Article

Improved Efficiency of Replica Exchange Simulations through Use of a Hybrid Explicit/Implicit Solvation Model

Asim Okur, Lauren Wickstrom, Melinda Layten, Raphel Geney, Kun Song, Viktor Hornak, and Carlos Simmerling

J. Chem. Theory Comput., **2006**, 2 (2), 420-433 • DOI: 10.1021/ct050196z

JCTC Journal of Chemical Theory and Computation

Improved Efficiency of Replica Exchange Simulations through Use of a Hybrid Explicit/Implicit Solvation Model

Asim Okur,[†] Lauren Wickstrom,[‡] Melinda Layten,[§] Raphäel Geney,[†] Kun Song,[†] Viktor Hornak,[||] and Carlos Simmerling*[,†,‡,||,⊥]

Department of Chemistry, Graduate Program in Biochemistry and Structural Biology, Graduate Program in Molecular and Cellular Biology, and Center for Structural Biology, Stony Brook University, Stony Brook, New York 11794, and Computational Science Center, Brookhaven National Laboratory, Upton, New York 11973

Received August 5, 2005

Abstract: The use of parallel tempering or replica exchange molecular dynamics (REMD) simulations has facilitated the exploration of free energy landscapes for complex molecular systems, but application to large systems is hampered by the scaling of the number of required replicas with increasing system size. Use of continuum solvent models reduces system size and replica requirements, but these have been shown to provide poor results in many cases, including overstabilization of ion pairs and secondary structure bias. Hybrid explicit/continuum solvent models can overcome some of these problems through an explicit representation of water molecules in the first solvation shells, but these methods typically require restraints on the solvent molecules and show artifacts in water properties due to the solvation interface. We propose an REMD variant in which the simulations are performed with a fully explicit solvent, but the calculation of exchange probability is carried out using a hybrid model, with the solvation shells calculated on the fly during the fully solvated simulation. The resulting reduction in the perceived system size in the REMD exchange calculation provides a dramatic decrease in the computational cost of REMD, while maintaining a very good agreement with results obtained from the standard explicit solvent REMD. We applied several standard and hybrid REMD methods with different solvent models to alanine polymers of 1, 3, and 10 residues, obtaining ensembles that were essentially independent of the initial conformation, even with explicit solvation. Use of only a continuum model without a shell of explicit water provided poor results for Ala_3 and Ala_{10}, with a significant bias in favor of the α-helix. Likewise, using only the solvation shells and no continuum model resulted in ensembles that differed significantly from the standard explicit solvent data. Ensembles obtained from hybrid REMD are in very close agreement with explicit solvent data, predominantly populating polyproline II conformations. Inclusion of a second shell of explicit solvent was found to be unnecessary for these peptides.

Introduction

The potential energy surfaces of biological systems have long been recognized to be rugged, hindering conformational transitions between various local minima. This sampling problem can preclude success even when a sufficiently accurate Hamiltonian of the system is used in the simulations. Thus, a significant effort has been put into devising efficient simulation strategies that locate low-energy minima for these

* Corresponding author e-mail: carlos.simmerling@stonybrook.edu.
[†] Department of Chemistry, Stony Brook University.
[‡] Graduate Program in Biochemistry and Structural Biology, Stony Brook University.
[§] Graduate Program in Molecular and Cellular Biology, Stony Brook University.
[||] Center for Structural Biology, Stony Brook University.
[⊥] Brookhaven National Laboratory.

10.1021/ct050196z CCC: $33.50
Published on Web 02/04/2006

complex systems. Conformational sampling was recently reviewed[1] and is also the subject of a recent special journal issue.[2]

One approach that has seen a recent increase in the use of biomolecular simulation is the replica exchange method.[3−5] In replica exchange molecular dynamics (REMD)[6] (also called parallel tempering[3]), a series of molecular dynamics simulations (replicas) are performed for the system of interest. In the original form of REMD, each replica is an independent realization of the system, coupled to a heat bath at a different temperature. The temperatures of the replicas span a range from low values of interest (such as 280 K or 300 K) up to high values (such as 600 K) at which the system can rapidly overcome potential energy barriers that would otherwise impede conformational transitions on the time scale simulated.

At intervals during the otherwise standard simulations, conformations of the system being sampled at different temperatures are exchanged based on a Metropolis-type criterion[7] that considers the probability of sampling each conformation at the alternate temperature (described in more detail in Methods). In this manner, REMD is hampered to a lesser degree by the local minima problem, since simulations at low temperature can escape kinetic traps by "jumping" directly to alternate minima being sampled at higher temperatures. Likewise, the structures sampled at high temperatures can anneal by being transferred to successively lower temperatures. Moreover, the transition probability is constructed such that the canonical ensemble properties are maintained during each simulation, thus providing potentially useful information about conformational probabilities as a function of temperature. Due to these advantages, REMD has been applied to studies of peptide and small protein folding.[3,6,8−16]

For large systems, however, REMD becomes intractable since the number of replicas needed to span a given temperature range increases with the square root of the number of degrees of freedom in the system.[17−20] Several promising techniques have been proposed[19,21−23] to deal with this apparent disadvantage to REM.

The method chosen to treat solvent effects can have a direct impact on the system size and thus the computational requirement of employing REMD. Explicit representation of solvent molecules significantly increases the number of atoms in the simulated system, particularly when the solvent box is made large enough to enclose unfolded conformations of peptides and proteins. The growth in system size results in the need for many more replicas to span the same temperature range. This increase in computational cost is in addition to that added by the need to calculate forces and integrate equations of motion for the explicit solvent molecules.

Continuum solvent models such as the semianalytical Generalized Born (GB) model[24] estimate the free energy of solvation of the solute based on coordinates of the solute atoms. The neglect of explicit solvent molecules can significantly reduce the computational cost of evaluating energies and forces for the system, but a larger effect with REMD can arise from the reduction in the number of replicas due to the fewer degrees of freedom. This factor can determine whether REMD is a practical approach to model the system. For example, in the 10-residue peptide model presented below, 40 replicas are needed when the solvent is included explicitly, while only 8 are sufficient for the same peptide with a continuum solvent model. Larger systems would be expected to show even greater differences; the number of peptide atoms increases approximately linearly with sequence length, while the volume of a sphere (and thus the number of solvent atoms) needed to enclose extended conformations increases with the peptide length to the third power. Thus one can roughly estimate that the difference in number of replicas required for explicit vs continuum solvation of a system will increase with the number of solute degrees of freedom to the $^3/_2$ power.

Continuum solvent models are thus an attractive approach to enabling the study of larger systems with REMD. Among the various models that have been developed, the GB approach is commonly used with molecular dynamics due to its computational efficiency, permitting use at each time step. However, these models can also have significant limitations. Since the atomic detail of the solvent is not considered, modeling specific effects of structured water molecules can be challenging. In the case of protein and peptide folding, it appears likely that the current generation of GB models do not have as good a balance between protein−protein and protein−solvent interactions as do the more widely tested explicit solvent models.[25,26] More particularly, it has been reported[12,26−28] that ion pairs were frequently too stable in the GB implicit water model, causing salt bridged conformations to be oversampled in MD simulations, thus altering the thermodynamics and kinetics of folding for small peptides. A clear illustration was given by Zhou and Berne[26] who sampled the C-terminal β-hairpin of protein G (GB1) with both a surface-GB (SGB)[29] continuum model and an explicit solvent. The lowest free energy state with SGB was significantly different from the lowest free energy state in the explicit solvent, with incorrect salt bridges formed at the core of the peptide, in place of hydrophobic contacts. Zhou extended this study on GB1 by examining several force field-GB model combinations, with all GB models tested showing erroneous salt-bridges.[27]

The more rigorous models based on Poisson−Boltzmann (PB) equations are generally considered to be more accurate. Historically, the increased cost of evaluating solvation free energy with these methods results in their use primarily to postprocess a small number of conformations, or snapshots sampled during an MD simulation in the explicit solvent.[30] However, some researchers have reported using PB as a solvent model for molecular dynamics simulation.[31,32] PB approaches do not necessarily overcome the difficulty of modeling nonbulk effects in the first solvation shells.

To benefit from the efficiency of implicit solvents while incorporating these first shell effects, several hybrid explicit/implicit models have been proposed. These typically employ the explicit solvent only for the first 1−2 solvation shells of the solute, often surrounded by a continuum representation of various types.[33−45] However, these methods have draw-

backs in that the explicit water typically must be restrained to remain close to the solute to avoid diffusion into the "bulk" continuum. These restraints as well as the boundary effects at the explicit/implicit interface can have a dramatic effect on solute behavior. In a recent implementation, Lee et al. employed a hybrid TIP3P/GB solvation model with excellent results,[41] but they pointed out drawbacks typical for these models, such as the need for a fixed solute volume and shape for the solvation cavity, preventing large-scale conformational changes of the type that is necessary for detailed analysis of conformational ensembles using enhanced sampling techniques such as REMD. In addition, they demonstrated that solvent properties such as radial density and dipole distributions showed significant artifacts due to boundary effects.

Recognizing that the main difficulty in applying REMD with the explicit solvent lies in the number of simulations required, rather than just the complexity of each simulation, we propose a new approach in which each replica is simulated in the explicit solvent using standard methods such as periodic boundary conditions and inclusion of long-range electrostatic interactions. However, the calculation of exchange probabilities (which determines the temperature spacing and thus the number of replicas) is handled differently. Only a subset of closest water molecules is retained, while the remainder is *temporarily* replaced by a continuum representation. The energy is calculated using the hybrid model, and the exchange probability is determined. The original solvent coordinates are then restored, and the simulation proceeds as a continuous trajectory with fully explicit solvation. This way the perceived system size for evaluation of exchange probability is dramatically reduced and fewer replicas are needed.

An important difference from the existing hybrid models is that our system is fully solvated throughout the entire simulation, and thus the distribution functions and solvent properties should not be affected by the use of the hybrid model in the exchange calculation. In addition, no restraints of any type are needed for the solvent, and the solute shape and volume may change since the solvation shells are generated for each replica on the fly at every exchange calculation. Nearly no computational overhead is involved since the calculation is performed infrequently as compared to the normal force evaluations. Thus the hybrid REMD approach can employ more accurate continuum models that are too computationally demanding for use in each time step of a standard molecular dynamics simulation.

In this study we have tested the hybrid REMD method on varying lengths of polyalanine peptides (dipeptide, tetrapeptide, and Ala_{10}). Many helical design studies have used polyalanines with charged residues,[46−48] N-capping,[49] and C-capping interactions[50] to solubilize the peptides and stabilize helical structure. Recently, experimental studies with CD, NMR, and UV resonance Raman have been able to characterize a primarily polyproline type II (P_{II}) structure in short polyalanines[51−53] and in the denatured state of longer alanine peptides.[54] MD simulations of polyalanines have further substantiated these experimental observations.[38,55] The quality of the solvent model is expected to be critically important since it has been proposed that specific solvation of backbone amide groups plays a key role in the stabilization of P_{II} conformations.[55,56]

For each peptide we first obtained conformation ensembles using standard REMD in explicit solvent. We used these data as a reference in order to remove the influence of the protein force field parameters from this study of solvation models. For each sequence, two sets of REMD simulations in the explicit solvent were run with different initial conformations until convergence was indicated by reasonable agreement between the data sets. For example, the populations of conformation clusters in the two Ala_{10} runs in the TIP3P solvent were highly correlated (R^2=0.974), demonstrating high similarity not only in the types of structures sampled in these two simulations but also in their probability in these independently generated ensembles. This level of convergence gives us confidence that the differences we observe between the various solvent models are predominantly due to solvation effects and not poorly converged ensembles with large uncertainties in the resulting data.

We then employed pure GB REMD simulation using both models available in Amber (GB^{HCT} [57] and GB^{OBC} [58,59]) as well as the hybrid REMD approach using the same GB models. We also performed REMD where only the first 1 or 2 solvation shells were retained for the exchange calculations (without a continuum model). Comparison of these results to each other and to the standard explicit solvent REMD results provides insight into the performance of the GB models, the improvement obtained by retaining the first solvation shell in the calculation of exchange probability (the hybrid model), and the need for the reaction field surrounding the solvation shells.

We compared ensemble distributions of properties such as chain end-to-end distance, backbone ϕ/ψ free energy maps, and cluster populations among the methods. While all of the solvation models provided similar results for alanine dipeptide, the GB models failed to reproduce the TIP3P ensemble data for Ala_3 and Ala_{10} even at a qualitative level, providing ensembles that were dominated by α-helical conformations. Simulations using hybrid REMD using GB^{OBC} and only a single shell of explicit water were in good accord with the reference simulations, with a high degree of similarity between structure populations (R^2=0.93), with lack of significant α-helix, and a strong preference for P_{II} conformation. This agreement was obtained despite a significant reduction in computational cost; for Ala_{10}, 40 replicas were used for standard REMD in TIP3P, while only 8 were needed for pure GB or hybrid GB/TIP3P REMD.

Methods

Replica Exchange Molecular Dynamics (REMD). We briefly summarize the key aspects of REMD as they relate to the present study. In standard Parallel Tempering or Replica Exchange Molecular Dynamics,[3,6] the simulated system consists of M noninteracting copies (replicas) at M different temperatures. The positions, momenta, and temperature for each replica are denoted by $\{q^{[i]}, p^{[i]}, T_m\}$, $i =$

1,...,*M*; *m* = 1,..., *M*. The equilibrium probability for this generalized ensemble is

$$W(p^{[i]},q^{[i]},T_m) = \exp\left\{-\sum_{i=1}^{M}\frac{1}{k_B T_m}H(p^{[i]},q^{[i]})\right\} \quad (1)$$

where the Hamiltonian $H(p^{[i]},q^{[i]})$ is the sum of kinetic energy $K(p^{[i]})$ and potential energy $E(q^{[i]})$. For convenience we denote $\{p^{[i]},q^{[i]}\}$ at temperature T_m by $x_m^{[i]}$ and further define $X = \{x_1^{[i(1)]}, ..., x_M^{[i(M)]}\}$ as one state of the generalized ensemble. We now consider exchanging a pair of replicas. Suppose we exchange replicas *i* and *j*, which are at temperatures T_m and T_n, respectively,

$$X = \{ ...;x_m^{[i]};...;x_n^{[j]};...\} \rightarrow X' = \{ ...;x_m^{[j]};...;x_n^{[i]};...\} \quad (2)$$

To maintain a detailed balance of the generalized system, microscopic reversibility has to be satisfied, thus giving

$$W(X)\rho(X\rightarrow X') = W(X')\rho(X'\rightarrow X) \quad (3)$$

where $\rho(X\rightarrow X')$ is the exchange probability between two states *X* and *X*′. With the canonical ensemble, the potential energy *E* rather than total Hamiltonian *H* will be used simply because the momentum can be integrated out. Inserting eq 1 into eq 3, the following equation for the Metropolis exchange probability is obtained:

$$\rho = \min\left(1, \exp\left\{\left(\frac{1}{k_B T_m} - \frac{1}{k_B T_n}\right)(E(q^{[i]}) - E(q^{[j]}))\right\}\right) \quad (4)$$

In practice, several replicas at different temperatures are simulated simultaneously and independently for a chosen number of MD steps. Exchange between a pair of replicas is then attempted with a probability of success calculated from eq 4. If the exchange is accepted, the bath temperatures of these replicas will be swapped, and the velocities will be scaled accordingly. Otherwise, if the exchange is rejected, each replica will continue on its current trajectory with the same thermostat temperature.

As we described above, one of the major limitations of REM is that the number of replicas needed to span a temperature range grows proportionally to the square root of number of degrees of freedom in the simulated system. While a more rigorous analysis of the acceptance probability in REM trials has been given recently using a Gaussian energy distribution model,[20,60] one can also approximate from eq 4 that the overall exchange probability P_{acc} is proportional to $\exp(-\Delta T^2/T^2)$, which implies that a greater acceptance ratio requires a smaller temperature gap ΔT or a more dense temperature distribution to reach. On the other hand, ΔT should be as large as possible so as to span a wide temperature range with a small number of replicas. The relationship can be estimated through consideration of potential energy fluctuations of two replicas sampling at the target temperature T_n and T_{n-1} (Figure 1). The instantaneous energy fluctuation δE in a given simulation at temperature *T* scales as $\sqrt{f}T$, and the average energy gap ΔE between two neighboring replicas is proportional to $f\Delta T$, where *f* is the number of degrees of freedom and $\Delta T = T_n - T_{n-1}$. Obtaining a reasonable acceptance ratio relies on keeping the replica energy gap comparable to the energy fluctuations, thus $\Delta E/\delta E$ should be near unity. Since $\Delta E/\delta E$ is proportional to $\Delta T\sqrt{f}/T$, the acceptable temperature gap between neighboring replicas therefore decreases with larger systems as $\Delta T \sim 1/\sqrt{f}$, and more simultaneous simulations are needed to cover the desired temperature range.

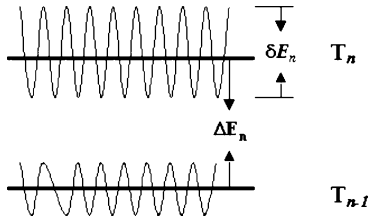

Figure 1. Schematic diagram illustrating the energy fluctuations for simulations at two temperatures for neighboring replicas. To obtain high exchange probabilities, the energy fluctuations δE in each simulation should be of comparable magnitude to the mean energy difference ΔE.

Model Systems and Simulation Details. We simulated three polyalanine sequences: alanine dipeptide (Ala_1), alanine tetrapeptide (Ala_3), and polyalanine (Ala_{10}), all with acetylated and amidated N- and C-termini, respectively. All simulations employed the Amber ff99 force field,[61,62] with modifications[63] to reduce α-helical bias. Explicit solvent and hybrid REMD used the TIP3P water model.[64] The standard REMD simulations in explicit solvent and in pure GB were run using our REMD implementation as distributed in Amber (version 8).[65] The hybrid solvent REMD calculations were performed with a locally modified version of Amber 8. All bonds involving hydrogen were constrained in length using SHAKE.[66] The time step was 2 fs. Temperatures were maintained using weak coupling[67] to a bath with a time constant of 0.5 ps^{-1}.

Secondary structure basin populations for central residues were calculated based on ϕ/ψ dihedral angle pairs. The dihedral angle ranges defining for those regions are provided in Table S1. The solvent accessible surface areas (SASA) for simulated peptides were calculated using the gbsa = 2 option in AMBER. The end-to-end distances for Ala_{10} were calculated between Cα atoms of Ala2 and Ala9 (omitting terminal residues) using the ptraj module of Amber. Cluster analysis for Ala_{10} was performed using moil-view,[68] using backbone RMSD for Ala2−9 and a similarity cutoff of 2.5 Å.

Explicit Solvent REMD. The Ala_{10} peptide in α-helical conformation was solvated in a truncated octahedral box using 983 TIP3P water molecules for a total of 3058 atoms. The system was equilibrated at 300 K for 50 ps with harmonic positional restraints on solute atoms, followed by minimizations with gradually reduced solute positional restraints and three 5 ps MD simulations with gradually reduced restraints at 300 K. Long-range electrostatic interactions were calculated using PME.[69] Simulations were run in the NVT ensemble.

Forty replicas were used at temperatures ranging from 267 K to 571 K, which were optimized to give a uniform exchange acceptance ratio of ∼30%. Exchange between neighboring temperatures was attempted every 1 ps, and each

REMD simulation was run for 50 000 exchange attempts (50 ns). The first 5 ns of each simulation was discarded to remove the initial structure bias.

To provide a stringent test of data convergence for greater conformational diversity expected for Ala_{10}, two sets of REMD simulations were performed, starting from different initial conformations. In one set, all replicas were started from a fully α-helical conformation; in the other an extended conformation was employed. In the case of Ala_1 and Ala_3, lower bounds for uncertainty were estimated by separating the full simulation data into halves and reporting the difference between values calculated for each half.

A similar procedure was used for Ala_1 and Ala_3. Ala_1 was solvated in a truncated octahedral box using 341 TIP3P water molecules. Ala_3 required 595 water molecules. For both systems the same equilibration procedure as used for Ala_{10} was employed. To cover the same temperature range 20 replicas for Ala_1 and 26 replicas for Ala_3 were needed. Both systems were simulated for ∼40 000 exchanges, and the first 5000 exchange attempts were discarded as equilibration.

Implicit Solvent REMD. Solvent effects were calculated through the use of two Generalized Born implementations in Amber (GB^{HCT} and GB^{OBC} (note that GB^{OBC} is model 2 in ref 59)). Two sets of intrinsic Born radii were used, both adopted from Bondi[70] with modification of hydrogen.[71] Unless otherwise noted, the GB^{HCT} model was used with the mbondi radii, and the GB^{OBC} model was employed with mbondi2 radii (as recommended in Amber). Scaling factors were taken from the TINKER modeling package.[72] No cutoff on nonbonded interactions was used. All other simulation parameters were the same as used in explicit solvent.

For Ala_{10}, the use of the continuum solvent model resulted in a total of 109 atoms considered explicitly in the simulations (∼28 times fewer than in the explicitly solvated system). The much smaller system size permitted the use of 8 replicas to cover the same temperature range that required 40 replicas in the explicit solvent, while obtaining the same 30% exchange acceptance probability. Exchanges were attempted every 1 ps, and the REMD simulation was run for 50 000 exchange attempts (50 ns). Simulations were initiated with the same two initial conformation ensembles as were used for the explicit solvent REMD calculations, with comparison of the two runs providing a lower bound for the uncertainty in resulting data. For Ala_1 and Ala_3 the same approach was used, with 4 replicas used to cover the temperature space for each system. Simulations were run for 50 000 exchange attempts, and the first 5000 exchanges were discarded.

Hybrid Solvent REMD. All simulation parameters in the hybrid solvent REMD simulations were the same as those employed for standard REMD in the explicit solvent, with the exception that the number of replicas (8 for Ala_1, Ala_3, and Ala_{10}) and the target temperatures were the same as those used for the pure GB REMD simulations for Ala_{10}. It is important to note that the hybrid solvent model was used *only* for calculation of exchange probability; the simulations themselves were performed on fully solvated systems with truncated octahedral periodic boundary conditions and PME for the calculation of long-range electrostatic interactions.

We determined the number of water molecules to retain in the hybrid model based on analysis of the number of waters in the first solvation shell of Ala_{10} in the ensemble of structures sampled in the standard REMD explicit solvent simulations. We found that 100 water molecules were sufficient even for the most extended conformations (data not shown). Thus this number was used for all replicas and all exchanges. For Ala_1, 30 water molecules were enough to incorporate the first solvation shell and 60 water molecules for the first and second solvation shells. These numbers increase to 50 waters and 100 waters for the first solvation shell and the first and second solvation shells of Ala_3, respectively. Ala_1 and Ala_3 hybrid simulations were run for ∼30 000 exchanges, and the first 5000 were discarded.

At each exchange step, the distance between the oxygen atom of each water molecule and all solute atoms was calculated. Water molecules were then sorted by their closest solute distance, and all water molecules except the X with the shortest solvent−solute distances were temporarily discarded (where X is the number of waters retained in each system, as described above). The energy of this smaller system was then recalculated using only these close waters and the GB solvent model. This energy was used to calculate the exchange probability, and then all waters were restored to their original positions and the simulations were continued (Figure 2). In this manner the simulations using the hybrid solvent model were continuous simulations with fully solvated PBC/PME, and the hybrid model was used only for the calculation of exchange probabilities.

Results and Discussion

Comparison of Exchange Efficiency for Hybrid and Standard REMD in Ala_{10}. Even though REMD has become a useful tool to improve conformational sampling, REMD simulations are highly computationally expensive, particularly when the solvent is treated explicitly. The increase in cost arises not only from the additional effort involved in calculating forces in a given simulation but also from the increase in the number of simulations (replicas) needed to span a particular temperature range. This increase is due to the much larger number of degrees of freedom present in the explicitly solvated system as compared to that in continuum solvent models. In the case of Ala_{10}, our largest model system, the number of replicas needed to span the range of 267 K to 571 K increases from 8 to 40 when switching from implicit to explicit solvation.

We evaluated the utility of the hybrid solvent model during the calculation of the exchange probability on several levels, using Ala_{10} as its size is most relevant to the larger systems that would benefit most from this method. First, we validated that fewer replicas were needed to obtain efficient exchange with the hybrid model as compared to the number required when retaining the full periodic box of explicit water molecules during the exchange probability calculation (eq 4). Efficient exchanges were obtained with the hybrid model even when using the same number of replicas as was needed for the pure continuum solvent REMD simulations. Next, we evaluated whether the use of the hybrid model affected the data obtained from the simulations, with particular

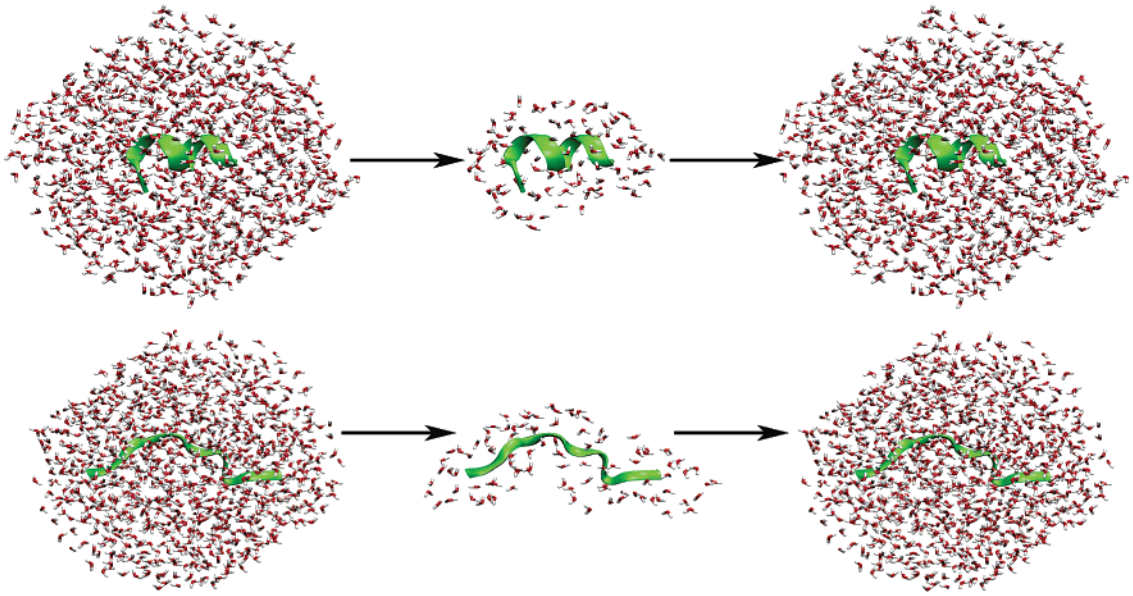

Figure 2. Schematic description of hybrid solvent REMD. The fully solvated Ala_{10} (with truncated octahedral boundary conditions) is simulated between exchanges (left). The exchange energy is calculated by retaining only the closest 100 waters (center), with bulk solvent properties calculated using the GB solvation model. After the exchange calculation the explicit solvent is restored, and the dynamics continues under periodic boundary conditions. This approach allows on the fly calculation of the solvation shell, whose shape adjusts automatically to the solute conformation (top: α-helical structure, bottom: extended structure). As a result, many fewer replica simulations are required.

emphasis on the conformational distributions sampled by the model peptides. These distributions were also compared to those obtained for REMD with only the continuum solvent model.

An important benefit of REMD is the ability to obtain improved sampling at low temperatures of interest by exchanging conformations with higher temperature simulations that have less likelihood to become kinetically trapped. As described in Methods, the probability of the successful exchange of conformations between two temperatures depends on the overlap in potential energy distributions at those temperatures. Figure 3 shows the potential energy distributions for each temperature for sets of simulations with explicit solvent (A) and those with GB (B) between 267 K and 571 K. The graph illustrates why fewer replicas are required for the GB model; the energy range spanned is smaller for the smaller system, and fewer replicas are still able to achieve the required overlap. In contrast, when the explicit solvent model is used with only the 8 replica temperatures that are successful with GB, no significant overlap in the distributions is observed (Figure 3C).

Based on Figure 3, exchanges between replicas at neighboring temperatures are expected to occur with a high probability when using 40 replicas in explicit solvent or 8 replicas with GB. No exchanges are expected for the explicit solvent with only 8 replicas. Figure 4 shows the temperature histories of the first 2 replicas in the same explicit solvent and GB REMD simulations as were shown in Figure 3. As expected, the replicas visited all available temperatures during the run (the other replicas showed similar behavior and are not shown for clarity). However, the explicit solvent REMD with only 8 replicas showed *no* exchanges even after 25 000 attempts (25 ns simulation), and all replicas remained at their initial temperatures. This REMD simulation is identical to

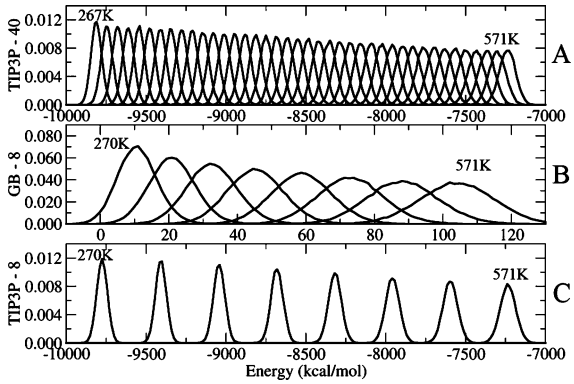

Figure 3. Potential energy distributions for Ala_{10} simulations over a range of temperatures using (A) explicit solvent REMD with 40 replicas, (B) GB REMD with 8 replicas, and (C) explicit solvent REMD with 8 replicas using the same temperature distribution as GB REMD. GB simulations involve fewer degrees of freedom and are able to span the energy range with fewer replicas. In contrast, no overlap is obtained when using explicit solvent with the same replica and temperature selection as GB. This implies that no exchanges would be permitted, and the benefits of REMD would be lost.

8 standard MD simulations at different temperatures, and therefore no sampling improvement is obtained. Thus, in order for replicas to sample a range of temperatures, more replicas (and significantly more computational resources) are required for simulations in the explicit solvent. Reducing this requirement while maintaining fully explicitly solvated simulations is the goal of our hybrid model.

These exchange efficiencies are all consistent with previously reported REMD simulations and the known scaling with system size of the number of replicas required for efficient exchange. In our case these data provide an

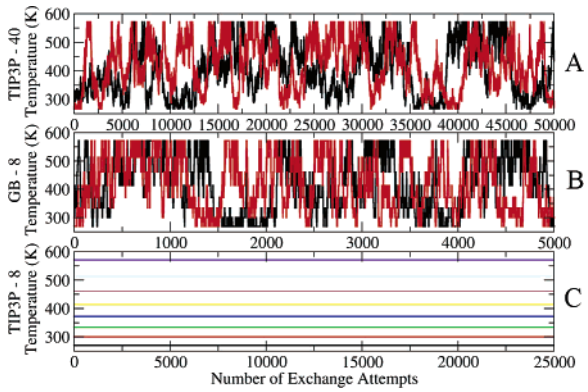

Figure 4. Temperature histories for Ala_{10} replicas using (A) explicit solvent with 40 replicas, (B) GB with 8 replicas, and (C) explicit solvent with 8 replicas. For clarity only the first two replicas for A and B and only the first 5000 exchanges of B are shown. Consistent with the potential energy distributions shown in Figure 3, exchanges are only obtained when sufficient overlap in potential energy distributions is present. If too few replicas are used (C), the result is a series of standard MD simulations.

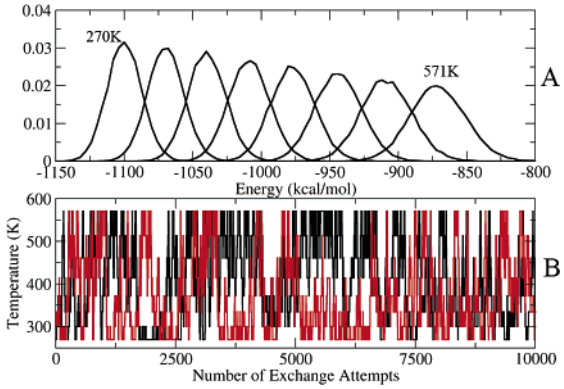

Figure 5. Potential energy distributions (A) and temperature histories of 2 Ala_{10} replicas (B) using 8 replicas in periodic boxes with fully explicit solvent but with the hybrid solvent model for calculation of exchange probability. Use of the hybrid model gives overlap between neighboring temperatures and allows replicas to span a range of temperatures, in sharp contrast to the total lack of exchanges for the *same* simulated system with standard REMD (Figures 3C and 4C). For clarity only the first 10 000 exchanges are plotted, and only 2 replicas are shown in the lower figure.

important context for evaluation of the use of hybrid solvation models during the calculation of exchange probability. We performed REMD simulations using the same explicitly solvated system as shown above, but with only the 8 replicas/temperatures that gave an efficient exchange with pure GB solvation. With standard REMD, this system showed no overlap in potential energy distributions and was unable to generate any successful exchanges (Figure 4C). We employed the hybrid solvent model only for calculation of the exchange probability (eq 4) for this fully explicit solvent system. The distributions of the potential energies for the different temperatures during 10 000 exchange attempts (10 ns) are shown in Figure 5. Use of the hybrid solvent model permits the simulations to achieve nearly the same level of energy distribution overlap as we obtained for the pure GB model. Consistent with this observation, multiple exchanges are observed despite the relatively small number of replicas employed. The replicas are able to traverse the entire temperature range on the nanosecond time scale. It is interesting to note that this is more rapid than seen for the standard REMD explicit solvent run, most likely due to the larger temperature step taken with each successful exchange with the hybrid solvent model (due to larger ΔT between neighboring replicas). The standard REMD run requires more exchanges to traverse the same total temperature range. This suggests that the hybrid calculation may have additional advantages beyond simply reducing the number of replicas as compared to the standard REMD; however, such an analysis is outside the scope of the present article.

Analysis of Conformational Sampling in Hybrid and Standard REMD. After establishing the ability of the hybrid REMD model to reduce the number of replicas required to obtain efficient exchanges, we examined the ability of the hybrid approach to reproduce ensemble data obtained with standard REMD in the explicit solvent. We also investigated whether the reaction field beyond the solvation shells is required, and the dependence of the results on the number of solvation shells included in the exchange calculation. For the larger Ala_{10}, the computational demands of obtaining high-precision data for various hybrid models (which require fully solvated simulations) prevented exhaustive testing. Thus, these more detailed tests were performed on the smaller models alanine dipeptide (blocked Ala_1) and alanine tetrapeptide (blocked Ala_3).

Alanine Dipeptide. We first compared results obtained for the standard REMD with TIP3P to those from 2 different GB models as well as to TIP3P but using the hybrid solvent model for calculation of exchange probability. The hybrid model employed either a first solvent shell (30 TIP3P waters) or first and second shells (60 waters). The population of minima corresponding to alternate secondary structure types (see Methods for details) are shown in Table 1. The largest population is found for the polyproline II basin (~35%), followed by an α-helix and a β-sheet (each ~25%), and a much lower population of a left-handed α-helix or turn conformation (1−3%). We make the observation that all of these solvent models provide essentially the same results. Use of either GB^{OBC} or GB^{HCT} with no explicit solvent either in MD or in the exchange calculation provides populations for each of the basins with an error of ~2% population as compared to the standard REMD in the explicit solvent. Similarly, the average SASA is nearly identical for all models. These data indicate that the hybrid model is at least performing adequately and does not have any obvious and serious problems and that similar results are obtained for either the first and second solvation shells or only the first shell. This insensitivity is expected since the GB simulations adequately reproduced the explicit solvent data with no explicit solvent shell. The insensitivity of the results to solvent model strongly indicates that alanine dipeptide is not a good test case for evaluation of the effects of inclusion of explicit solvent.

Alanine Tetrapeptide. We next turn to results from alanine tetrapeptide to evaluate whether the agreement

Table 1. Populations of Basins on the Alanine Dipeptide ϕ/ψ Energy Landscape Corresponding to Alternate Secondary Structures, along with Average Solvent Accessible Surface Areas[a]

alanine dipeptide	α	β	P^{II}	α^L	SASA
explicit solvent	28.1 ± 1.0	25.1 ± 0.1	36.2 ± 0.5	2.6 ± 0.1	355.8 ± 0.0
GB^{OBC}	29.3 ± 0.8	26.5 ± 0.5	35.1 ± 0.2	0.7 ± 0.1	356.5 ± 0.0
GB^{HCT}	28.5 ± 0.2	27.6 ± 0.1	34.0 ± 0.2	0.8 ± 0.2	356.5 ± 0.1
hybrid first shell + GB^{OBC}	29.7 ± 1.8	24.7 ± 0.4	35.0 ± 1.5	2.5 ± 0.1	355.8 ± 0.1
hybrid first and second shells + GB^{OBC}	30.3 ± 1.5	24.7 ± 0.3	36.0 ± 0.2	1.3 ± 0.8	355.9 ± 0.1

[a] The results for the pure GB and hybrid REMD models are all similar to those obtained using standard REMD with full explicit solvent.

Table 2. Data for the Central Alanine in Alanine Tetrapeptide (Blocked Ala_3)[a]

alanine tetrapeptide	α	β	P^{II}	α^L	SASA
explicit solvent	23.6 ± 0.1	23.4 ± 1.3	40.2 ± 1.4	5.1 ± 0.1	565.3 ± 0.1
GB^{OBC}	50.5 ± 2.4	17.5 ± 0.9	22.9 ± 0.6	1.1 ± 0.4	557.4 ± 1.0
GB^{HCT}	57.8 ± 1.0	15.2 ± 0.2	18.2 ± 0.4	1.2 ± 0.1	552.4 ± 0.4
hybrid first shell noGB	41.4 ± 0.8	13.5 ± 0.9	23.4 ± 1.0	13.1 ± 0.8	552.7 ± 0.1
hybrid first and second shells noGB	29.5 ± 0.2	14.1 ± 0.2	24.1 ± 0.5	23.4 ± 0.3	550.8 ± 0.2
hybrid first shell GB^{OBC}	21.6 ± 0.9	21.2 ± 0.2	41.1 ± 0.3	7.6 ± 1.0	563.2 ± 0.1
hybrid first and second shells GB^{OBC}	28.3 ± 1.7	22.2 ± 0.9	37.7 ± 0.2	3.8 ± 0.1	563.8 ± 0.2
hybrid first shell + GB^{HCT}	23.5 ± 1.1	22.1 ± 0.8	42.8 ± 1.0	2.3 ± 0.0	566.4 ± 0.2
hybrid first and second shells + GB^{HCT}	14.9 ± 0.2	25.6 ± 0.1	49.4 ± 0.4	1.9 ± 0.4	569.6 ± 0.1

[a] Populations of basins on the ϕ/ψ energy landscape corresponding to alternate secondary structures are shown, along with average solvent accessible surface areas. Data are discussed in the text.

between all solvent models tested for alanine dipeptide is maintained in larger systems. In Table 2 we show populations for secondary structure basins for the central alanine residue using standard REMD with explicit solvent, GB^{OBC} or GB^{HCT}. Data are also shown for several hybrid models, as discussed below.

For standard REMD in explicit solvent, we observe that the populations have not changed significantly from those obtained for alanine dipeptide, with a slight increase in population of the polyproline II conformation that dominates the ensemble. In this case, however, we observe that both of the pure GB models are in significant disagreement with TIP3P, with α-helical conformations dominating the ensemble (over 50% for each GB model). The two GB models are similar to each other. Overstabilization of salt bridges in GB has been reported,[12,26,27] but no salt bridges are present in this system.

Next, we performed REMD simulations in explicit solvent, but retain only the first (50) or the first and second (100) solvation shells in the exchange calculation. Importantly, no GB model was included in these simulations. Using only a single solvation shell results in a significant bias in favor of α-helical conformations (41% vs ∼24% for standard REMD), much too little polyproline II conformation and nearly three times the α^L/turn conformation than was sampled in standard REMD. Inclusion of a second shell (without GB) resulted in an even greater shift of the ensemble toward turn structures. Notably, both of these shell models show a significantly smaller average SASA than obtained with standard REMD in the explicit solvent, consistent with a drive toward compact conformations that reduce the water/vacuum interface that is present without a reaction field to surround the solvent shells.

We next examine the data obtained from the hybrid model in which GB solvation was employed in addition to shells of explicit solvation. We note that all of these models are in significantly better agreement with the standard TIP3P REMD data, regardless of the GB method or number of shells. The more recent GB^{OBC} model performed best, with errors in population of only ∼3% for all basins with the exception of the α-helix conformation with the first and second shell model, which had an error that was less than 5%. The average SASA was also in excellent agreement with standard REMD. We conclude that this hybrid model is significantly better than the pure GB REMD or inclusion of only the solvation shells with no reaction field. The addition of a second shell in the exchange calculation appears to make no significant difference as compared to a single shell.

As described above, the MD simulations between exchanges in the hybrid model are performed with full explicit solvation. We thus do not need to restrain the explicit water, and since the solvation shells are surrounded by bulk explicit solvent, we expect no effect on the water geometries as have been reported when using a hybrid GB+explicit water model for dynamics.[41] To test this hypothesis, we calculated the radial distribution function for water oxygens around the carbonyl oxygen in the central Ala2 and found that the function obtained in the hybrid model was indistinguishable from that in the standard REMD in the explicit solvent (Figure S1). Since these data are obtained from the entire set of structures, this close agreement is also a further indicator of the similarity of the ensembles obtained using hybrid or standard REMD.

The hybrid model using GB^{HCT} performed comparably to GB^{OBC} when only a single shell was used, but the first+second shell model showed a marked reduction in the α-helix conformation (from 23.5% to 14.9%). This was accompanied by an increase in the average SASA. These effects with GB^{HCT} are even more apparent in Ala_{10} and will be discussed in more detail below.

Polyalanine (Ala_{10}). The conformational variability available to Ala_{10} is significantly greater than for alanine dipeptide

Table 3. Data for the Central Ala5 in Blocked Ala_{10}[a]

Ala_{10}	α	β	P^{II}	α^L	SASA
explicit solvent	24.9 ± 0.8	19.5 ± 0.6	39.5 ± 0.4	8.4 ± 2.0	1195.4 ± 5.6
GB^{OBC}	67.8 ± 1.8	8.3 ± 0.7	12.5 ± 0.8	4.2 ± 0.1	1098.6 ± 0.4
GB^{HCT}	83.1 ± 0.1	3.2 ± 0.1	5.0 ± 0.0	2.3 ± 0.1	1038.3 ± 1.6
hybrid GB^{OBC} +first shell	35.7 ± 6.2	17.3 ± 0.2	29.0 ± 5.3	6.6 ± 0.7	1140.8 ± 4.4
hybrid GB^{HCT} + first shell	12.3 ± 0.2	28.3 ± 0.3	50.5 ± 1.2	2.1 ± 1.1	1275.4 ± 2.5
hybrid $GB^{OBC'}$ + first shell	29.8 ± 1.6	18.5 ± 1.6	34.3 ± 0.5	8.9 ± 0.3	1167.8 ± 2.5

[a] Populations of basins on the ϕ/ψ energy landscape corresponding to alternate secondary structures are shown, along with average solvent accessible surface areas. $GB^{OBC'}$ refers to the hybrid model using GB^{OBC} with slight adjustment of the Born radius on H bonded to O. Uncertainties reflect differences between independent simulations from different initial structures. Data are discussed in the text.

or tetrapeptide. We thus performed a more stringent evaluation of data convergence in this case to ensure that the differences we observe between the different solvent models are statistically significant. We performed two completely independent REMD simulations for each of the solvent models, in each case starting from 2 different initial ensembles (fully extended or fully helical). This allows us to evaluate the influence of the solvent model within the context of intrinsic uncertainties in each data set.

We also consider separately the local ϕ/ψ conformations and more global properties of this larger peptide, such as end-to-end distance distributions and conformation cluster analysis.

Comparison of Local Conformational Preferences. In Table 3 we show secondary structure basin populations for the central Ala5 residue. Free energy surfaces for these simulations are provided in Figure S2. For the reference standard REMD simulations in explicit solvent, the polyproline II conformation is again favored with the same ∼40% population as we obtained for alanine dipeptide and tetrapeptide. In comparison, both GB models show a very large bias in favor of α-helix conformations (∼70−80%).

Consistent with the results obtained for alanine tetrapeptide, the GB^{HCT} hybrid model favors extended conformations with large SASA too strongly (β and P_{II}), despite the bias in favor of an α-helix for the pure GB^{HCT} simulations. This suggests that the explicit water shell is solvated too strongly by this GB model. The GB^{OBC} hybrid model shows a more balanced profile in good agreement with the full TIP3P data. The strong bias favoring an α-helix in the pure GB^{OBC} model is nearly completely eliminated when a single solvent shell is retained, although some remains with approximately 10% too much α-helix present in the GB^{OBC} hybrid.

In addition to differences in the method for calculating GB effective Born radii, the GB^{HCT} and GB^{OBC} simulations employed different intrinsic Born radii (denoted in Amber as mbondi and mbondi2 sets, respectively), consistent with recommendations for these models. To determine the relative influence of these two differences, we repeated the calculations, swapping the GB models and radii (GB^{HCT} with mbondi2, GB^{OBC} with mbondi). We found that the results depended nearly exclusively on the set of radii and were less sensitive to the GB models themselves (data not shown). This is consistent with the aim of the GB^{OBC} model, which was designed to provide improved properties for larger systems than our current model.[58] We note that the strong bias toward extended structures seen in the hybrid models using mbondi radii likely arises from the use of 0.8 Å for hydrogen atoms bonded to oxygen. In the more recent mbondi2 set, this value was restored to the default Bondi value of 1.2 Å. This larger value appears to have an improved balance of hydrogen bonding of the explicit solvent to the solute or to the bulk (continuum) solvent.

Comparison of Global Structural Properties. Our analysis of alanine dipeptide and tetrapeptide focused on local backbone conformation; in the larger Ala_{10} we supplement this analysis with more global properties of the chain. We calculated the end-to-end distance distributions for Ala_{10} in the 300 K ensembles obtained from each of the different REMD simulations. In Figure 6 we show the results of the 2 explicit solvent REMD simulations that were initiated from fully α-helical or extended conformations, respectively. A broad distribution of distances is observed, suggesting that no particular conformation is preferred, consistent with the local backbone preferences for the central Ala5. Consistent with the small uncertainties in the ϕ/ψ basin populations, we observe that the initial conformation has essentially no effect on the distribution, indicating that the REMD simulations are well-converged on this time scale. Similar behavior is observed for other temperatures. As expected, standard MD simulations at 300 K were trapped near the initial conformation on this time scale (data not shown).

In Figure 6, we show the distance distributions at 300 K obtained from GB REMD using the two GB models (HCT and OBC). In contrast to the relatively flat profiles seen in the explicit solvent REMD data, a sharp peak near 11 Å is obtained using either GB model, with essentially no sampling of extended conformations with end-to-end distances greater than ∼15−20 Å, unlike the explicit solvent REMD that shows a nearly flat distribution out to ∼22 Å. This is consistent with the strong bias toward α-helix in the pure GB models as shown in Table 3. The bias is somewhat less pronounced with the GB^{OBC} model than with GB^{HCT}. We note that these differences between the various solvent models are much larger than the differences obtained from alternate initial conformations using the same solvent model.

In Figure 6 we also show end-to-end distance distributions at 300 K obtained from REMD with the same hybrid variations shown in Table 3, each of which retained only the first shell (100 closest) water molecules combined with different GB models in the exchange calculation. When GB^{HCT} was used in the hybrid model (Figure 6C), the distributions differ significantly from the reference explicit solvent REMD data, consistent with the large increase in polyproline II backbone conformations and average SASA

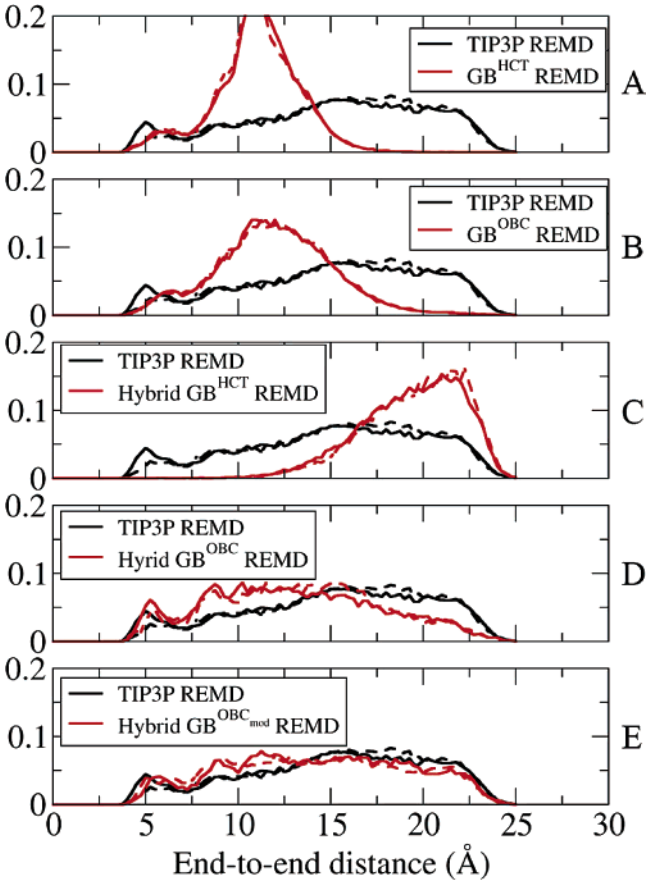

Figure 6. Ala_{10} end-to-end distance distributions at 300 K obtained in REMD using alternate solvent models (red): (A) pure GB^{HCT}, (B) pure GB^{OBC}, (C) hybrid REMD with GB^{HCT} and mbondi radii, (D) hybrid REMD with GB^{OBC} and mbondi2 radii ($H^O = 1.2$ Å), and (E) hybrid REMD with $GB^{OBC'}$ (mbondi2 radii with $H^O = 1.15$ Å). In each case the results are independent of initial conformation (solid/dashed lines). Data from standard REMD with explicit solvent are shown in each graph for comparison (black).

for this model shown in Table 3. This bias toward more extended conformations in the hybrid using GB^{HCT} is also consistent with what we observed for alanine tetrapeptide (Table 2).

We next analyzed the distributions obtained from the GB^{OBC} hybrid model (Figure 6D). In this case, much better agreement with the reference data is seen than with either GB^{OBC} alone or the explicit/GB^{HCT} hybrid. However, the sampling of the most extended conformations (longest end-to-end distances) is slightly reduced in the hybrid REMD simulations.

The good convergence of our data suggested the possibility of using it for minor empirical adjustment of the mbondi2 values for use with the GB^{OBC} hybrid model. We adjusted the radii of hydrogen bonded to either N or O by 0.05 Å. Modification of H on N had little effect on the resulting distributions (data not shown), but reduction of the radius of H on O from 1.2 to 1.15 Å ($GB^{OBC'}$) resulted in an end-to-end distance distribution in improved agreement with standard explicit solvent REMD data (Figure 6E and Table 3). This slight reduction in the hydrogen radius is consistent with the increased electronegativity of oxygen.[71] This change

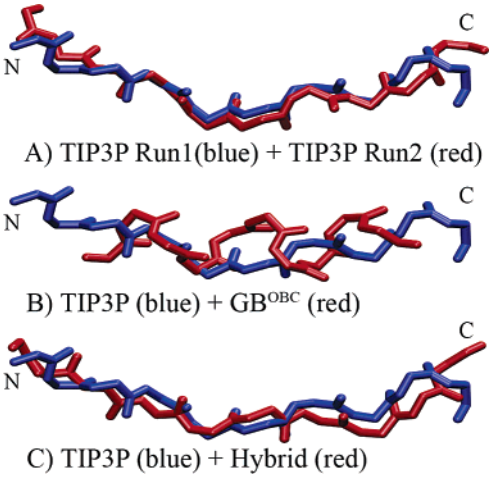

Figure 7. Representative structures for the most populated clusters in 300 K ensembles obtained using various solvent models. (A) Very similar P_{II} structures are obtained from 2 independent standard REMD simulations with explicit solvent, initiated in extended and fully helical conformations. (B) Comparison of structures from GB^{OBC} and TIP3P. GB^{OBC} prefers α-helical conformations, in disagreement with explicit solvent simulations. (C) Using $GB^{OBC'}$ with the hybrid model provides structures in close agreement with standard REMD in TIP3P. Terminal residues were not included in the cluster analysis.

does not affect the pure GB calculations since Ala_{10} has no H bonded to O.

The $GB^{OBC'}$ hybrid model showed improved agreement with the pure TIP3P data, with all basin populations within 5% of the standard explicit solvent REMD. Some slight bias favoring an α-helix at the expense of some polyproline II conformation remains in this model and will be the subject of future investigation. We repeated the simulations of alanine dipeptide and tetrapeptide using this modified radius and found that the populations (Table S2) remained in good agreement with standard REMD with explicit solvent.

Since the backbone conformation populations suggest that the P_{II} basin is the global free energy minimum in both the standard explicit solvent and the hybrid solvent models (Table 3 and Figure S2), we performed cluster analysis to determine the extent to which this local preference was reflected in the conformation of the entire polymer chain. Once again we compare results from independent ensembles generated by REMD with different initial conformations to ensure the convergence of our data.

The most populated cluster for Ala_{10} at 300 K in both standard explicit solvent REMD runs was an extended P_{II} conformation (over 98% of the local backbone conformations in this cluster are P_{II}, data not shown). This fully P_{II} cluster comprised ~20% of the overall ensemble in both explicit solvent simulations (19.5% vs 21.2%). Representative structures for the clusters obtained from the independent simulations differed by only 1.3 Å in backbone RMSD (Figure 7A). Once again, the high level of consistency between the data sets and independence of not only the conformation but also the absolute population of the clusters give us confidence in the converged nature of our data. The relatively low population of this cluster in both simulations is also consistent

with the broad distribution of end-to-end distances (Figure 6). A more detailed analysis of the ensemble of structures sampled by Ala_{10} will be presented elsewhere, but this preference for P_{II} conformations is consistent with the experimental and simulation reports described previously.

As was demonstrated with the analyses presented above, the pure GB^{HCT} and GB^{OBC} REMD simulations do not reproduce the data obtained in the explicit solvent, nor are they consistent with experimental data. The most populated cluster in both cases is fully α-helical (Figure 7B shows the GB^{OBC} structure), comprising ∼48% of the overall ensemble for GB^{HCT} and 25.4% for GB^{OBC}. This analysis is consistent with the α-helical bias apparent in the Ramachandran free energy surfaces shown in Figure S2.

We next performed cluster analysis on the ensembles obtained with the $GB^{OBC'}$ hybrid model with modified mbondi2 radii. Consistent with the standard explicit solvent REMD runs, the most populated cluster at 300 K was also an extended P_{II} conformation. Representative structures were within 1.5 Å backbone RMSD from those obtained in the explicit solvent (Figure 7C), again suggesting that the hybrid model is able to capture the dominant effects of the explicit solvent in the exchange calculation despite the need for many fewer replicas.

Since the most populated clusters were in close agreement between both TIP3P REMD simulations and the $GB^{OBC'}$ hybrid model, we compared the populations of all clusters observed. Smith et al. showed[73] that cluster analysis of simulations was a much more stringent test of convergence than other measures that they tested, including energy, RMSD, or diversity of hydrogen bonds sampled. This was particularly useful when analyzing coordinate sets obtained by merging two independent trajectories. They examined the 5 ns dynamics of an 11-residue peptide and showed that the two trajectories sampled essentially none of the same clusters.

We adapted this approach to our analysis, but we emphasize not only just the existence of conformation families in two data sets but also the fractional population of each cluster in 300 K ensembles sampled in independent simulations. All trajectories from TIP3P REMD, GB^{OBC} REMD, and hybrid $GB^{OBC'}$ simulations were combined, and the resulting data set was clustered. A total of 44 clusters contained 99% of the structures; the fraction of the ensemble corresponding to each cluster was calculated for each REMD simulation. We compared the population of each cluster in the different ensembles, including those generated with the same or different solvent models.

First we evaluated the convergence of our standard REMD simulations with TIP3P by comparing cluster sizes between the independent runs with different initial conformations (extended and fully α-helical). Not only were the same conformations sampled in each run (20.3±0.9%), but the populations of clusters in each ensemble were highly correlated (Figure 8A, R^2=0.974 and a slope of 1.02). This indicates that the relative population of each structure type is highly converged in these data sets.

In stark contrast, when the TIP3P and GB^{OBC} ensembles are compared, no correlation between cluster populations is observed (Figure 8B, R^2=0.075), and the largest cluster in each (∼20%) has less than 2% population in the other model. Much better results are obtained from the $GB^{OBC'}$ hybrid data, with a correlation coefficient of 0.935 with the standard TIP3P REMD data (Figure 8C). All clusters larger than 5% have the same rank order in the two models. There is a relatively small difference in the size of the single cluster that is the largest for both models (15.9±0.6% and 20.3±0.9% for hybrid and standard TIP3P REMD, respectively). This corresponds to an error of only 0.15 kcal/mol for the free energy of this cluster between the two models, compared to the 0.05 kcal/mol difference obtained between data sets from the same model. For comparison, the error in the free energy of this conformation using GB was more than 10 times larger (1.6 kcal/mol).

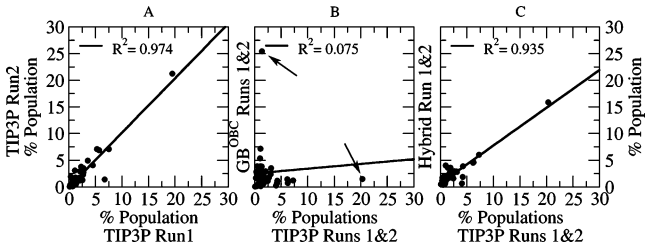

Figure 8. Cluster populations at 300 K from REMD for TIP3P Run1 vs Run2 (A), TIP3P Runs 1&2 vs GB^{OBC} Runs 1&2 (B), and TIP3P Runs 1&2 vs hybrid $GB^{OBC'}$ Runs 1&2. High correlations between individual TIP3P simulations and between TIP3P and hybrid simulations are observed, with the difference in the largest cluster in (C) corresponding to an error in free energy of only 0.15 kcal/mol. No correlation between TIP3P and GB^{OBC} is observed; note also in plot (B) that the largest cluster in each solvent model has very low population in the other model (indicated by arrows).

Since the standard explicit solvent REMD and hybrid solvent using $GB^{OBC'}$ have the same most populated cluster, we investigated the time scale required for each model to adopt this conformation as the dominant member of their ensemble. This is important since the standard REMD simulation employed many more replicas, possibly facilitating an earlier location of the P_{II} conformation that would then be adopted in the lowest temperature ensembles. In Figure 9 we show the fractional size of this cluster in the structures sampled as a function of time for the standard REMD and the hybrid REMD, including data from both initial conformations in each model. Data are shown at 300 K, and the first 5 ns were discarded in each case to remove biasing of the populations by the initial conformations that were not sampled at later points. The level of agreement is impressive; the long-time averages for both simulations of the 2 models are all ∼20%, with convergence to this value occurring at approximately 5 ns in all cases (in addition to the 5 ns that were discarded).

Conclusions

We introduced a new variant of replica exchange molecular dynamics in which simulations are performed with a fully explicit representation of the solvent, but those solvent molecules beyond the first solvation shell are replaced with a continuum description only for the purpose of calculating the exchange probability. This reduces the effective system

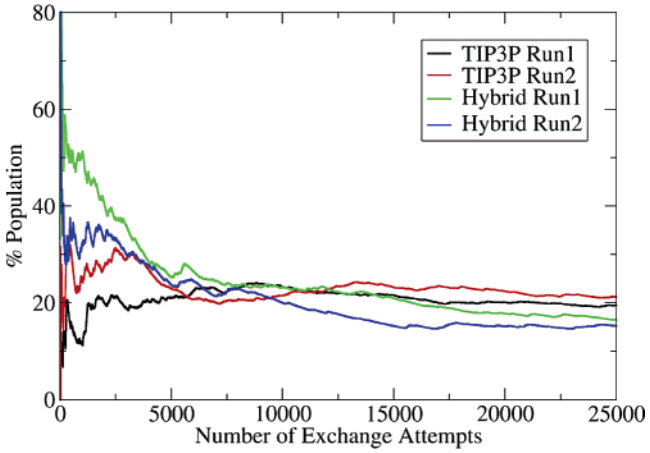

Figure 9. Population of the cluster corresponding to polyproline II helix (Figure 7) as a function of time for REMD simulations in explicit solvent, with the 2 independent simulations using the full system energy in the exchange calculation shown in black/red and the GB^{OBC} hybrid shown in green/blue. At ~5 ns, all four simulations converge to a population of 16−20% (the largest cluster in each of the ensembles), with a slightly lower population in the hybrid models that is consistent with Figure 8C.

size governing the number of replicas required to span a given temperature range and therefore significantly reduces the computational cost of REMD simulations. This approach is similar in spirit to hybrid explicit/continuum models that have been proposed for use during each step of MD simulation; in the present case, however, the solvent is fully explicit during the dynamics, and no restraints are needed to maintain a solvation shell. However, since the Hamiltonian used for the exchange differs from that employed during dynamics, these simulations are approximate and are not guaranteed to provide correct canonical ensembles. It is important to determine the extent to which this approximation affects the resulting ensembles; in this article we introduce the method and investigate some of these effects on several short alanine-based peptides.

Recently, another approach to reducing the number of replicas required for explicit solvent REMD simulations was proposed[74] in which the water−water interaction energy was temperature-dependent. That study employed alanine dipeptide as a model to show that their less computationally demanding method provided a similar ensemble to that obtained with the standard REMD. In the present work we show that alanine dipeptide conformations are nearly insensitive to the solvent models that we tested, with results from the full explicit solvent, two different GB models, and several hybrid models all providing similar ensembles. In contrast, several of these models provided ensembles for the longer peptides that were in significant disagreement with the standard REMD in the explicit solvent, indicating that larger model systems should be included in evaluation of solvent models.

We further tested the method by calculation of conformational ensembles of Ala_{10} using the TIP3P explicit solvent model, two GB models available in Amber, and hybrid variants using TIP3P and each GB model, all using the same underlying protein force field parameters. Ensembles from standard REMD in the explicit solvent were considered the standard, and convergence of this data set was validated by a high correlation (R^2=0.974) between the fractional populations of conformation families in simulations initiated with completely different initial structure ensembles. While a broad distribution of conformations was sampled, the predominant cluster for Ala_{10} adopted a P_{II} structure. This preference is consistent with reported experimental and computational results for short polyalanine peptides.[75]

Simulations using the hybrid model with GB^{OBC} were in excellent agreement with the reference data for local backbone conformations, end-to-end distance, SASA, and populations of each conformation family in the ensemble. The difference in population in the largest cluster indicates that the hybrid model introduced an error of less than 0.2 kcal/mol in free energy while reducing the computational expense by a factor of 5.

In contrast, REMD using only the GB models provided ensembles that bore no resemblance to the reference data, with the GB ensembles incorrectly dominated by α-helical conformations. This may be indicative of general errors in these GB models, or they may arise from neglect of the structure in the first solvation shells of the peptide. Mezei et al. recently reported[55] free energy calculations using explicit solvent, showing that solvation strongly favors the P_{II} conformation over an α-helix. Solvation free energy was shown to be highly correlated with the energy of interaction between the peptide and its first solvation shell.

It is important to note that several challenges remain for more general use of the proposed hybrid approach. In particular, the present work studied the effects on alanine-based peptides. Future studies should be performed on other sequences with a more diverse representation of functional groups in the side chains. In particular, it will be important to determine whether the hybrid model is able to overcome known issues with GB models and ions pair interactions. The inclusion of explicit counterions in the exchange calculation may also be problematic. Additionally, we demonstrated that inclusion of a single shell of explicit water was sufficient for alanine dipeptide and analine tetrapeptide. In both cases similar results were obtained using one or two shells, but we were unable to perform these comparisons for Ala_{10}. Although our approach reduces the number of replicas required for REMD, the simulations are still fully solvated during each step of MD and obtaining well converged data requires a significant investment of computational resources.

The results obtained from these model systems provide additional evidence that explicit representation of water in the first solvation shell can significantly improve the performance of the GB continuum models, providing data similar to standard REMD with a fully explicit solvent but at a greatly reduced cost. This reduction in computational requirements can enable simulations on longer time scales for the same system size or permit application of REMD to the study of much larger systems. We also showed that use of one or two explicit solvent shells alone was inadequate and that adding a reaction field was essential for obtaining

reasonable results. Adaptation of this method to other continuum models (such as the more rigorous PB) should be straightforward. Since the continuum solvent is only used for the infrequent exchange calculations, models that are too complex for use at each step of dynamics can be readily employed.

Acknowledgment. The authors thank Adrian Roitberg for helpful feedback, John Mongan and Alexey Onufriev for valuable discussions concerning the GB models, and Guanglei Cui for help with Amber REMD. Roberto Gomperts provided important code optimizations for Amber. Supercomputer time at NCSA (NPACI MCA02N028) and financial support from the National Institutes of Health (NIH GM6167803) and Department of Energy (Contract DE-AC02-98CH10886) are gratefully acknowledged. Additional computer time was generously provided by the SGI Engineering group. C.S. is a Cottrell Scholar of Research Corporation.

Supporting Information Available: Basin populations for Ala_1 and Ala_3 using hybrid $GB^{OBC'}$, free energy profiles for backbone conformations in Ala5 for Ala_{10}, radial distribution functions for water near Ala_3, and ranges for definition of secondary structure basins. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Tai, K. *Biophys. Chem.* **2004**, *107* (3), 213−220.

(2) Roitberg, A.; Simmerling, C. *J. Mol. Graphics Modell.* **2004**, *22* (5), 317−317.

(3) Hansmann, U. H. E. *Chem. Phys. Lett.* **1997**, *281* (1−3), 140−150.

(4) Swendsen, R. H.; Wang, J. S. *Phys. Rev. Lett.* **1986**, *57* (21), 2607−2609.

(5) Tesi, M. C.; vanRensburg, E. J. J.; Orlandini, E.; Whittington, S. G. *J. Stat. Phys.* **1996**, *82* (1−2), 155−181.

(6) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314* (1−2), 141−151.

(7) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087−1092.

(8) Feig, M.; Karanicolas, J.; Brooks, C. L. *J. Mol. Graphics Modell.* **2004**, *22* (5), 377−395.

(9) Garcia, A. E.; Sanbonmatsu, K. Y. *Proteins: Struct., Funct., Genet.* **2001**, *42* (3), 345−354.

(10) Garcia, A. E.; Sanbonmatsu, K. Y. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99* (5), 2782−2787.

(11) Karanicolas, J.; Brooks, C. L. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100* (7), 3954−3959.

(12) Pitera, J. W.; Swope, W. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100* (13), 7587−7592.

(13) Sugita, Y.; Kitao, A.; Okamoto, Y. *J. Chem. Phys.* **2000**, *113* (15), 6042−6051.

(14) Zhou, R.; Berne, B. J.; Germain, R. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98* (26), 14931−6.

(15) Kinnear, B. S.; Jarrold, M. F.; Hansmann, U. H. E. *J. Mol. Graphics Modell.* **2004**, *22* (5), 397−403.

(16) Roe, D. R.; Hornak, V.; Simmerling, C. *J. Mol. Biol.* **2005**, *352* (2), 370−381.

(17) Rathore, N.; Chopra, M.; de Pablo, J. J. *J. Chem. Phys.* **2005**, *122* (2), 024111.

(18) Fukunishi, H, W. O., Takada, S. *J. Chem. Phys.* **2002**, *116* (20), 9058−9067.

(19) Cheng, X. L.; Cui, G. L.; Hornak, V.; Simmerling, C. *J. Phys. Chem. B* **2005**, *109* (16), 8220−8230.

(20) Kofke, D. A. *J. Chem. Phys.* **2002**, *117* (15), 6911−6914.

(21) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **2000**, *329* (3−4), 261−270.

(22) Mitsutake, A., S. Y., Okamoto, Y. *J. Chem. Phys.* **2003**, *118* (14), 6664−6688.

(23) Jang, S.; Shin, S.; Pak, Y. *Phys. Rev. Lett.* **2003**, *91* (5), 58305.

(24) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112* (16), 6127−6129.

(25) Nymeyer, H.; Garcia, A. E. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100* (24), 13934−13939.

(26) Zhou, R.; Berne, B. J. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99* (20), 12777−82.

(27) Zhou, R. *Proteins* **2003**, *53* (2), 148−61.

(28) Simmerling, C.; Strockbine, B.; Roitberg, A. *J. Am. Chem. Soc.* **2002**, *124* (38), 11258.

(29) Ghosh, A.; Rapp, C. S.; Friesner, R. A. *J. Phys. Chem. B* **1998**, *102* (52), 10983−10990.

(30) Srinivasan, J.; Cheatham, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A. *J. Am. Chem. Soc.* **1998**, *120* (37), 9401−9409.

(31) Luo, R.; David, L.; Gilson, M. K. *J. Comput. Chem.* **2002**, *23* (13), 1244−1253.

(32) Jeancharles, A.; Nicholls, A.; Sharp, K.; Honig, B.; Tempczyk, A.; Hendrickson, T. F.; Still, W. C. *J. Am. Chem. Soc.* **1991**, *113* (4), 1454−1455.

(33) Alper, H.; Levy, R. M. *J. Chem. Phys.* **1993**, *99* (12), 9847−9852.

(34) Beglov, D.; Roux, B. *Biopolymers* **1995**, *35* (2), 171−178.

(35) Beglov, D.; Roux, B. *J. Chem. Phys.* **1994**, *100* (12), 9050−9063.

(36) Brooks, C. L.; Brunger, A.; Karplus, M. *Biopolymers* **1985**, *24* (5), 843−865.

(37) Brooks, C. L.; Karplus, M. *J. Chem. Phys.* **1983**, *79* (12), 6312−6325.

(38) Kentsis, A.; Mezei, M.; Gindin, T.; Osman, R. *Proteins: Struct., Funct., Bioinformatics* **2004**, *55* (3), 493−501.

(39) King, G.; Warshel, A. *J. Chem. Phys.* **1989**, *91* (6), 3647−3661.

(40) Lee, M. S.; Olson, M. A. *J. Phys. Chem. B* **2005**, *109* (11), 5223−5236.

(41) Lee, M. S.; Salsbury, F. R.; Olson, M. A. *J. Comput. Chem.* **2004**, *25* (16), 1967−1978.

(42) Das, B.; Helms, V.; Lounnas, V.; Wade, R. C. *J. Inorg. Biochem.* **2000**, *81* (3), 121−131.

(43) Topol, I. A.; Tawa, G. J.; Burt, S. K.; Rashin, A. A. *J. Chem. Phys.* **1999**, *111* (24), 10998−11014.

(44) van der Spoel, D.; van Maaren, P. J.; Berendsen, H. J. C. *J. Chem. Phys.* **1998**, *108* (24), 10220−10230.

(45) Vorobjev, Y. N.; Hermans, J. *Biophys. Chem.* **1999**, *78* (1−2), 195−205.

(46) Errington, N.; Doig, A. J. *Biochemistry* **2005**, *44* (20), 7553−8.

(47) Groebke, K.; Renold, P.; Tsang, K. Y.; Allen, T. J.; McClure, K. F.; Kemp, D. S. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93* (9), 4025−9.

(48) Marqusee, S.; Robbins, V. H.; Baldwin, R. L. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86* (14), 5286−5290.

(49) Maison, W.; Arce, E.; Renold, P.; Kennedy, R. J.; Kemp, D. S. *J. Am. Chem. Soc.* **2001**, *123* (42), 10245−54.

(50) Heitmann, B.; Job, G. E.; Kennedy, R. J.; Walker, S. M.; Kemp, D. S. *J. Am. Chem. Soc.* **2005**, *127* (6), 1690−704.

(51) Chen, K.; Liu, Z. G.; Kallenbach, N. R. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101* (43), 15352−15357.

(52) McColl, I. H.; Blanch, E. W.; Hecht, L.; Kallenbach, N. R.; Barron, L. D. *J. Am. Chem. Soc.* **2004**, *126* (16), 5076−5077.

(53) Shi, Z. S.; Olson, C. A.; Rose, G. D.; Baldwin, R. L.; Kallenbach, N. R. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99* (14), 9190−9195.

(54) Asher, S. A.; Mikhonin, A. V.; Bykov, S. *J. Am. Chem. Soc.* **2004**, *126* (27), 8433−8440.

(55) Mezei, M.; Fleming, P. J.; Srinivasan, R.; Rose, G. D. *Proteins: Struct., Funct., Bioinformatics* **2004**, *55* (3), 502−507.

(56) Garcia, A. E. *Polymer* **2004**, *45* (2), 669−676.

(57) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *Chem. Phys. Lett.* **1995**, *246* (1−2), 122−129.

(58) Feig, M.; Onufriev, A.; Lee, M. S.; Im, W.; Case, D. A.; Brooks, C. L. *J. Comput. Chem.* **2004**, *25* (2), 265−284.

(59) Onufriev, A.; Bashford, D.; Case, D. A. *J. Phys. Chem. B* **2000**, *104* (15), 3712−3720.

(60) Kofke, D. A. *J. Chem. Phys.* **2004**, *121* (2), 1167−1167.

(61) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117* (19), 5179−5197.

(62) Wang, J. M.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21* (12), 1049−1074.

(63) Hornak, V.; Simmerling, C. Manuscript in preparation.

(64) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926−935.

(65) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26* (16), 1668−1688.

(66) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23* (3), 327−341.

(67) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Dinola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81* (8), 3684−3690.

(68) Simmerling, C.; Elber, R.; Zhang, J., MOIL-View − A Program for Visualization of Structure and Dynamics of Biomolecules and STO − A Program for Computing Stochastic Paths. In *Modelling of Biomolecular Structures and Mechanisms*; Pullman et al., A., Ed.; Kluwer Academic Publishers: Netherlands, 1995; pp 241−265.

(69) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98* (12), 10089−10092.

(70) Bondi, A. *J. Phys. Chem.* **1964**, *68* (3), 441−451.

(71) Tsui, V.; Case, D. A. *J. Am. Chem. Soc.* **2000**, *122* (11), 2489−2498.

(72) Ponder, J. W.; Richards, F. M. *J. Comput. Chem.* **1987**, *8* (7), 1016−1024.

(73) Smith, L. J.; Daura, X.; van Gunsteren, W. F. *Proteins: Struct., Funct., Genet.* **2002**, *48* (3), 487−496.

(74) Liu, P.; Kim, B.; Friesner, R. A.; Berne, B. J. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102* (39), 13749−13754.

(75) Shi, Z.; Woody, R. W.; Kallenbach, N. R. *Adv. Prot. Chem.* **2002**, *62*, 163−240.

CT050196Z