THE JOURNAL OF PHYSICAL CHEMISTRY B LETTERS

2008, *112,* 657−660
Published on Web 12/29/2007

Prediction of Partition Coefficients by Multiscale Hybrid Atomic-Level/Coarse-Grain Simulations

Julien Michel, Mario Orsi, and Jonathan W. Essex*

School of Chemistry, University of Southampton, Highfield, Southampton SO17 1BJ, U.K.

Received: August 1, 2007

Coarse-grain models are becoming an increasingly important tool in computer simulations of a wide variety of molecular processes. In many instances it is, however, desirable to describe key portions of a molecular system at the atomic level. There is therefore a strong interest in the development of simulation methodologies that allow representations of matter with mixed granularities in a multiscale fashion. We report here a strategy to conduct mixed atomic-level and coarse-grain simulations of molecular systems with a recently developed coarse-grain model. The methodology is validated by computing partition coefficients of small molecules described in atomic detail and solvated by water or octane, both of which are represented by coarse-grain models. Because the present coarse-grain force field retains electrostatic interactions, the simplified solvent particles can interact realistically with the all-atom solutes. The partition coefficients computed by this approach rival the accuracy of fully atomistic simulations and are obtained at a fraction of their computational cost. The present methodology is simple, robust and applicable to a wide variety of molecular systems.

Atomic-level (AL) computer simulations have become an established tool to probe the structure and dynamics of molecular systems by Monte Carlo or molecular dynamics techniques.[1] However, owing to their computational expense, AL simulations are generally limited in size (hundreds of thousands of atoms) and time (hundreds of nanoseconds) scales. To overcome this difficulty, the field of molecular simulations has witnessed a renewed interest in coarse-grain (CG) models that reduce significantly the computational cost by subsuming several atomic sites into single particles.[2−5] In recent years, CG strategies have been applied to the study of complex molecular processes that are not yet within reach of AL simulations.[6,7]

In many instances, it would be desirable to retain an atomic level description of parts of a molecular system, for example, the amino acids that form a channel in a transmembrane protein and the solutes that permeate through it, while the membrane and solvent are represented with a CG model. Indeed, a few mixed atomistic coarse-grain (AL-CG) schemes have been recently proposed in the literature.[8,9] For the sake of efficiency, CG potentials are often highly simplified. Typically, electrostatics are highly simplified[10,11] or even neglected.[12,13] Although these CG models can reproduce thermodynamic properties satisfactorily, this usually requires extensive parametrization and it is not clear how, if, and which CG force fields could be compatible with atomistic potentials. Because of the wide availability and reliability of AL force fields, it is highly desirable to develop an AL-CG strategy that allows for realistic interactions between both representations.

Recently, our group has developed a CG approach to conduct molecular dynamics simulations of lipid bilayers.[14] Unlike many alternatives, this CG model fully retains electrostatic interactions through inclusion of point dipoles and point charges. The lipid head groups consist of Lennard-Jones sites that carry a partial charge, and the glycerol group and hydrocarbon tails are represented by the Gay−Berne potential.[15] The Gay−Berne model has been employed for decades in simulations of liquid crystals and allows the simulation of ellipsoidal particles, which capture more realistically than spheres the elongated shape of lipid tails. The soft sticky dipole potential is employed to model water.[16] This single-site model has been shown to reproduce the main physical properties of water as accurately as popular three- or four-site models, while being an order of magnitude more efficient. Using this approach, a CG model that reproduces quantitatively many key properties of a fluid phase dimyristoylphosphatidylcholine (DMPC) bilayer has been constructed.[14] The methodology proved to be about 2 orders of magnitude more efficient than traditional AL simulations. Because this CG model retains electrostatic interactions, without requiring the imposition of an explicit dielectric constant for screening, we sought to determine whether or not it could be combined with atomistic models of small molecules. This feature would be particularly advantageous to allow, for instance, computational studies of small molecule permeation across a lipid bilayer. Subtle differences in functional groups can impact significantly the rates of permeation and it is unlikely that these could be reproduced by a fully CG model without lengthy parametrization.

Equation 1 summarizes the key terms of the potential energy function employed in this study

$$U = U(\text{intra}) + U(\text{AL,AL}) + U(\text{SSD,SSD}) + U(\text{GB,GB}) + U(\text{SSD,AL}) + U(\text{GB,AL}) \quad (1)$$

U(intra) and U(AL,AL) account for the intramolecular and

* To whom correspondence should be addressed. Phone: +44 238059 2794. Fax: +44 238059 3781. E-mail: J.W.Essex@soton.ac.uk.

10.1021/jp076142y CCC: $40.75

intermolecular energies of the all-atom solutes and solvent molecules. U(SSD,SSD) is the interaction energy between soft sticky dipole waters. We use the parametrization of Fennel and Gezelter, denoted SSD/E.[17] U(GB,GB) is the interaction energy between two Gay−Berne particles. Detailed equations for these terms can be found elsewhere.[15,16]

The last two terms are introduced to handle interactions between AL solutes and CG solvents. Equation 2 details their functional form for a pair of atom i and CG site j

$$U_{ij}(\mathrm{SSD,AL}) = 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \alpha\frac{q_i\vec{\mu}_j\cdot\vec{r}_{ij}}{r_{ij}^{3}}$$

$$U_{ij}(\mathrm{GB,AL}) = 4\beta\epsilon(\vec{r}_{ij},\vec{u}_j)\left[\left(\frac{\sigma_{\mathrm{GB}}}{r-\sigma(\vec{r}_{ij},\vec{u}_j)+\sigma_{\mathrm{GB}}}\right)^{12} - \left(\frac{\sigma_{\mathrm{GB}}}{r-\sigma(\vec{r}_{ij},\vec{u}_j)+\sigma_{\mathrm{GB}}}\right)^{6}\right] \quad (2)$$

The interaction between atomistic particles and SSD water, U_{ij}(SSD,AL), consists of a Lennard-Jones and monopole-dipole term. The Lennard-Jones cross terms σ_{ij} and ϵ_{ij} are computed using the standard combining rule.[1] The partial charge, q_i, and water dipole, μ_j, are obtained from the AL or SSD force field. The potential term U_{ij}(GB,AL) accounts for the mixed Lennard-Jones/Gay−Berne interactions using the generalized Gay−Berne potential; σ and ϵ are functions of the separation $\vec{r}_{ij}$ of the two particles i,j and the orientation $\vec{u}_j$ of particle j, and σ_{GB} depends on the parameters of the Gay−Berne particle. Detailed formulas can be found elsewhere.[18] The terms α and β will be discussed later in the text.

To demonstrate the viability of the present potential energy function, we have computed partition coefficients between water and octane for a set of 15 analogues of neutral amino acid side chains, listed in the Supporting Information. These molecules have been studied extensively by fully atomistic simulations[19−22] and are particularly relevant to determine if AL-CG simulations of membrane proteins can be reliably conducted. Water is represented by the SSD potential discussed above. Octane is modeled by two Gay−Berne ellipsoids connected by a weak harmonic spring; the parameters used correspond to those of the tail sites in the CG DMPC lipid model.[14] All atom models of the solutes were constructed using the program Molden[23] and parametrized with the GAFF force field[24] and AM1/BCC atomic partial charges.[25] Monte Carlo free energy simulations were performed with a modified version of the program ProtoMS2.1.[26] Equilibrated solvent boxes of TIP4P water,[27] SSD water, and GB octane were prepared by extensive equilibration in the NPT ensemble at 1 atm and 25 °C. Absolute solvation free energies were computed with an implementation of a double annihilation protocol and replica exchange thermodynamic integration (see Supporting Information).[28,29] Each simulation was repeated five times with a different random number seed; free energies were taken as the mean of these five simulations, and one standard error was taken as an estimate of the statistical error. From the solvation free energies calculated in both solvents, partition coefficients are then derived according to eq 3:[30,31]

$$\log P = \frac{\Delta G_{\mathrm{solv}}(\mathrm{water}) - \Delta G_{\mathrm{solv}}(\mathrm{octane})}{2.303RT} \quad (3)$$

Extensive experimental data for solvation free energies of small molecules in several solvents has been compiled previously.[32] Unfortunately, solvation free energies in octane for all the present compounds are lacking. However, solvation free energies in cyclohexane are available for all the compounds, and for those molecules where data in octane and cyclohexane exist, the solvation free energies are similar to within 0.3 kcal mol^{-1} on average.[33] Given that the present CG force field would not distinguish easily octane and cyclohexane, we make the approximation that the simulation results can be compared to the experimental data compiled for cyclohexane.

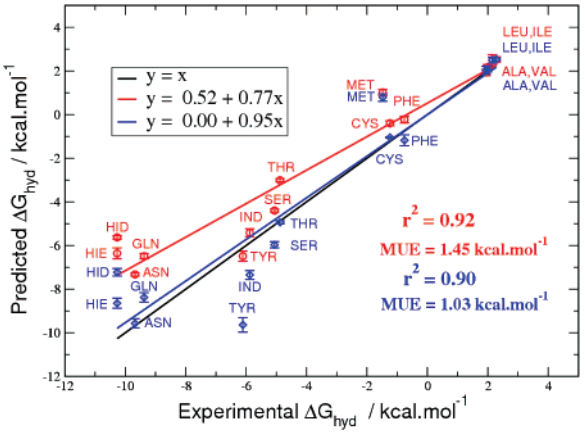

Figure 1. Comparison between predicted and experimental hydration free energies for a dataset of analogues of amino acid side chains in a box of water represented by the soft sticky dipole potential. In red, are the results obtained without any empirical scaling parameter ($\alpha = 1.0$); in blue, the results obtained with α set to 1.10. Linear regression equations are also plotted.

We consider first the simulation results for water. Figure 1 plots the predicted hydration free energies of the atomistic solutes in SSD water against those measured experimentally. The mean unsigned error is 1.45 kcal mol^{-1}, the coefficient of determination r^2 is 0.92 and the slope is 0.77. Because similar results for the GAFF/AM1 BCC force field in TIP4P water have not been reported, we have also carried out the same calculations in TIP4P water (see Supporting Information). The mean unsigned error is 1.64 kcal mol^{-1}, the coefficient of determination r^2 is 0.94, and the slope is 0.97. The results with TIP4P are consistent with those of Shirts et al. obtained for the AMBER force field and TIP3P without a long range correction term.[20] This is not unexpected as the AM1/BCC atomic partial charges have been developed to mirror the electrostatic potentials obtained by the RESP/HF-6-31G* method from which the atomic partial charges of the AMBER force field were derived.[34] Mobley et al. recently reported hydration free energies for the GAFF/AM1-BCC force field in TIP3P water that are somewhat more accurate than the results reported here (MUE 0.89 kcal mol^{-1}, $r^2 = 0.98$, slope = 1.00).[35] We attribute the differences to their use of a long range correction term and the particle mesh Ewald method[36] as opposed to a spherical residue based cutoff in our simulations.

Although the MUE in SSD water is lower than that for TIP4P water, it is apparent that solute−solvent polar interactions are systematically underestimated with the present combination of AL and CG force fields. On the other hand, the hydration free energy of apolar compounds, where electrostatic interactions are negligible, are well predicted.

Because we wish to ensure that AL and CG models can be mixed without extensive re-parametrization, we did not attempt to modify CG or AL parameters and sought instead to calibrate the magnitude of the AL-CG interactions by introducing an empirical parameter α that scales the solute−solvent charge-dipole interactions. Figure 1 shows that this simple modification

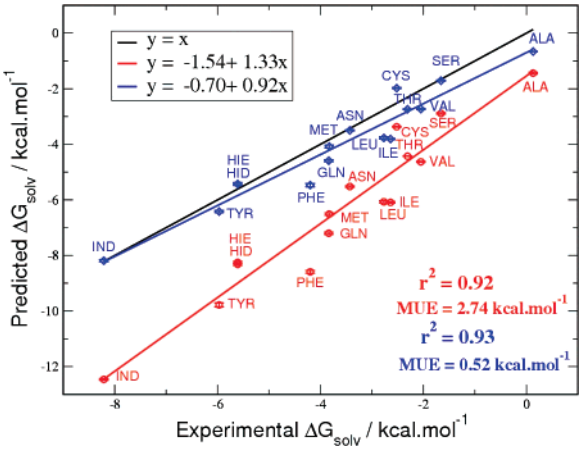

Figure 2. Comparison between predicted and experimental solvation free energies for a dataset of analogues of amino acid side chains in a box of octane represented by two Gay−Berne ellipsoids. In red, are the results obtained without any empirical scaling parameter ($\beta = 1.0$); in blue, the results obtained with β set to 0.80. Linear regression equations are also plotted.

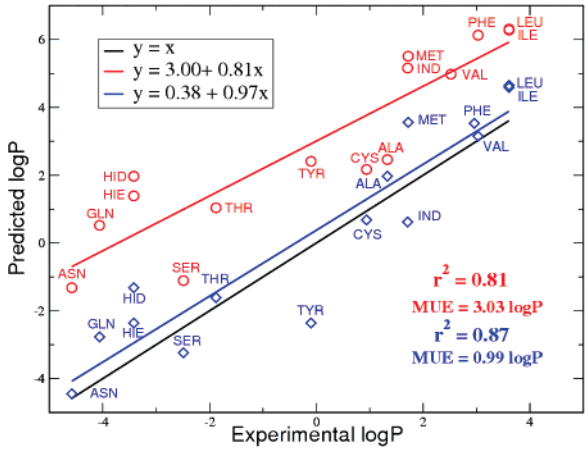

Figure 3. Comparison between predicted and experimental partition coefficients between water and octane for a dataset of analogues of amino acid side chains. In red, are the results obtained without any empirical scaling parameter α and β; in blue, the results obtained with α set to 1.10 and β set to 0.80. Linear regression equations are also plotted.

significantly increases the agreement between experiment and predictions when α is set to 1.10. Indeed, the predictions are of comparable accuracy to those obtained with TIP4P. This is very encouraging given that the soft sticky dipole potential is about 1 order magnitude more efficient than TIP4P.[16]

We now consider the simulations results for octane. An all-atom model of octane would have 26 atomic sites, but the present CG model has only two sites. Typical nonbonded interactions scale as N^2, and one can expect a 2 orders of magnitude increase in efficiency upon switching to this CG representation. Despite the simplicity of the CG model, the density of liquid octane at 25 °C is reproduced to within 7%. Figure 2 shows that the predicted solvation free energies of the solutes in octane are systematically overestimated. It is also apparent that the heavier the solutes, the larger the discrepancy; thus an empirical scaling of the GB-AL interactions similar to the scaling of the SSD-AL electrostatic interactions is expected to improve the quality of the predictions. It is indeed found that with a scaling parameter of β set to 0.80, the discrepancy between simulation and experiment can be considerably reduced.

The statistical error on the calculated solvation free energies in octane is extremely low, of the order of 0.01−0.05 kcal mol^{-1}. Thus, the simulations, already computationally efficient, could have been shortened considerably and yet provide solvation free energies with acceptable precision. On the other hand, the simulations conducted in SSD water are on average only slightly more precise than those conducted (on the order of 0.10−0.30 kcal mol^{-1}) in TIP4P.

From the solvation free energies computed in water and octane, we now compute partition coefficients according to eq 3. Figure 3 plots partition coefficients with and without correction terms for this set of small molecules. With the unadjusted interactions, the solutes favor too much the apolar solvent. Introduction of the parameters α and β improves considerably the agreement with experiment and the overall inaccuracy is about one log P unit, which is comparable to results from fully atomistic simulations.[21]

As further validation, we have computed partition coefficients for ten other solutes (see Supporting Information) that carry functional groups different from those present in the amino acid side chain analogues. Using the parameters α and β set to 1.10 and 0.80, for the SSD simulations, the mean unsigned error is reduced from 1.66 to 1.13 kcal mol^{-1}. For the octane simulations the mean unsigned error is reduced from 4.01 to 1.32 kcal mol^{-1}. Overall, the mean unsigned error on the partition coefficients is improved from 4.16 to 1.57 log P units. This strongly suggests that the simple adjustments we propose are transferable across functional groups commonly encountered in organic molecules. Clearly, some artifacts remain; for instance, the hydration free energy of *p*-cresol is too large after adjustments. Although we could possibly improve agreement with additional correction terms, it would be difficult to justify them in the absence of a very large validation set. We find the simplicity of the present empirical adjustments appealing, robust, and transferable and further improvements should be sought with caution. The present work demonstrates that it is feasible to conduct hybrid atomistic coarse-grain simulations with accuracy comparable to fully atomistic representations. A unique aspect of the present methodology is that the CG force field retains electrostatic interactions,[14] without requiring an explicit dielectric constant. This makes the embedding of atomistic particles relatively straightforward and unlike other existing AL-CG schemes,[8,9] without a need for extensive reparameterization. This work paves the way for permeation studies of small atomistic molecules across a coarse-grain lipid bilayer and, ultimately, mixed AL-CG simulations of membrane proteins. The implications of this work go beyond studies of biomembranes and the availability of this multiscale AL-CG methodology should prove useful to the molecular simulations community as a whole.

Acknowledgment. We thank the University of Southampton and BBSRC for funding this work and EPSRC (GR/R06137/01) for providing computational resources.

Supporting Information Available: Details of the simulation protocol and plots of computed hydration free energies with the TIP4P model. Name of the amino acid side chains analogues. Table of computed solvation free energies and partition coefficients for ten organic compounds: aniline, ethylamine, diethylamine, methylethanoate, ethylethanoate, 2-methylpyrazine, propanol, propanone, butanone, 1-nitropropane. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Allen, M. P.; Tildesley, D. J. In *Computer Simulations of Liquids*; Oxford University Press: Oxford, U.K., 1987.

(2) Nielsen, S. O.; Lopez, C. F.; Srinivas, G.; Klein, M. L. *J. Phys. Condens. Matter* **2004**, *16*, R481−R512.
(3) Kolinski, A.; Skolnick, J. *Polymer* **2004**, *45*, 511−524.
(4) Tozzini, V. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144−150.
(5) Orsi, M.; Sanderson, W.; Essex, J. W. *Molecular Interactions - Bringing Chemistry to Life*; Beilstein-Institut: Berlin, in press.
(6) Elcock, A. H. *PLOS Comput. Biol.* **2006**, *2*, 824−841.
(7) Arkhipov, A.; Freddolino, P. L.; Imada, K.; Namba, K.; Schulten, K. *Biophys. J.* **2006**, *91*, 4589−4597.
(8) Shi, Q.; Izvekov, S.; Voth, G. A. *J. Phys. Chem. B* **2006**, *110*, 15045−15048.
(9) Neri, M.; Anselmi, C.; Cascella, M.; Maritan, A.; Carloni, P. *Phys. Rev. Lett.* **2005**, *95*, 218102.
(10) Marrink, S.-J.; de Vries, A. H.; Mark, A. E. *J. Phys. Chem. B* **2004**, *108*, 750−60.
(11) Shelley, J. C.; Shelley, M. Y.; Reeder. R. C.; Ramdyopadhyay, S.; Klein, M. L. *J. Phys. Chem. B* **2001**, *105*, 4464−4470
(12) Izvekov, S.; Voth, G.A. *J. Phys. Chem. B* **2005**, *109*, 2469−2473.
(13) Kranenburg, M.; Nicolas, J. P.; Smit, B. *Phys. Chem. Chem. Phys.* **2004**, *6*, 4142−4151.
(14) Orsi, M.; Haubertin, D. Y.; Sanderson, W. E.; Essex, J. W. *J. Phys. Chem. B*, in press, DOI: 10.1021/jp076139e.
(15) Gay, J. G.; Berne, B. J. *J. Chem. Phys.* **1981**, *74*, 3316−3319.
(16) Liu, Y.; Ichiye, T. *J. Phys. Chem.* **1996**, *100*, 2723−2730.
(17) Fennell, C. J.; Gezelter, J. D. *J. Chem. Phys.* **2004**, *120*, 9175−9184.
(18) Cleaver, D. J.; Care, C. M.; Allen, M. P.; Neal, M. P. *Phys. Rev. E* **1996**, *54*, 559−567.
(19) Villa, A.; Mark, A. E. *J. Comput. Chem.* **2002**, *23*, 548−553.
(20) Shirts, M. R.; Pitera, J. W.; Swope, W. C.; Pande, V. S. *J. Chem. Phys.* **2003**, *119*, 5740−5761.
(21) MacCallum, J. L.; Tieleman, D. P. *J. Comput. Chem.* **2003**, *24*, 1930−1935.
(22) Deng, Y.; Roux, B. *J. Phys. Chem. B* **2004**, *108*, 16567−16576.
(23) Schaftenaar, G.; Noordik, J. *J. Comput.-Aided. Mol. Design* **2000**, *14*, 123−134.
(24) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157−1174.
(25) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. *J. Comput. Chem.* **2000**, *21*, 132−146.
(26) Woods, C. J.; Michel, J. "ProtoMS2.1", in house Monte Carlo software.
(27) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926−935.
(28) Woods, C. J.; Essex, J. W.; King, M. A. *J. Phys. Chem. B* **2003**, *107*, 13703−13710.
(29) Michel, J.; Verdonk, M. L.; Essex, J. W. *J. Chem. Theory Comput.* **2007**, *3*, 1645−1655.
(30) Essex, J. W.; Reynolds, C. A.; Richards, W. G. *J. Am. Chem. Soc.* **1992**, *114*, 3634−3639.
(31) Essex, J. W.; Reynolds, C. A.; Richards, W. G. *J. Chem. Soc., Chem. Commun.* **1989**, *16*, 1152−1154.
(32) Zhu, T. H.; Li, J. B.; Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Phys.* **1998**, *109*, 9117−9133.
(33) Radzicka, A.; Wolfenden, R. *Biochemistry* **1988**, *27*, 1664−1670.
(34) 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*, 5179−5197.
(35) Mobley, D. L.; Dumont, E.; Chodera, J. D.; Dill, K. A. *J. Phys. Chem. B* **2007**, *111*, 2242−2254.
(36) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577−8593.