algorithms as described above were implemented within the Monte Carlo module[44] of the biomolecular simulation program CHARMM,[45] and tested by calculating and comparing the probability distributions of the molecule number statistics $P(n)$ with bulk argon and water at low densities.

III. APPLICATIONS

In this section, we describe the implementation of the algorithms for the GCMC simulations of water in a protein/solute environment, and test it with a few simple model systems. To perform GCMC simulations with Eqs. (21), an appropriate method to determine the potential of mean force W for the system under study is necessary. If one chooses periodic boundary conditions, the GCMC volume v becomes the total volume of the cell, and W can be replaced with the potential energy (since there is effectively no other influence on the system). Figure 3 shows an example of the results of GCMC simulations of bulk water in periodic boundary conditions with various insertion algorithms, plotted as functions of both the Monte Carlo steps and the CPU time on a 1.8 GHz Pentium chip. The grid-insertion algorithm with orientational bias is seen to perform best. The GCMC simulations accurately reproduce the correct number of water molecules $\langle n \rangle = \bar{n}$ for the given set of input parameters, $\rho_0 = 0.0334\ \text{Å}^{-3}$ and $\mu_{ex} = -5.8$ kcal/mol, which correspond to the properties of the TIP3P water model. The GCMC with dynamical-grid insertion-deletion method is very efficient. From Fig. 3, we estimate that the characteristic relaxation time of GCMC simulations of pure water with orientational bias is roughly 2–3 times smaller than the simple grid insertion, whereas that of random insertion appears to be much larger. A more detailed study of the equilibration time of the algorithms could be performed for example by the method of Yang, Bitetti-Putzer, and Karplus[46] for free energy simulations.

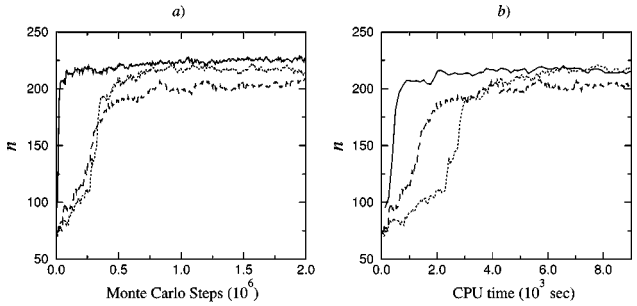

FIG. 3. Convergences of the number of water molecules n inside a cubic box of length 18.856 Å in periodic boundary conditions for $\mu_{ex} = -5.8$ kcal/mol, $T = 298$ K, and $\rho_0 = 0.0334\ \text{Å}^{-3}$, as functions of the Monte Carlo steps and the CPU time on a 1.8 GHz Pentium chip. The dashed, dotted, and solid lines are results with random insertion, grid insertion, and grid insertion with orientational bias, respectively. The expected number of water molecules is $\bar{n} = 224$.

In biological systems, one is often interested in considering only a small region within a large system containing proteins, water, and other ligands. The use of an effective solvent boundary potentials, which incorporate the influence of the distant solvent or solute molecules, is necessary. As a test of the use of GCMC algorithm, we consider the GCMC simulation of a small sphere of water, using GSBP.[11] In GSBP, the collective electrostatic effects of the solvent and solute atoms outside a fixed central volume (static field and reaction field) is approximated assuming that the surrounding solvent can be represented by a featureless dielectric continuum. Both the electrostatic static field and the reaction field (expressed in the form of expansion coefficients for a basis set) are calculated and stored prior to the simulation for efficiency.[11] The empirical function based on the reference interaction site model calculations[10] can be used for the nonpolar contribution to the PMF. Since the boundary of the central region is fixed during the simulation, the number of explicit solvent water in the subvolume fluctuates according to the statistics of the grand canonical ensemble, and the use of GCMC simulations is necessary.

As a test of the GCMC algorithm with GSBP, we simulated a water sphere of radius $R_s = 10$ Å using the three different insertion algorithms (Figs. 4 and 5). Compared to the random insertion method, the cavity-bias and the dynamical-grid algorithms considerably enhance the acceptance probabilities of the GCMC insertion-deletion attempts, which were found to be 0.20%, 0.70%, 0.85% for the three algorithms, respectively. The relative efficiencies of the grid-insertion versus cavity bias are expected to depend on the parameter σ_{ex} (Ref. 47) as well as the number of cavity candidate points generated for each insertion attempts. With the same value of σ_{ex} as in Fig. 5, compared with respect to the number of Monte Carlo steps, the grid insertion is expected to be slightly more efficient since cubic grids tile the space completely whereas spheres do not.

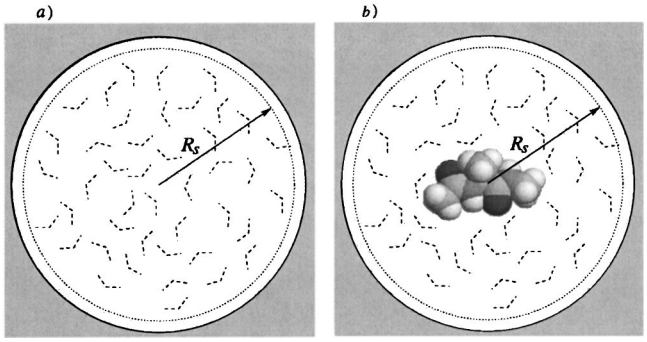

FIG. 4. A spherical system of radius R_s containing explicit water molecules is simulated with GCMC and the influence of the surrounding solvent is approximated by the GSBP.[11] GSBP includes a reaction field arising from the polarization of the dielectric continuum at $r > R_s + \Delta r_{\text{diel}}$, where $\Delta r_{\text{diel}} = 2.8$ Å. Water insertions and deletions are attempted only within the sphere $r < R_s$. (a) Pure water sphere; (b) water sphere solvating a small peptide.

With respect to the computational time required, the grid insertion is considerably more efficient for densely packed systems where $f_n \ll 1$. As an example of the use with a small solute solvated by the water molecules inserted by GCMC with GSBP, Fig. 5 also shows similar results with a dialanine as the solute at the center of the sphere.

As a last illustration of water solvating a protein cavity, we consider the KcsA potassium channel. The KcsA channel is a membrane protein tetramer with a narrow selectivity filter on the extracellular side, and a wide nonpolar cavity region.[4,48] The flux of ions through the narrow selectivity filter into and out of the central cavity is achieved by alternating sequences of ion occupation states at four locally stable positions along the filter.[49,50] Although the nonpolar cavity, containing a potassium ion at the center, is large enough to contain tens of water molecules, only a few are sufficiently ordered to be visible in the crystal structure.[51] The proper equilibration of the number of water molecules inside the pore is highly nontrivial, since it can only be achieved physically via the diffusion of water molecules through the narrow inner vestibule, which would only occur within a time scale inaccessible with brute-force MD simulations.[52,53] The GCMC simulation therefore provides an ideal method to obtain well-equilibrated solvated structures of the channel pore. The crystallization condition of the channel includes the use of salts and a 40%–50% polyethylene-glycol solution,[4] implying that the excess chemical potential of water would be lower than in the bulk at ambient conditions. To approximate these effects, we used a value of excess chemical potential of $\mu_{\text{ex}} = -6.5$ kcal/mol for the GCMC simulations of water in the central cavity of the potassium channel.

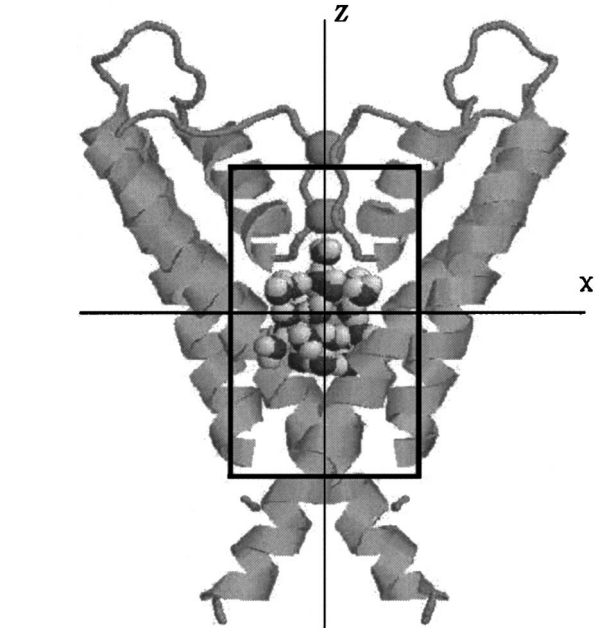

FIG. 6. GCMC simulation box of the KcsA channel. The potassium ions fixed at the crystal coordinates during the simulation are shown as spheres. Only two channel backbones are shown with ribbons, while the crystal waters outside are not shown. GCMC simulations were run without any crystal waters inside the box. Space-filled waters (29 in total) inside show a typical configuration obtained with the GCMC.

The GCMC region of volume v has been chosen as a rectangular box such that $|x| \leqslant 8.5$ Å $|y| \leqslant 8.5$ Å and $|z| \leqslant 15$ Å (Fig. 6). As illustrated in Fig. 7, much of the space within the box is occupied with the channel atoms as well as the water molecules already present. The grid-insertion method facilitates successful insertion attempts by efficiently

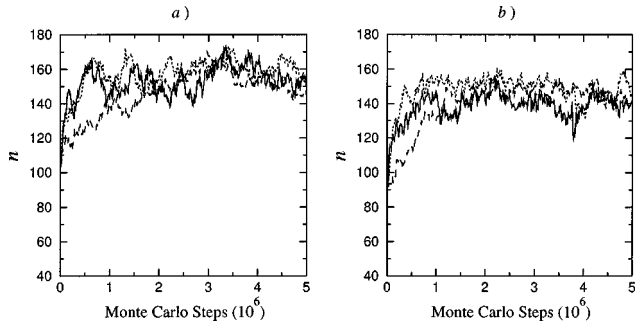

FIG. 5. Convergences of number of water molecules n inside a sphere of radius $R_s = 10$ Å simulated with the various insertion algorithms with GSBP potential; (a) pure water sphere; (b) water sphere containing a dialanine at the center. Solid, dotted, and dashed lines are results with grid-insertion/orientational bias, cavity bias, and random insertion, respectively. See Fig. 4.

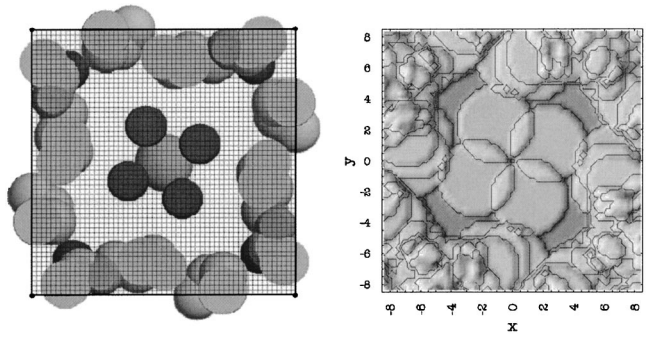

FIG. 7. Grid variable of a configuration of the KcsA channel (crystal structure) within a plane at $z = 3$ Å (see Fig. 6). The image on the left shows the nonhydrogen atoms whose centers lie within the distance $\sigma_{\text{ex}} = 2.5$ Å below the plane. Four water oxygens are shown around the central potassium ion. The contours on the right represent the grid variables $s_{\mathbf{x}} = 0,...,5$. The cavity sites are the grids with $s_{\mathbf{x}} = 0$, the lowest contour level shown. The grid size corresponds to 0.25 Å. Insertions are attempted only within the cavity sites.

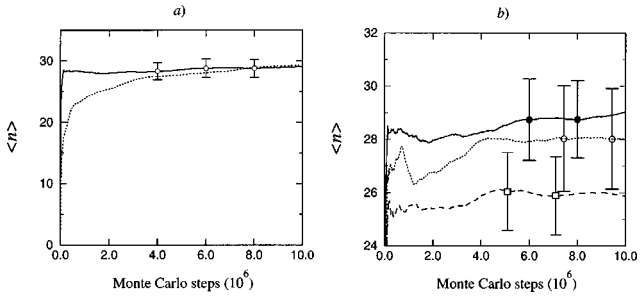

FIG. 8. Cumulative averages of the number of water $\langle n \rangle$ within the central pore of the KcsA potassium channel from the GCMC simulations. The channel atoms as well as the potassium ions were fixed during the simulation. Solid and dotted lines in (a) are the results obtained from the grid-insertion algorithm with orientational bias, and from the random insertion, respectively. In (b), the results with (solid line) and without (dotted line) the potassium ion in the cavity, and with its charge set to zero (dashed line), are compared using the grid-insertion method. The error bars represent standard deviations.

keeping track of the complex geometry of the free space available during the simulation. Figure 8 shows the equilibration behavior of typical GCMC simulations of water in the central pore region of the KcsA channel. The acceptance ratios of the GCMC attempts were 0.06% and 0.81% for the random insertion, and grid insertion with orientational bias, respectively. In total, 29 water molecules fill the cavity with the potassium ion, whereas in the absence of the ion inside, the cavity accommodates 28. Shown together in Fig. 8 is the result with the central ion, but with its charge set to zero. Both the electrostatic stabilizations as well as the steric effects due to the central ion seem to play roles in the equilibrium configurations of solvating water molecules. In addition, the overall number of water molecules is expected to depend on the bulk excess chemical potential of water molecules in the mother liquor bathing the crystals. One simulation in Fig. 8 for 10^6 Monte Carlo steps took $\approx$6 h of CPU time on a 1.8 GHz Pentium chip. By comparison, prohibitively long MD or canonical Monte Carlo simulations would be necessary to allow equivalent levels of equilibration of the number of water molecules if they can only enter/leave the pore via diffusion.

IV. SUMMARY

In this paper, we have formulated the GCMC algorithm in a form most appropriate to simulate the solvation of a finite subvolume in open equilibrium with a large reservoir of bulk liquid. The GCMC algorithm with a dynamical-grid technique was implemented and tested with a few simple model systems. The excess chemical potential and the number density of the bulk water serve as known input parameters, and the inhomogeneous water density distribution inside the simulation volume in equilibrium with the bulk water reservoir is obtained as a result of the simulation. The grid-insertion algorithm combined with the orientational bias facilitates the overall efficiency of the simulation considerably, especially in the presence of a solute.

As a possible application of the simulation methodology presented, it will be of interest to combine the GCMC simulation with the alchemical free energy perturbation (FEP) method[8,37,43,54] where two different states of a system are connected through unphysical intermediates using a thermodynamic coupling parameter λ. While FEP calculations can be safely performed with a fixed number of solvent water molecules when a large volume of solvent is simulated, a fixed number of water molecules can create difficulties when FEP calculations are carried out with finite systems as depicted in Fig. 1(b). For such systems, a mixture of the GCMC and MD simulations allowing the number of water to adapt and vary during the simulations of the various intermediate λ states could provide a more accurate and realistic results.

[1] C. L. Brooks III, B. M. Pettitt, and M. Karplus, in *Proteins: A Theoretical Perspective of Dynamics, Structure, and Thermodynamics*, Advances in Chemical Physics Vol. 71, edited by I. Prigogine and S. A. Rice (Wiley, New York, 1988).
[2] F. Merzel and J. C. Smith, Proc. Natl. Acad. Sci. U.S.A. **99**, 5378 (2002).
[3] G. N. Phillips, Jr and B. M. Pettitt, Protein Sci. **4**, 149 (1995).
[4] D. A. Doyle *et al.*, Science **280**, 69 (1998).
[5] R. M. Stroud *et al.*, Curr. Opin. Struct. Biol. **13**, 424 (2003).
[6] M. S. P. Sansom, I. D. Kerr, J. Breed, and R. Sankararamakrishnan, Biophys. J. **70**, 693 (1996).
[7] B. Roux and T. Simonson, Biophys. Chem. **78**, 1 (1999).
[8] P. Kollman, Chem. Rev. (Washington, D.C.) **93**, 2395 (1993).
[9] B. Honig, Curr. Opin. Struct. Biol. **5**, 203 (1995).
[10] D. Beglov and B. Roux, J. Chem. Phys. **100**, 9050 (1994).
[11] W. Im, S. Bernèche, and B. Roux, J. Chem. Phys. **114**, 2924 (2001).
[12] N. K. Banavali, W. Im, and B. Roux, J. Chem. Phys. **117**, 7381 (2002).
[13] M. Berkowitz and J. A. McCammon, Chem. Phys. Lett. **90**, 215 (1982).
[14] C. L. Brooks III and M. Karplus, J. Chem. Phys. **79**, 6312 (1983).
[15] A. Brunger, C. L. Brooks III, and M. Karplus, Chem. Phys. Lett. **105**, 495 (1984).
[16] J. W. Essex and W. L. Jorgensen, J. Comput. Chem. **16**, 951 (1995).
[17] S. T. Russell and A. Warshel, J. Mol. Biol. **185**, 389 (1985).
[18] J. Aqvist and A. Warshel, Biophys. J. **56**, 171 (1989).
[19] J. A. C. Rullmann and P. T. Vanduijnen, Mol. Phys. **61**, 293 (1987).
[20] D. J. Adams, Mol. Phys. **29**, 307 (1975).
[21] M. Mezei, Mol. Phys. **40**, 901 (1980).
[22] M. Mezei, Mol. Phys. **47**, 1307 (1982).
[23] G. L. Deitrick, L. E. Scriven, and H. T. Davis, J. Chem. Phys. **90**, 2370 (1989).
[24] M. R. Stapleton and A. Z. Panagiotopoulos, J. Chem. Phys. **92**, 1285 (1990).
[25] D. Frenkel and B. Smit, Mol. Phys. **75**, 983 (1992).
[26] J. C. Shelley and G. N. Patey, J. Chem. Phys. **102**, 7656 (1995).
[27] H. Resat, M. Mezei, and J. A. McCammon, J. Phys. Chem. **100**, 1426 (1996).
[28] P. Jedlovszky and M. Mezei, J. Chem. Phys. **111**, 10770 (1999).
[29] H. Resat and M. Mezei, J. Am. Chem. Soc. **116**, 7451 (1994).
[30] H. Resat and M. Mezei, Biophys. J. **71**, 1179 (1996).
[31] G. S. Heffelfinger and D. M. Ford, Mol. Phys. **94**, 659 (1998).
[32] P. I. Pohl and G. S. Heffelfinger, J. Membr. Sci. **155**, 1 (1999).
[33] A. P. Thompson and G. S. Heffelfinger, J. Chem. Phys. **110**, 10693 (1999).
[34] W. Im, S. Seefeld, and B. Roux, Biophys. J. **79**, 788 (2000).
[35] B. Roux, Biophys. J. **77**, 139 (1999).
[36] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, J. Chem. Phys. **79**, 926 (1983).
[37] Y. Deng and B. Roux J. Phys. Chem. B (in press).
[38] J. Hermans, A. Pathiaseril, and A. Anderson, J. Am. Chem. Soc. **110**, 5982 (1988).
[39] R. Pomes, E. Eisenmesser, C. B. Post, and B. Roux, J. Chem. Phys. **111**, 3387 (1999).
[40] W. L. Jorgensen, J. F. Blake, and J. K. Buckner, Chem. Phys. **129**, 193 (1989).
[41] L. R. Pratt and D. Chandler, J. Chem. Phys. **67**, 3683 (1977).
[42] G. Hummer, S. Garde, A. E. García, and L. R. Pratt, Chem. Phys. **258**, 349 (2000).
[43] L. R. Pratt, Annu. Rev. Phys. Chem. **53**, 409 (2002).
[44] A. R. Dinner, PhD thesis, Harvard University, 1999.
[45] B. R. Brooks *et al.*, J. Comput. Chem. **4**, 187 (1983).

[46] W. Yang, R. Bitetti-Putzer, and M. Karplus, J. Chem. Phys. **120**, 2618 (2004).
[47] S. H. Northrup and J. A. McCammon, J. Chem. Phys. **72**, 4569 (1980).
[48] B. Roux and R. MacKinnon, Science **285**, 100 (1999).
[49] J. H. Morais-Cabral, Y. F. Zhou, and R. MacKinnon, Nature (London) **414**, 37 (2001).
[50] S. Bernèche and B. Roux, Nature (London) **414**, 73 (2001).
[51] Y. Zhou, J. H. Morais-Cabral, A. Kaufman, and R. MacKinnon, Nature (London) **414**, 43 (2001).
[52] S. Bernèche and B. Roux, Biophys. J. **78**, 2900 (2000).
[53] I. H. Shrivastava and M. S. P. Sansom, Biophys. J. **78**, 557 (2000).
[54] J. A. McCammon, Curr. Opin. Struct. Biol. **8**, 245 (1998).