Theor Chem Acc (2012) 131:1162
DOI 10.1007/s00214-012-1162-6

REGULAR ARTICLE

A coarse-grained model for *β*-D-glucose based on force matching

Sergiy Markutsya · Yana A. Kholod · Ajitha Devarajan · Theresa L. Windus · Mark S. Gordon · Monica H. Lamm

Received: 20 October 2011 / Accepted: 17 January 2012 / Published online: 2 March 2012

Abstract Cellulosic ethanol production is a two-stage process that involves the hydrolysis of cellulose to form simple sugars and the fermentation of these sugars to ethanol. Hydrolysis of cellulose is the rate-limiting step, and there is a great need to characterize the process with numerical simulations to better understand the complex mechanisms involved. The ultimate goal is to generate accurate coarse-grained molecular models that are capable of predicting the structure of lignocellulose before and after pretreatment so that subsequent ab initio calculations can be performed to probe the degradation pathways. As a first step toward that goal, the force-matching method is used to derive coarse-grained models for *β*-D-glucose molecules in aqueous solution. Using the same reference, an all-atom molecular dynamics simulation trajectory, two sets of three- and six-site coarse-grained models of *β*-D-glucose are developed using two definitions of the coarse-grained center site location: center of mass (CG-CM) and geometric center (CG-GC). The performance of these coarse-grained models is evaluated by comparing the coarse-grained predictions for bond-length distributions and radial distribution functions to those obtained from the all-atom reference simulation. The six-site coarse-grained models retain more structural details than the three-site coarse-grained models. Comparison between center site definitions shows that CG-CM models generally predict local ordering better, while CG-GC models predict long-range structure better.

Keywords Coarse-grain force fields · Glucose · Glucopyranose · Molecular dynamics simulation

1 Introduction

Biomass conversion of cellulosic material, such as grasses and wood chips, is a promising route for producing renewable energy and chemical feedstocks [1]. The most important steps in the biomass-to-energy production process are acidic or enzymatic hydrolysis to break the crystalline cellulose structure into glucose molecules, and subsequent fermentation to convert glucose to ethanol as the final product [2]. The hydrolysis step is a bottleneck in this process due to the long reaction time and low efficiency. To understand how to overcome this bottleneck, accurate and detailed studies of the hydrolysis reaction are required. Numerical simulations are one tool that is able to deliver the desired level of detail for the process and also to satisfy accuracy requirements.

Molecular dynamics simulations for an all-atom crystalline cellulose system solvated with water molecules ($\sim 10^9$ atoms) is a very challenging computational task. An alternative is to design a numerical simulation of the equivalent coarse-grained system to reduce the number of degrees of freedom. Such coarse-grained models should be

Published as part of the special collection of articles: From quantum mechanics to force fields: new methodologies for the classical simulation of complex systems.

S. Markutsya · M. H. Lamm
Department of Chemical and Biological Engineering,
Iowa State University, Ames, IA, USA

S. Markutsya · Y. A. Kholod · A. Devarajan ·
T. L. Windus · M. S. Gordon (✉) · M. H. Lamm
Ames Laboratory, Ames, IA 50011, USA
e-mail: mark@si.fi.ameslab.gov

Y. A. Kholod · A. Devarajan · T. L. Windus · M. S. Gordon
Department of Chemistry, Iowa State University,
Ames, IA, USA

able to accurately represent most important physical and chemical properties of the system while significantly reducing the computational cost. Coarse-grained models are usually created by combining multiple atoms into one group and then representing this group as a single effective coarse-grained site, thereby significantly reducing the complexity of the system. There are generally many alternative choices for the coarse-graining sites, and one must make sensible choices, usually based on physical, chemical, and structural properties of the system. However, it is often difficult to predict whether a particular set of coarse-grained sites are the best equivalent representation of the corresponding all-atom system.

Once the coarse-grained sites have been chosen, the effective forces/potentials between these coarse-grained sites must be derived and several coarse-graining strategies have been applied. Energy-based coarse-graining methods [3, 4] parameterize the coarse-grained potentials to match thermodynamic properties, such as the partitioning of a species between oil and water. In other strategies, such as Boltzmann inversion (BI) [5], iterative Boltzmann inversion (IBI) [6], and inverse Monte Carlo (IMC) [6–9], the coarse-grained potentials are parameterized to reproduce the average structure observed in all-atom simulations. The force-matching [10–12] scheme takes a different approach and bases the coarse-grained model on the reference inter- and intra-atom forces. In force matching, the pairwise effective force field is derived from a trajectory obtained with all-atom molecular dynamics simulations. The original force-matching approach [10] is not suitable for high molecular weight biological molecules because of the large number of coarse-grained sites in the system. To overcome this problem, a new multiscale coarse-graining (MS-CG) approach has been developed [11, 12]. With the MS-CG approach, the force field is assumed to depend linearly on the fitting parameters and the problem is reduced to the solution of an over-determined system. This over-determined system is solved for smaller sets of all-atom configurations (subsets), and the final solution is obtained by averaging over all subsets. The coarse-grained potential obtained in this way is a potential of mean force (PMF), which is the derivative of the free energy in the phase space with reduced degrees of freedom. Thus, the coarse-grained model can only be used for the identical thermodynamic state point at which the all-atom system is simulated. In any coarse-graining approach, configurational entropy is lost as the degrees of freedom in the system are reduced. The loss of configurational entropy was quantified for hydrocarbon chains and shown to increase as the flexibility of the chain increases [13]. The configurational entropy loss is significant when the degrees of freedom are drastically reduced, such as when an implicit solvent coarse-grained model is derived for a polymer or protein. Kim and Lamm recently demonstrated a method for restoring the lost configurational entropy to coarse-grained models of flexible macromolecules in solution when the solvent degrees of freedom have been removed [14].

A coarse-grained model for malto-oligosaccharides and their aqueous mixtures has been developed by Molinero and Goddard [15, 16] using an IMC method to derive non-bonded interactions and the BI method to extract bonded interactions. In this model, the glucose monomer is represented by three sites, and each water molecule by a single site. In their approach, the non-bonded interactions are described with two-body Morse potentials, and valence interactions between two coarse-grained sites are described as conventional bond, angle, and torsion interactions. However, the Morse function is not necessarily optimal for interactions between coarse-grained sites. Therefore, their model does not always accurately predict such structural properties of the system as the radial distribution function, bond lengths, angles, and dihedrals.

Izvekov and Voth have used their multiscale coarse-graining (MS-CG) method [11, 12] that is based on force matching to build a three-site coarse-grained model for α-D-glucose and α-(1 → 4)-D-glucan with 14 glucose units in aqueous solution [17]. The MS-CG method was used to build non-bonded interactions, whereas the bonded interactions are obtained with a Boltzmann statistical analysis of the all-atom trajectory. With this approach, they were able to reproduce many experimental structural and thermodynamical properties in the constant NPT ensemble.

A hybrid of the force-matching and Boltzmann inversion (BI) methods was used by Hynninen et al. to develop a three-site model to describe β-D-glucose, cellobiose, and cellotetraose [18]. They used a pure force-matching method for β-D-glucose and hybrid methods for cellobiose and cellotetraose in aqueous solution. In their hybrid method, the non-bonded interactions were obtained from the force-matching method, while a BI was used to fit the bond distances, angles, and dihedral parameters. Their coarse-grained model yields a good match with all-atom simulations for structural properties.

Recently, Srinivas et al. [19] have used the Boltzmann inversion method to develop a one-site model for a cellulose crystalline fibril containing 36 chains with 40 cellobiose units in each chain in aqueous solution. In addition, a coarse-grained simulation of fully amorphous cellulose fibrils was performed. The method provides an accurate and constraint-free approach to derive coarse-grained models for cellulose with a wide range of crystallinity.

To develop a coarse-grained model that can accurately represent physical and chemical properties of all-atom systems, it is important to have both an accurate, robust course-graining method and representative coarse-grained sites that are able to provide an acceptable level of accuracy. Coarse-grained models have been built for many systems. However, there appears to be no systematic study available on how the choice of the number and locations of coarse-grained sites might influence the predictions of a coarse-grained model. Moreover, it is difficult to predict a priori the most appropriate and effective coarse-graining parameters.

In this work, the force-matching approach [11, 12] is employed to develop a coarse-grained model of β-D-glucose in water solution as a structural unit of cellulose. For this system, two different coarse-grained mapping schemes (three-site and six-site) are investigated. In addition to varying the number of atoms in a coarse-grained site, two definitions of the center site location for the coarse-grained sites, center of mass (CM) and geometric center (GC), were considered and evaluated in detail. This provides a systematic comparative analysis of these coarse-grained mapping alternatives.

The paper is organized as follows: In Sect. 2, the force-matching method is briefly described, the coarse-grained mapping schemes are defined, and the parameters used for the molecular dynamics (MD) simulations are given. In Sect. 3, the results obtained from all-atom MD and coarse-grained MD simulations are presented and discussed. Section 4 contains the conclusions to the work.

2 Methods

2.1 Force-matching method

Coarse-grained potentials for β-D-glucose and water were derived using a multiscale coarse-graining approach based on force matching [11, 12]. A detailed description of this method can be found elsewhere [20–23]. Briefly, in the force-matching (FM) method, a coarse-grained pairwise effective force field is derived from the corresponding all-atom trajectory and force data obtained from a reference all-atom molecular dynamics simulation. For a given configuration from the reference all-atom simulation, the positions of N coarse-grained sites and the net forces $\mathbf{F}_i^{\mathrm{ref}}$ acting along them are computed. The force acting on the ith coarse-grained site due to the jth coarse-grained site is modeled as $\mathbf{f}_{ij}(\mathbf{r}_i, \mathbf{r}_j, p_1, p_2, \ldots, p_m)$ where $p_1, p_2, \ldots p_m$ are the m unknown parameters that need to be determined. Since the functional form for the force between coarse-grained sites is not known a priori, cubic splines are chosen to conveniently and systematically construct $\mathbf{f}_{ij}(\mathbf{r}_i, \mathbf{r}_j, p_1, p_2, \ldots, p_m)$ as a linear function of unknowns. Hence, the following system of N linear equations with m unknowns is obtained.

$$\sum_{j=1}^{N} \mathbf{f}_{ij}(r_i, r_j) = \mathbf{F}_i^{\mathrm{ref}}, \quad i = 1, 2, 3, \ldots, N \qquad (1)$$

The system of equations is overdetermined ($N > m$), and the values for the m unknown parameters are found using a singular value decomposition [24]. The parameters obtained in this way from each all-atom configuration are averaged over the total number of all-atom configurations sampled.

2.2 Coarse-grain mapping

Because a goal of this work is to evaluate coarse-grain mapping strategies, four different coarse-grain models for β-D-glucose were evaluated. The first two coarse-grained models consist of the same three coarse-grain sites as shown in Fig. 1 and two different definitions for the center of each coarse-grained site: center of mass (three-site CG-CM) and geometric center (three-site CG-GC). The 3-site coarse-grain mapping scheme is similar to coarse-grained models for glucose used previously [17, 18]. The advantage of this mapping is that it contains the minimum number of coarse-grain sites that preserve the anisotropic shape of glucose. The disadvantage of this mapping scheme is that it cannot represent the chair conformation that is the preferred conformation of β-D-glucose. The third and fourth coarse-grained models consist of six coarse-grain sites as shown in Fig. 2, again with two definitions for the center of each coarse-grained site: six-site CG-CM and six-site CG-GC. With this mapping scheme, the chair conformation

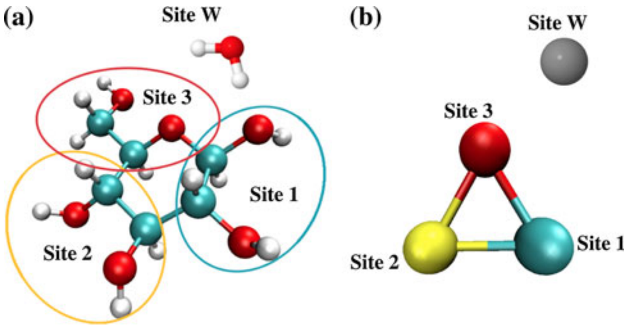


Fig. 1 **a** All-atom structures for β-D-glucose and water molecules; **b** three-site coarse-grained representation of β-D-glucose and single-site water molecules

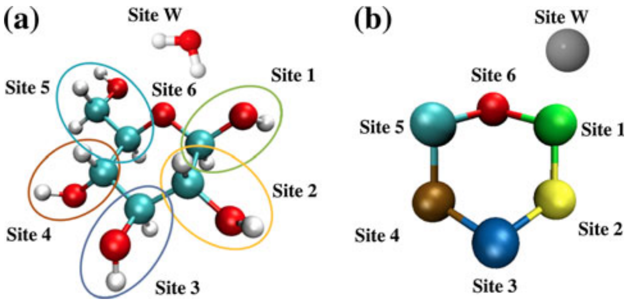


Fig. 2 **a** All-atom structures for β-D-glucose and water molecules; **b** six-site coarse-grained representation of β-D-glucose and single-site water molecules

of glucose can be sampled; however, this mapping doubles the number of coarse-grain sites and requires additional computational time. For all of the coarse-grained simulations, the water molecules were modeled with a single coarse-grained site that has been shown to accurately capture the liquid state structure of water [25]. The center site location of each coarse-grain water molecule was defined as either the center of mass or the geometric center in accordance with the particular coarse-grained model for β-D-glucose being used.

2.3 Molecular dynamics simulations

All MD simulations in this study were run using LAMMPS [26, 27]. The all-atom MD simulations used a CHARMM-style force field for glucose and the TIP3P [28] model for water. The CHARMM force field parameters for glucose were taken from parameters derived for carbohydrates [29]. The long-range electrostatic interactions were treated with the particle mesh Ewald (PME) method [30], and all hydrogen bond lengths were fixed by the SHAKE algorithm [31]. A Nose–Hoover thermostat was implemented to control the temperature at 300 K for all simulations [32]. In addition, the NVT ensemble is used for all simulations in this study. A mixture of 64 β-D-glucose molecules and 27,224 water molecules was placed in a cubic box with side length 9.6 nm and periodic boundary conditions. The system was equilibrated for 800 ps at $T = 300$ K, followed by 8 ns of simulation. A total of 4,000 configurations with positions, velocities, and forces were collected every 2 ps.

The coarse-grained MD simulations used the tabulated form of the coarse-grained force field as input to LAMMPS. Coarse-grained MD simulations were initiated from the same initial system as was used for the all-atom case. The coarse-grained system of β-D-glucose and water molecules was first equilibrated for 800 ps, followed by 8 ns of simulation.

3 Results and discussion

This section describes the results from the coarse-grained models discussed above for β-D-glucose in aqueous solution, beginning with an evaluation of the three-site coarse-grain model for β-D-glucose. Bond-length distributions and radial distribution functions from MD simulations using the

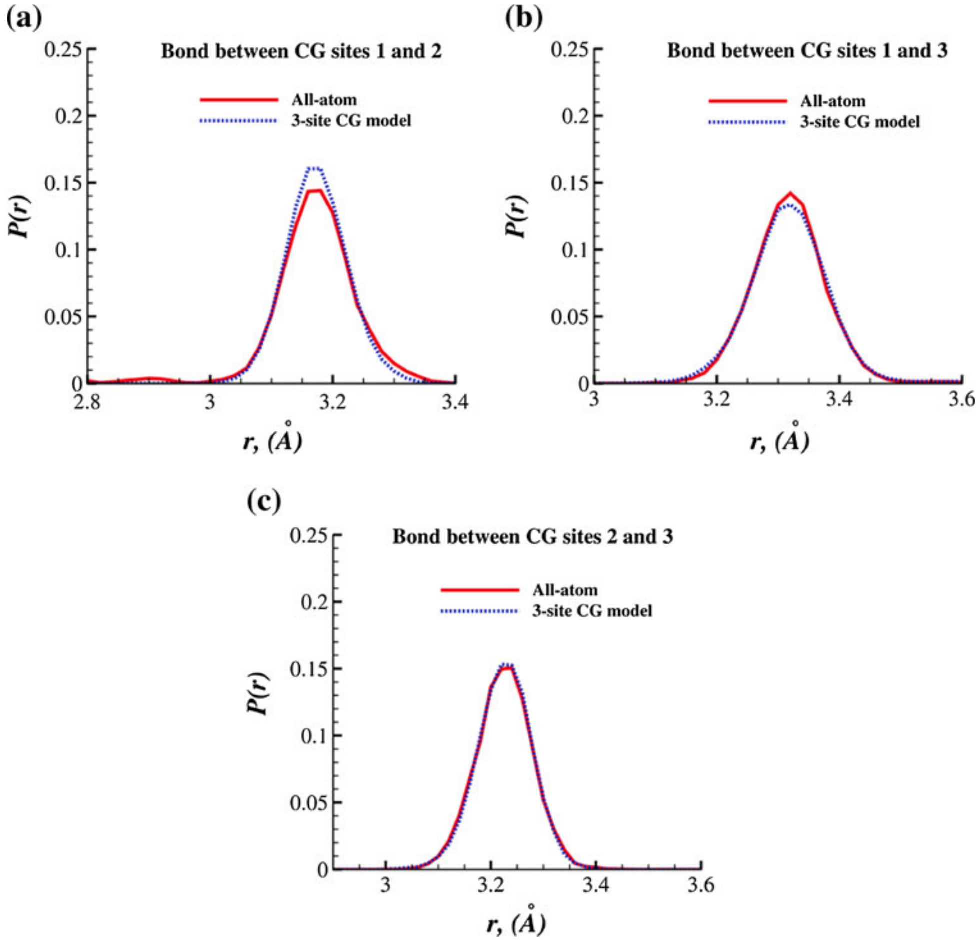


Fig. 3 The bond-length distribution $P(r)$ for the three-site CG-CM model for β-D-glucose: **a** bond between CG sites 1 and 2, **b** bond between CG sites 1 and 3, **c** bond between CG sites 2 and 3. The coarse-grain sites are defined as shown in Fig. 1, and the center of each coarse-grain site is located at the center of mass of the corresponding atoms. For comparison, the bond-length distribution between the coarse-grain sites taken directly from the reference all-atom MD simulation is shown

three-site coarse-grained β-D-glucose model are compared with MD simulation results obtained using an all-atom model. This discussion is followed by an analogous analysis of the six-site coarse-grained β-D-glucose model. When comparing bond-length distributions and radial distribution functions, the data are considered to be in "good agreement" if the peak positions differ by less than 0.1 Å, and the magnitude of peaks differs by not more than 30%. The data are then analyzed to demonstrate how the definition of the center of each coarse-grained site (center of mass vs. geometric center) changes the level of agreement for intramolecular and intermolecular structural properties between the coarse-grained simulation and the all-atom reference simulation. It is also shown how different degrees of coarsening affect the structural predictions derived from simulations with the coarse-grained model.

3.1 Evaluation of the three-site coarse-grained β-D-glucose model

Consider a coarse-grained model where each β-D-glucose molecule is represented by three coarse-grain sites and each water molecule is represented by one coarse-grain site, as shown in Fig. 1. This mapping yields the simplest coarse-grained model consisting of the minimum number of coarse-grain sites that can capture the planar structure and spatial anisotropy of the glucose molecule. With this three-site coarse-grained model for β-D-glucose, the aqueous solution can be fully described with 10 non-bonded pair interactions (six glucose–glucose, three glucose–water, one water–water) and three bonded interactions. When selecting a coarse-grain mapping scheme for an all-atom system, the center of the coarse-grain site must be specified. Two logical choices are the center of mass of all the atoms associated with the coarse-grain site (CG-CM) or the geometric center (CG-GC) of all the atoms associated with the coarse-grain site. Both cases are considered here to investigate how the center site location influences the outcome obtained from simulations using the coarse-grained model.

Figures 3 and 4 show results from MD simulations of aqueous β-D-glucose solutions and compare the three-site CG-CM model with the all-atom reference model. Figure 3 shows a comparison of bond-length distributions, $P(r)$,

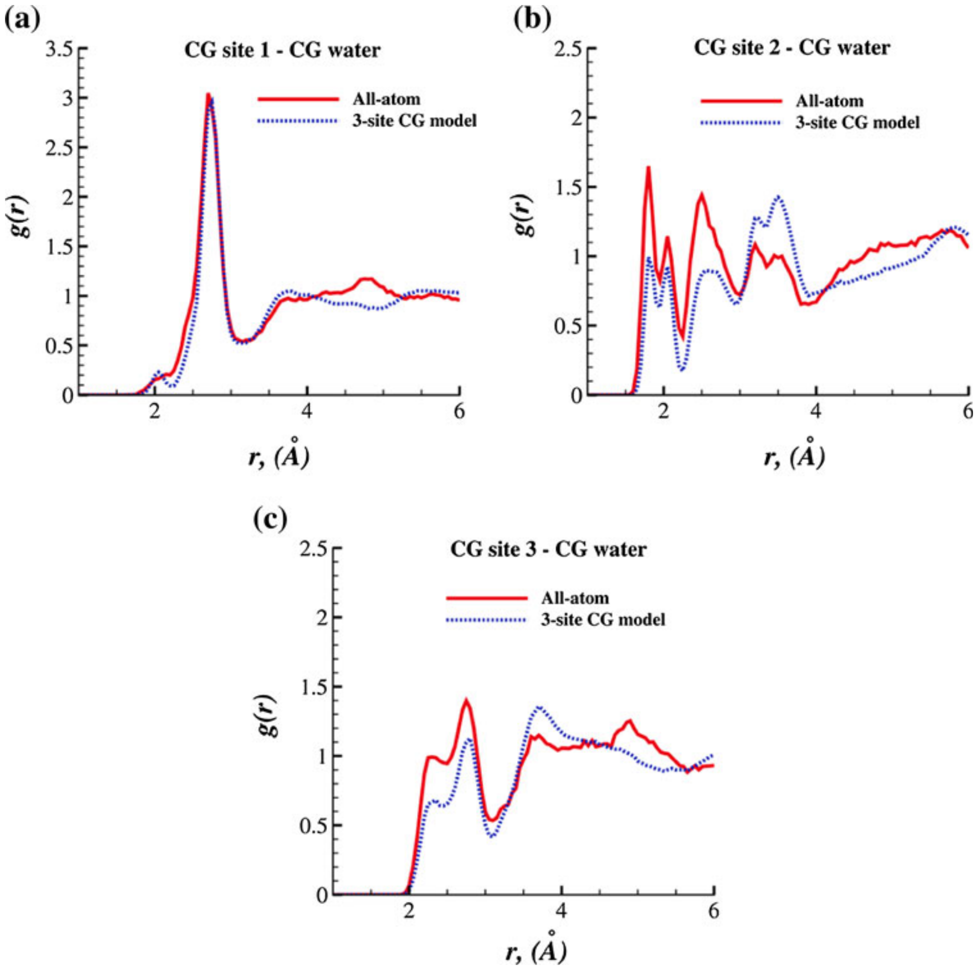


Fig. 4 The radial distribution function $g(r)$ for the three-site CG-CM model for β-D-glucose with the center of each coarse-grain site defined at the center of mass: **a** CG site 1-CG water, **b** CG site 2-CG water, **c** CG site 3-CG water. The coarse-grain sites are defined as shown in Fig. 1. For comparison, the radial distribution function taken directly from the reference all-atom MD simulation is shown

between the three coarse-grain β-D-glucose sites. Here, $P(r) = N(r)/N_{\mathrm{tot}}$ is the probability of finding a bond of length r, $N(r)$ is the number of bonds of length r, and N_{tot} is the total number of bonds. To make direct comparisons between the coarse-grained model and the all-atom model, the configurations from the all-atom MD simulation are first reduced to the three-site CG-CM mapping before calculating the bond-length distributions. Good agreement between the bond-length distributions from the three-site CG-CM MD simulation and the all-atom MD simulation is observed. For the bond between coarse-grain sites 1 and 2, the CG-CM model overestimates the maximum in the bond-length distribution by 15%. This discrepancy can be attributed to the fact that the three-site CG-CM model does not recover the minor peak at $r = 2.9$ Å (Fig. 3a) observed in the bond-length distribution derived from the all-atom simulation.

Figure 4 shows the radial distribution function, $g(r)$, for water around the three CG-CM β-D-glucose sites. The radial distribution function between β-D-glucose sites was not calculated because statistics for intermolecular interactions between β-D-glucose molecules are poor due to its low concentration in the system. As for the bond-length distribution case, the radial distributions functions for the all-atom system are obtained by reducing the configurations from the all-atom MD simulations to the three-site CG-CM mapping before calculating $g(r)$. Good agreement between the radial distribution functions from the three-site CG-CM simulations and the all-atom simulations is observed in the first neighbor peak for coarse-grain site 1. For coarse-grain sites 2 and 3, the three-site CG-CM model underestimates the magnitude of the peaks in $g(r)$ at short-range (<3 Å) distances and overestimates the midrange (3 Å $< r <$ 4 Å). However, the peak positions agree well with those derived from the reference all-atom simulation.

Now consider the three-site CG-GC model with the all-atom reference model. There is good agreement in the bond-length distributions obtained from the three-site CG-GC model and the all-atom model (Fig. 5). Figure 6 shows the radial distribution functions for water around the three CG-GC β-D-glucose sites. The three-site CG-GC model overestimates the magnitude of the first neighbor peak for all three coarse-grained sites. The first neighbor peak for coarse-grain site 1 is sharper and shifted about

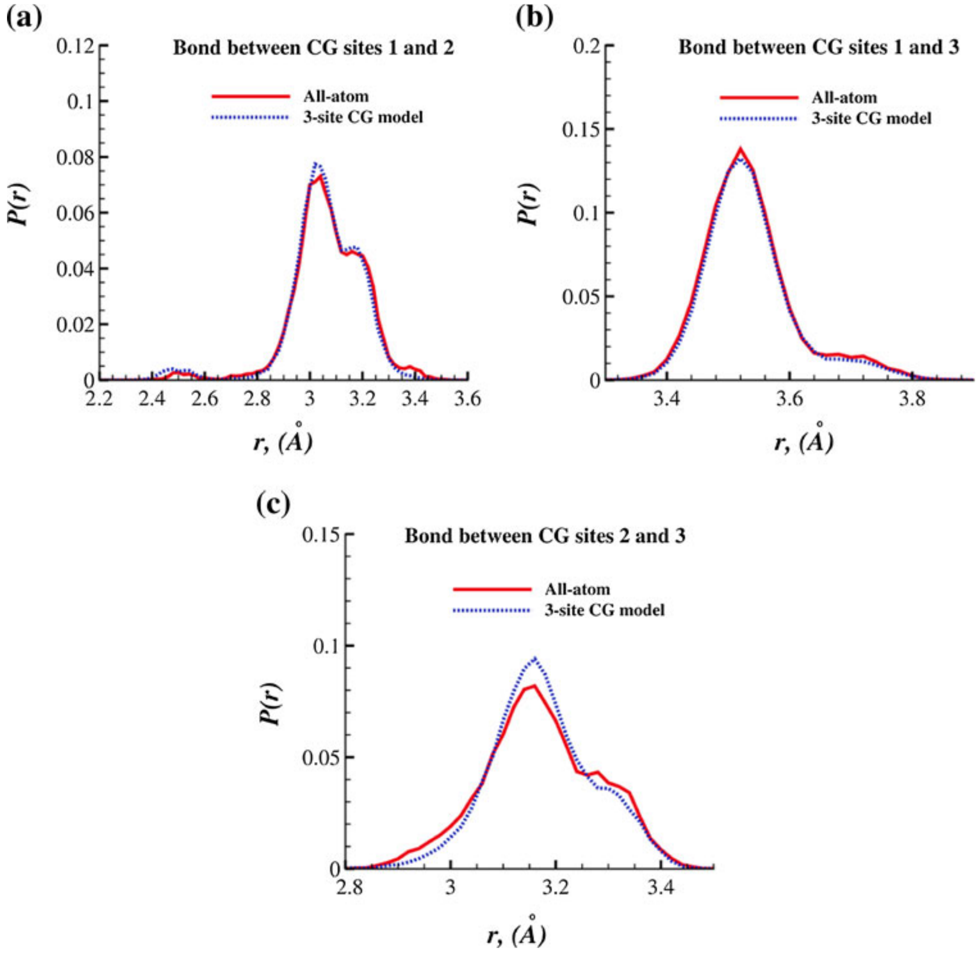


Fig. 5 The bond-length distribution $P(r)$ for the three-site CG-GC model for β-D-glucose: **a** bond between CG sites 1 and 2, **b** bond between CG sites 1 and 3, **c** bond between CG sites 2 and 3. The coarse-grain sites are defined as shown in Fig. 1, and the center of each coarse-grain site is located at the geometric center of the corresponding atoms. For comparison, the bond-length distribution between the coarse-grain sites taken directly from the reference all-atom MD simulation is shown

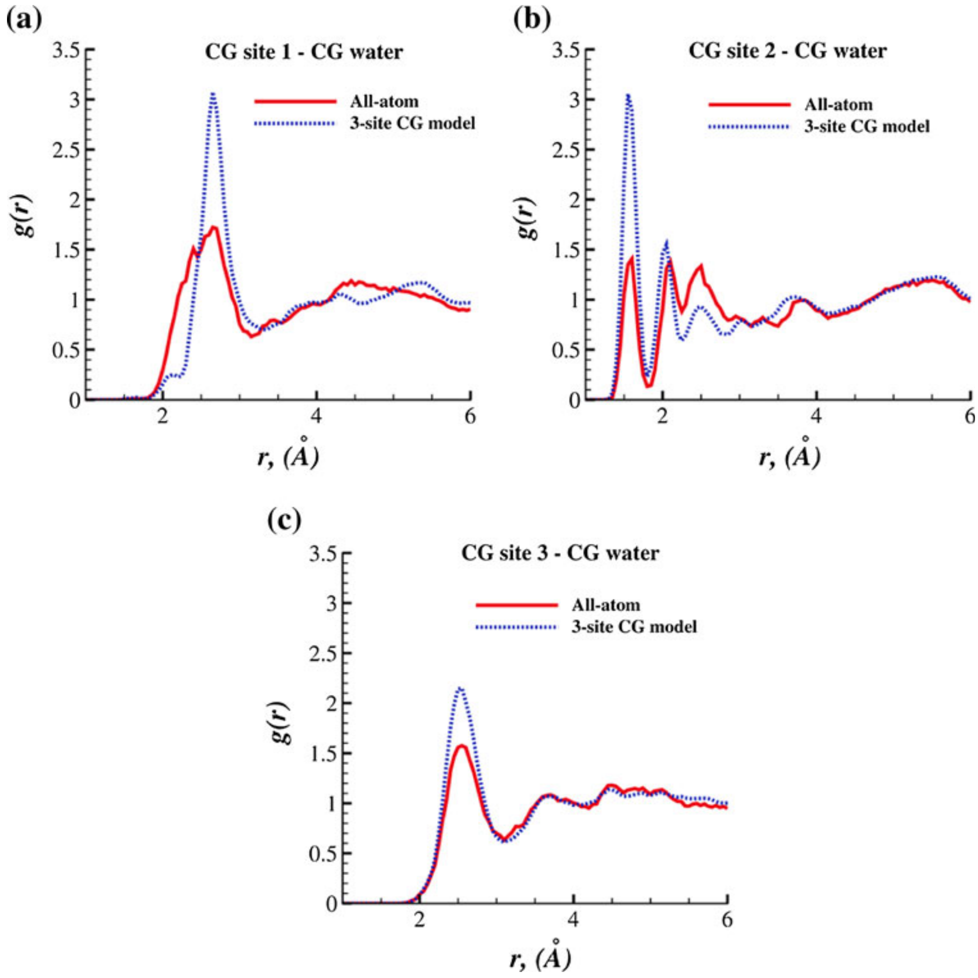


Fig. 6 The radial distribution function $g(r)$ for the three-site CG-GC model for β-D-glucose with the center of each coarse-grain site defined at the geometric center: **a** CG site 1-CG water, **b** CG site 2-CG water, **c** CG site 3-CG water. The coarse-grain sites are defined as shown in Fig. 1. For comparison, the radial distribution function between the coarse-grain sites taken directly from the reference all-atom MD simulation is shown

0.2 Å higher than the all-atom model, while the first neighbor peak positions for coarse-grain sites 2 and 3 do agree well with the all-atom model.

The results for the three-site coarse-grained models represented in the present work can be qualitatively compared with the three-site coarse-grained glucose models developed previously [17, 18]. Direct quantitative comparison is not possible since in these studies, different all-atom force fields are applied, different glucose conformers are used, and different concentrations of glucose are considered. However, qualitatively the three-site coarse-grained models represented in this paper show a good match in the radial distribution functions and bond distribution computed from the all-atom reference simulation and the coarse-grained simulation. In the present study, all coarse-grain interactions (non-bonded, bonds, and angles) are obtained from the coarse-graining procedure. The previously developed models [17, 18] used a hybrid scheme, whereby non-bonded interactions were obtained from force-matching method, and bonds, angles, and dihedrals were derived using Boltzmann inversion methods. Thus, the three-site CG models in this work are developed with fewer assumptions about the all-atom trajectory.

3.2 Evaluation of the six-site coarse-grained glucose model

Next, consider a coarse-grained model in which each β-D-glucose molecule is represented by six coarse-grain sites and each water molecule is represented by one coarse-grain site, as shown in Fig. 2. With this six-site coarse-grained model for β-D-glucose, the aqueous solution can be fully described with 28 non-bonded pair interactions, six bonded interactions, and three distinct dihedral angles. Figure 7 shows a comparison of bond-length distributions between the six CG-CM β-D-glucose sites. Figure 8 shows the radial distribution functions for water around the six CG-CM β-D-glucose sites. There is good agreement between the radial distribution functions obtained for the six-site CG-CM model and the all-atom model for the first neighbor peaks around coarse-grain sites 1–5. However, the CG-CM model does not recover the sharp secondary

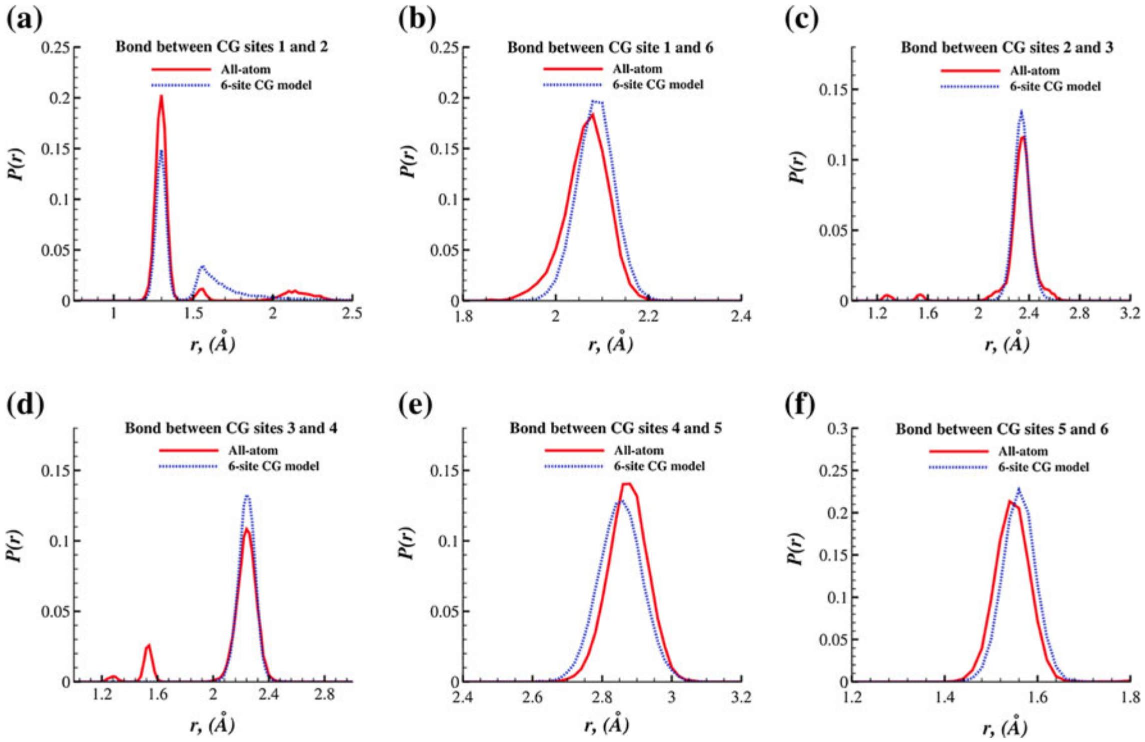


Fig. 7 The bond-length distribution $P(r)$ for the six-site CG-CM model for β-D-glucose: **a** bond between CG sites 1 and 2, **b** bond between CG sites 1 and 6, **c** bond between CG sites 2 and 3, **d** bond between CG sites 3 and 4, **e** bond between CG sites 4 and 5, **f** bond between CG sites 5 and 6. The coarse-grain sites are defined as shown in Fig. 2, and the center of each coarse-grain site is located at the center of mass of the corresponding atoms. For comparison, the bond-length distribution between the coarse-grain sites taken directly from the reference all-atom MD simulation is shown

peaks observed in the all-atom model for coarse-grain sites 1 and 2. Figure 9 shows a comparison of bond-length distributions between the six CG-GC β-D-glucose sites. Good agreement between the bond-length distributions from the six-site CG-GC MD simulation and the all-atom MD simulation is observed for bonds between coarse-grain sites 1–2, 1–6, 2–3, and 4–5. For the bond between coarse-grain sites 1–2, and 3–4, the CG-GC model overestimates the maximum in the bond-length distribution by 40 and 80%, respectively. As discussed above, this discrepancy can be attributed to the fact that the six-site CG-GC model does not recover the secondary peaks observed in the bond-length distribution derived from the all-atom simulation. Figure 10 shows the radial distribution functions for water around the six CG-GC β-D-glucose sites. There is good agreement between the six-site CG-GC model and the all-atom model for coarse-grain sites 2 and 6. The six-site CG-GC model overestimates the magnitude of the first neighbor peaks for coarse-grain sites 1, 3, 4, and 5, but the positions of first and secondary peaks are in close agreement with the all-atom model.

In Fig. 11, the distribution of conformations for the six-site coarse-grain model of β-D-glucose is compared to the all-atom reference data. The conformation distribution is defined by the angles ϕ, calculated between planes formed by coarse-grained sites 6–1–2 and 2–6–5, and ψ, calculated between planes formed by coarse-grained sites 3–4–5 and 3–5–6. The coarse-grained and the all-atom simulation data show occurrences of the chair conformation (where both $\phi, \psi < 180°$) and the boat conformation (where ϕ or $\psi > 180°$). For both CG-CM and CG-GC systems, the boat conformation is found in 43% of cases. In systems where the all-atom reference data are reduced to the 6-site CG models, about 57% of all conformations are boat conformations. However, when the conformation distributions are calculated for the β-D-glucose rings using all-atom data with no coarse-grain reduction, the boat conformation is found only in 20% cases (not shown). This increase in the presence of boat conformations can be explained by the fact that during the coarse-grain procedure, the positions of the coarse-grained sites in the β-D-glucose rings are significantly different from the carbon centers of the all-atom β-D-glucose rings. Thus, in the coarse-grained models, the occurrence of boat conformations is artificially increased, in this case by a factor of three.

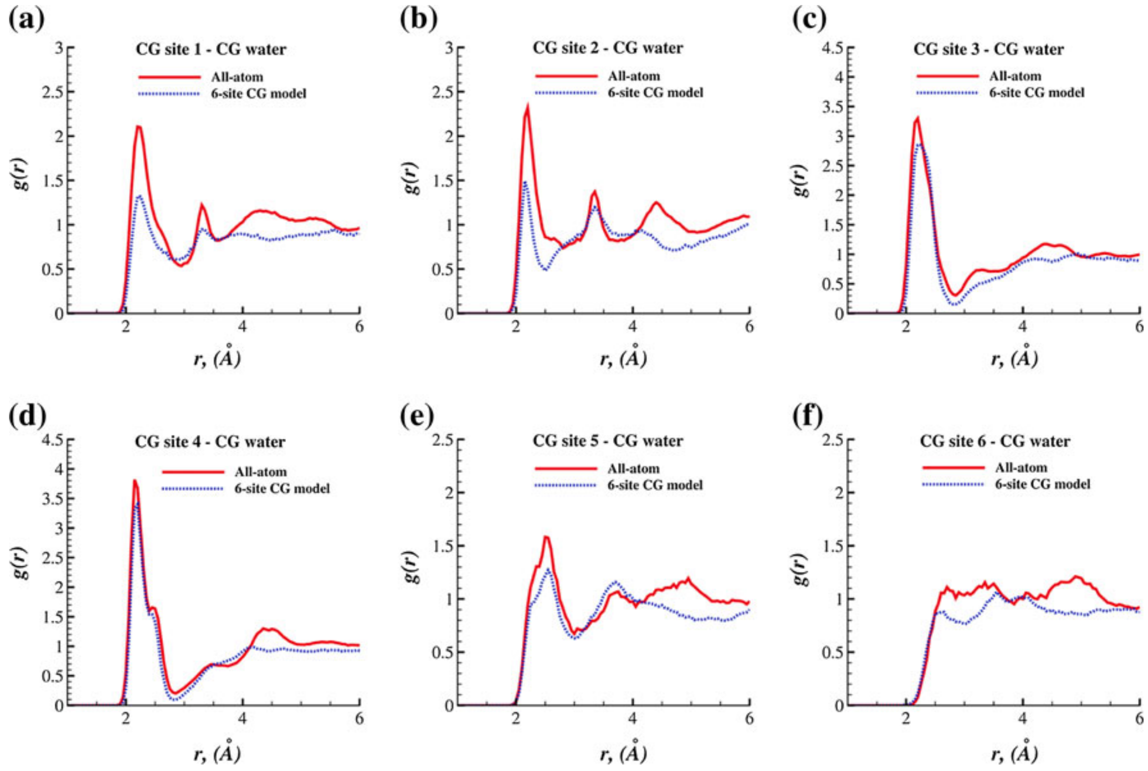


Fig. 8 The radial distribution function $g(r)$ for the six-site CG-CM model for β-D-glucose with the center of each coarse-grain site defined at the center of mass: **a** CG site 1-CG water, **b** CG site 2-CG water, **c** CG site 3-CG water, **d** CG site 4-CG water, **e** CG site 5-CG water, **f** CG site 6-CG water. The coarse-grain sites are defined as shown in Fig. 2. For comparison, the radial distribution function between the coarse-grain sites taken directly from the reference all-atom MD simulation is shown

3.3 Evaluation of the center site location

To facilitate direct comparison between center site locations, difference curves showing $\Delta g(r) = g(r)^{\mathrm{CG}} - g(r)^{\mathrm{AA}}$ for CG-CM and CG-GC are shown for the three-site and six-site models in Figs. 12 and 13, respectively. The CG-CM models reproduce the first neighbor peak in the radial distribution better than the CG-GC models, while the CG-GC models reproduce secondary peaks better than the CG-CM models. Exceptions are that the first neighbor peaks for coarse-grain sites 1 and 2 are better represented by the CG-GC model.

Such differences in the radial distribution functions suggests that defining the center site at different positions results in different effective interactions between coarse-grain sites. To more closely examine the effect that shifting the center site position has on the overall system properties, the difference between the center site for the two definitions (CM and GC) of coarse-grain site centers is calculated for β-D-glucose sites and for water. Table 1 shows the position difference, $R_{\mathrm{CM-GC}}$, between the CM and the GC sites calculated from the all-atom simulations. The distance between the CM and the GC center site position varies from 0.306 to 0.404 Å for most of the CG sites on glucose; this is comparable to the distance of 0.325 Å for water molecules. It can therefore be concluded that the difference between CG-CM and CG-GC models impacts both β-D-glucose coarse-grain sites and water coarse-grain sites. In the present study, a very dilute solution of β-D-glucose in aqueous solution is considered; therefore, it is difficult to estimate the degree to which the center site location impacts interactions between β-D-glucose molecules. However, the effect of the center site location on water molecules can be investigated in detail. Looking at the water molecule shown in Fig. 14, its center of mass is very close to the center of an oxygen atom (Fig. 14a). On the other hand, the geometric center of the water molecule is relatively far from the center of the oxygen atom and shifted toward the two hydrogen atoms (Fig. 14b). Recall that in this coarse-grain model of water, three atoms are replaced with a single spherical coarse-grain site. Thus, the CG-CM model does not take into account the presence of two hydrogen atoms (in a topological sense) since the center of the coarse-grain site is very close to the center of the oxygen atom (Fig. 14a). With this model, two CG-CM water molecules approaching each other would interact as

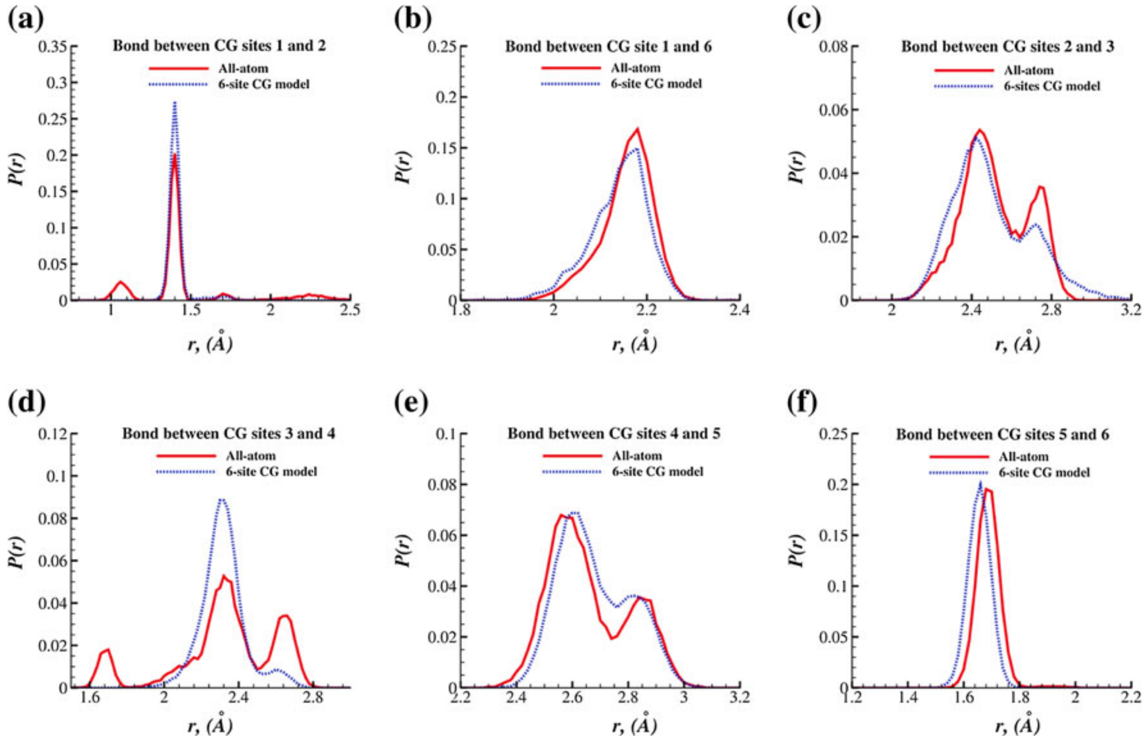


Fig. 9 The bond-length distribution $P(r)$ for the six-site CG-GC model for β-D-glucose: **a** bond between CG sites 1 and 2, **b** bond between CG sites 1 and 6, **c** bond between CG sites 2 and 3, **d** bond between CG sites 3 and 4, **e** bond between CG sites 4 and 5, **f** bond between CG sites 5 and 6. The coarse-grain sites are defined as shown in Fig. 2, and the center of each coarse-grain site is located at the geometric center of the corresponding atoms. For comparison, the bond-length distribution between the coarse-grain sites taken directly from the reference all-atom MD simulation is shown

hard spheres due to the strong repulsion between oxygen atoms. In the CG-GC model (Fig. 14b), the presence of the two hydrogen atoms is taken into account. As two CG-GC water molecules are brought together, they approach from the hydrogen atom sites. As such, this interaction will be more like an elastic sphere interaction.

These arguments are fully supported by the effective force data for the CG water–CG water interaction obtained from the coarse-graining procedure as shown in Fig. 15a. The CG-CM model predicts strong repulsion between CG water molecules for separations that are <2.8 Å—a signature of hard sphere-like behavior. The CG-GC model predicts a shoulder with moderate repulsion at separations between 2.1 and 2.8 Å that can be interpreted as an elastic sphere interaction. Similar behavior is observed for β-D-glucose CG sites–CG water interactions (Fig. 15b). The radial distribution function between the CG water–CG water molecules for the CG-CM and CG-GC water models in comparison to the all-atom reference system are presented in Fig. 16. The positions of the first peak are well-matched for both models and located at 2.8 Å, which is in good agreement with the mean van der Waals diameter of water (2.82 Å) [33]. However, the CG-CM water model is not able to represent secondary peaks, while the CG-GC water model gives a closer match for the secondary peaks (Fig. 16). This explains why at distances beyond the first neighbor peak, the CG-GC models reproduce the secondary peaks better than the CG-CM models (Figs. 12, 13). Likewise, for systems of glucose and water, the effective elastic sphere interactions that result from coarse-graining to the GC is that the number of nearest neighbor coarse-grained water molecules around a given coarse-grained glucose site increases, thereby overestimating the first neighbor peak in the radial distribution compared to the all-atom reference system (Figs. 12, 13).

Based on this analysis, the CG-GC model for water molecules more accurately represents the overall water–water spatial distribution compared to the CG-CM model. Great attention must be given to the development of an accurate coarse-grained water model. Because water molecules are non-symmetrical, a coarse-grained model that is built based on statistical averaging might not be generally applicable to all systems or even to all state points for a given system. When accurate and reliable coarse-grained structures are required, such as when the intended use of the coarse-grained model is to efficiently generate equilibrium configurations and then restore the all-atom degrees of freedom (e.g., reverse-mapping), it may be

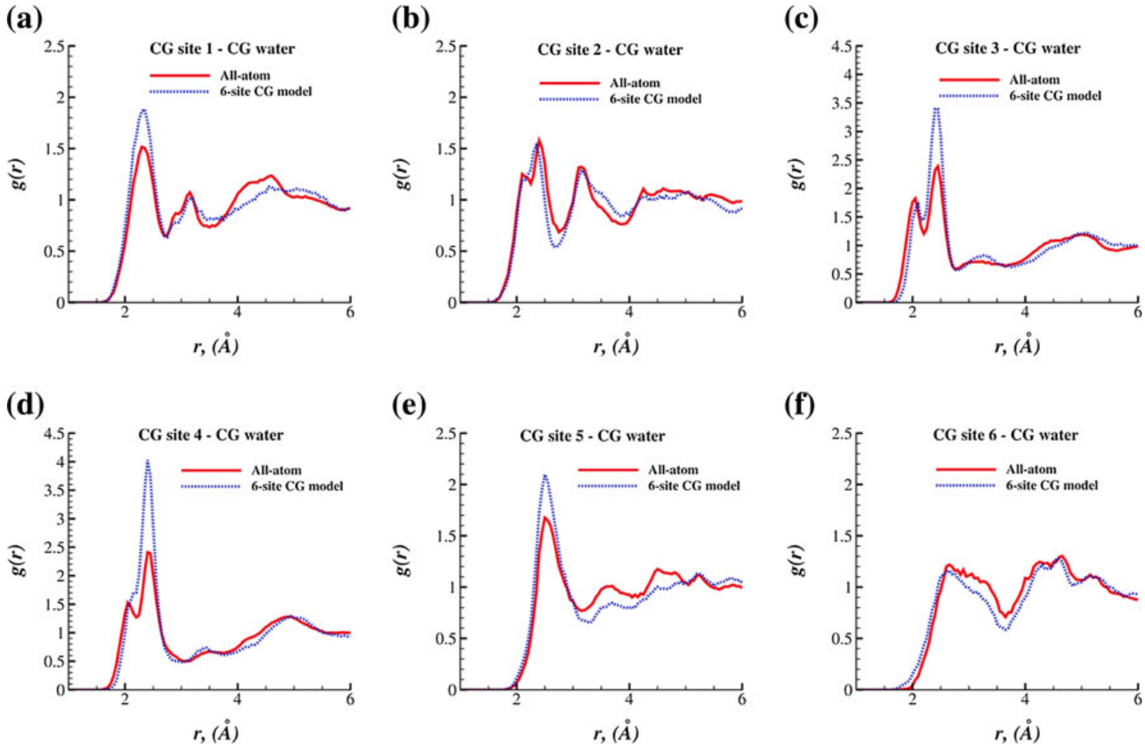
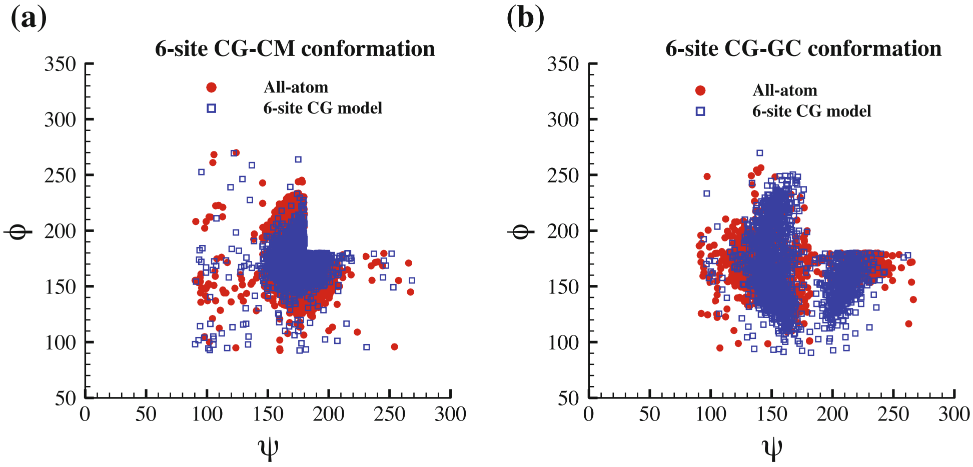


Fig. 10 The radial distribution function $g(r)$ for the six-site CG-GC model for β-D-glucose with the center of each coarse-grain site defined at the geometric center: **a** CG site 1-CG water, **b** CG site 2-CG water, **c** CG site 3-CG water, **d** CG site 4-CG water, **e** CG site 5-CG water, **f** CG site 6-CG water. The coarse-grain sites are defined as shown in Fig. 2. For comparison, the radial distribution function between the coarse-grain sites taken directly from the reference all-atom MD simulation is shown

Fig. 11 Conformation of 6-site CG molecules represented by the angle ϕ calculated between planes formed by CG sites 6–1–2 and 2–6–5, and the angle ψ calculated between planes formed by CG sites 3–4–5 and 3–5–6. **a** data for CG-CM model, **b** data for CG-GC model. Blue dots represent data calculated for CG system. For comparison, the angle distribution data taken directly from the reference all-atom MD simulation is shown



prudent to develop a coarse-grained water molecule model along with the solute molecule, rather than using a generalized coarse-grained water model.

3.4 Evaluation of the coarse-grain mapping

Comparing Fig. 4 to Fig. 8 (CG-CM) and Fig. 6 to Fig. 10 (CG-GC) shows how the mapping scheme affects the accuracy of the coarse-grain model in predicting the structure observed in the all-atom reference simulations. A three-site CG model (Fig. 1) is fully determined when non-bonded interactions, bonds, and angles are defined. A six-site CG model (Fig. 2) is much more complex since the number of unique interaction pairs is significantly larger. Moreover, a dihedral interaction must be included to obtain an accurate coarse-grained model.

Despite its relative simplicity, a three-site CG model is able to represent the structural properties of the system

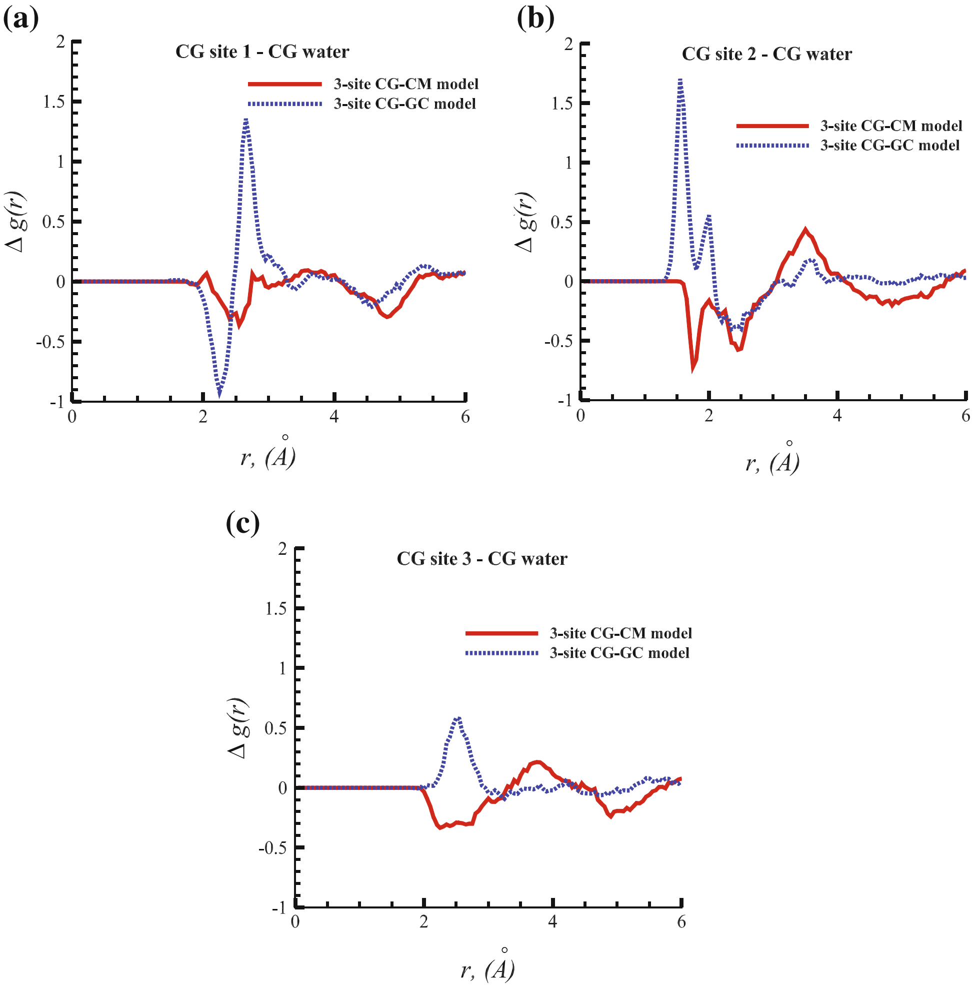


Fig. 12 The differences of the radial distribution functions calculated between the radial distribution function for the three-site CG-CM and CG-GC models and for the three-site all-atom MD simulation for β-D-glucose. The *positive values* represent overestimation of $g(r)$ by the CG model, and the *negative values* represent underestimation of $g(r)$ by the CG model. **a** CG site 1-CG water, **b** CG site 2-CG water, **c** CG site 3-CG water. The coarse-grain sites are defined as shown in Fig. 1

predicted in the all-atom reference simulations well (Figs. 3, 4, 5, 6). The positions of the primary peaks in the radial distribution functions are accurately matched (Figs. 4, 6), while the magnitude of the primary peaks are better predicted by the CG-CM model (Fig. 4). Bond-length distributions show a perfect match with all-atom reference data for both CG-CM and CG-GC models (Figs. 3, 5, respectively). A six-site CG model is able to accurately capture the structural properties of β-D-glucose in a water environment (Figs. 7, 8, 9, 10). The positions and the magnitudes of the primary peaks in the radial distribution function are accurately predicted (Figs. 8, 10). However, the magnitudes for site 1 and site 2 for the CG-CM model are underestimated (Fig. 8), and the magnitudes for site 3 and site 4 for the CG-GC model are overestimated (Fig. 10). The positions and the magnitudes of the secondary peaks are better described by the CG-GC model (Fig. 10) due to the more accurate CG model of water molecules (Fig. 16b). Bond-length distributions for the six-site CG models are in good agreement with the all-atom reference data (Figs. 7, 9). This match is not as perfect as for the three-site CG models (Figs. 3, 5) especially for the CG-GC model where bond-length distributions have secondary peaks (Fig. 9). However, by taking into account the increased complexity of a six-site CG model, the positions and magnitude of primary and secondary peaks match reasonably well (Figs. 7, 9).

Based on this comparison, it is possible to say that both the three-site CG models and the six-site CG models can be successfully used to accurately predict the structure observed in all-atom reference simulations. If not many details are required for the CG model then a simple three-site CG-CM model can be used. In cases, when more structural details are needed then a more complex six-site CG-GC model should be used.

4 Conclusions

In this work, coarse-grained models of β-D-glucose in water solution were developed using the force-matching approach. For this system, two different coarse-grained

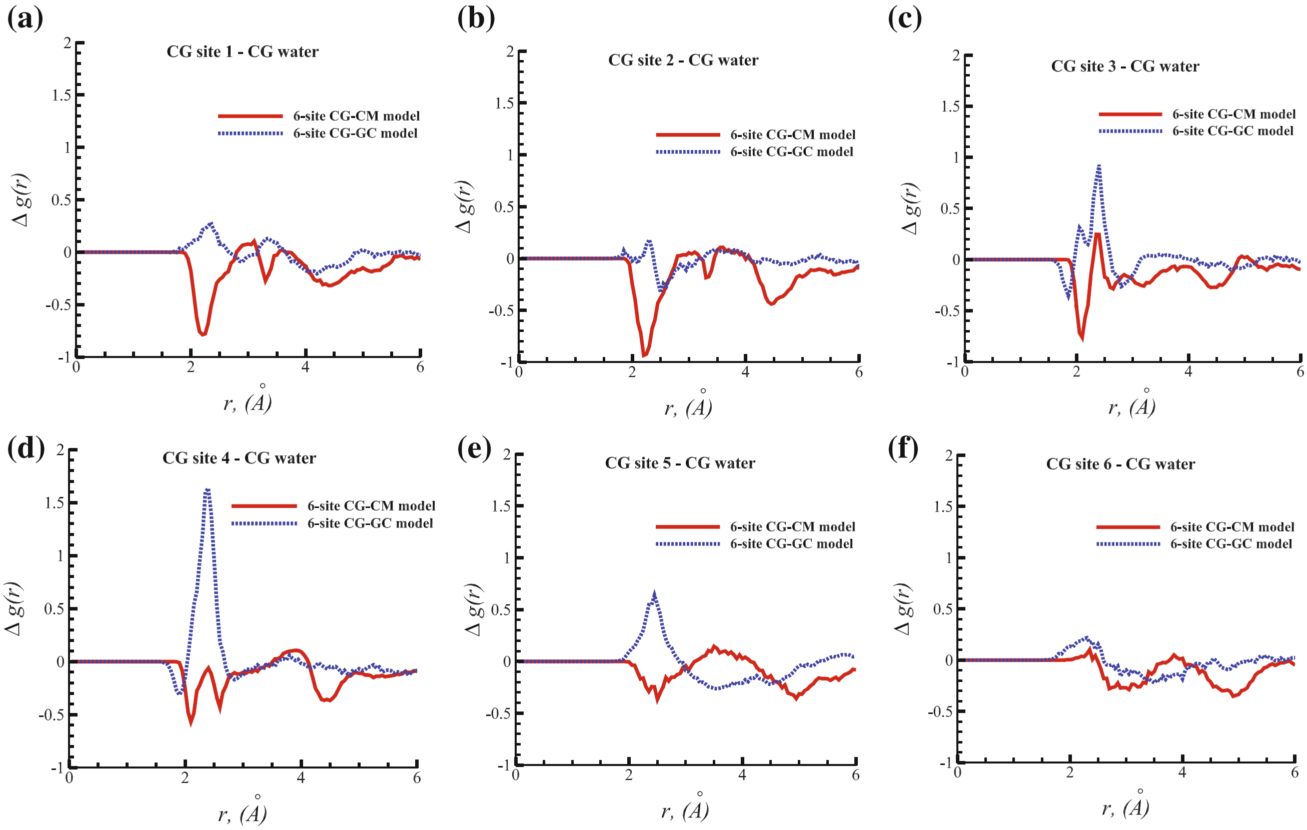


Fig. 13 The differences of the radial distribution functions calculated between the radial distribution function for the six-site CG-CM and CG-GC models and for the six-site all-atom MD simulation for β-D-glucose. The positive values represent overestimation of $g(r)$ by the CG model, and the negative values represent underestimation of $g(r)$ by the CG model. **a** CG site 1-CG water, **b** CG site 2-CG water, **c** CG site 3-CG water, **d** CG site 4-CG water, **e** CG site 5-CG water, **f** CG site 6-CG water. The coarse-grain sites are defined as shown in Fig. 2

Table 1 The distance between the center of mass and geometric center calculated based on all-atom MD simulation results and correspondent standard deviation for all models used in the present study

	$R_{\text{CM-GC}}$, Å	Stdv., Å
3-site model		
1	0.108	0.025
2	0.132	0.052
3	0.404	0.013
6-site model		
1	0.378	0.053
2	0.378	0.060
3	0.365	0.073
4	0.349	0.092
5	0.306	0.020
6	0.000	0.000
1-site water		
1	0.325	0.023

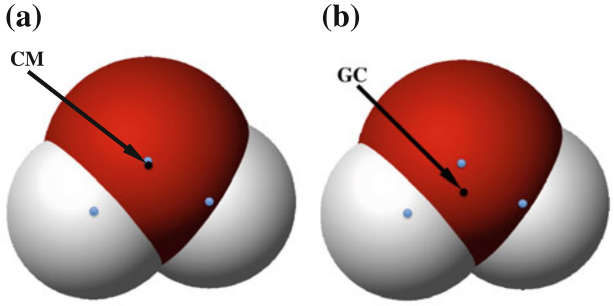


Fig. 14 Water molecule representation. *Red sphere* represents oxygen atom, *white spheres* represent hydrogen atoms. *Blue dots* represent the center of each atom in the molecule. *Black dots* represent **a** center of mass of the water molecule, **b** geometric center of the water molecule

mapping schemes (three-site and six-site) were investigated. In addition to varying the number of atoms in a coarse-grained site, two definitions of the center site location for the coarse-grained sites, center of mass and geometric center, were considered and evaluated in detail.

The CG-CM models generally reproduce the primary peak in the radial distribution function better than the CG-GC models. However, the CG-GC models accurately reproduce secondary peaks in the radial distribution

Fig. 15 Effective force–distance curves obtained with coarse-graining method. *Red solid lines* correspond to the CG-CM model and *blue dotted lines* correspond to the CG-GC model. **a** force–distance data for CG water–water interaction, **b** force–distance data for CG site 3–CG water molecules in 3-site CG β-D-glucose model

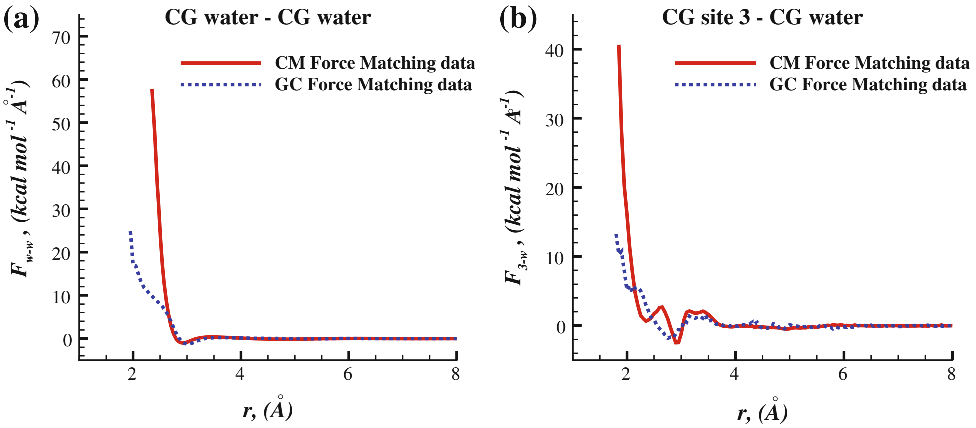
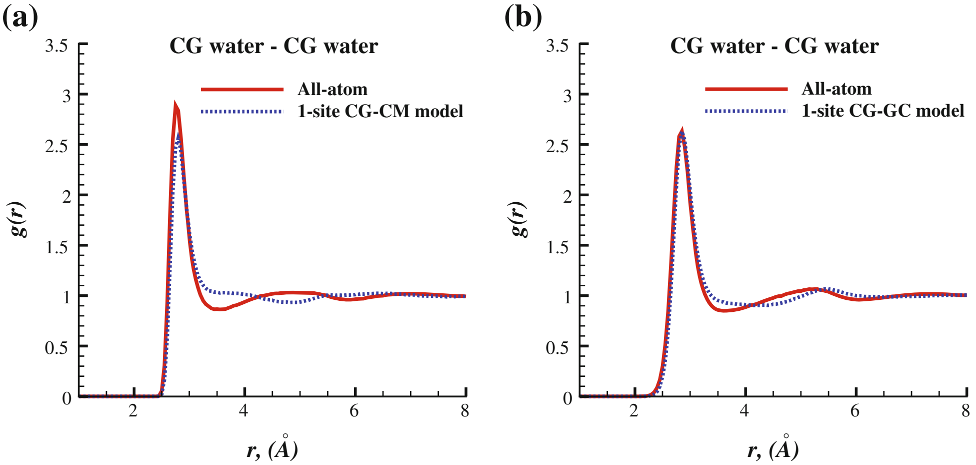


Fig. 16 The radial distribution function $g(r)$ for the water model **a** the center of coarse-grain site defined at the center of mass, **b** the center of coarse-grain site is defined at the geometric center. For comparison, the radial distribution function between the coarse-grain sites taken directly from the reference all-atom MD simulation is shown



functions. This difference in the RDFs for CG-CM and CG-GC models can be explained by the different interactions captured by the two definitions of center site. The coarse-grained sites in the CG-CM water molecules interact like hard spheres, while the CG-GC water molecules interact like soft-spheres. Thus, the choice of using CG-CM or CG-GC depends on whether the desired outcome of the coarse-grained model is to predict local or long-range structure more accurately.

Three-site coarse-grained models reproduce structural properties like radial distribution functions and bond-length distributions observed in all-atom reference simulations reasonably well, although there are some cases where primary peaks are significantly overestimated. For β-D-glucose in water, the three-site CG-CM model gives better predictions than the three-site CG-GC model. The three-site CG-CM model is suitable for situations in which a simple and fast coarse-grained model capable of generating long length scale equilibrium structures is favored over accuracy in local ordering.

The six-site CG models reveal a higher level of structural detail than the three-site CG models. For β-D-glucose in water, the six-site CG-GC model is a more accurate model than the CG-CM model for most of primary peaks and it yields a nearly perfect match for the secondary peaks in the radial distribution function.

The detailed analysis of coarse-grain modeling shown here underscores the importance of evaluating multiple coarse-grain mapping schemes for a given application. For example, the β-D-glucose molecules considered in the present work are the building blocks for cellulose molecules, and thus, the insight gained about coarse-grained model development for glucose is a critical first step before proceeding to the development of coarse-grain models for the more complicated crystalline cellulose structure.

Acknowledgments This research is sponsored by U.S. Department of Energy's (USDOE) Scientific Discovery through Advanced Computing (SciDAC) program through USDOE's Office of Advanced Scientific Computing Research (ASCR) and Biological and Environmental Research (BER), and performed at the Ames Laboratory, FWP AL-08-330-039. Ames Laboratory is managed by Iowa State University for the U.S. Department of Energy under contract DE-AC02-07CH11358. The authors thank Professor G. A. Voth for providing the multiscale coarse-graining (MS-CG) software developed by his research group.

References

1. Rubin EM (2008) Genomics of cellulosic biofuels. Nature 454:841–845
2. Sanderson K (2011) Lignocellulose: a chewy problem. Nature 474:S12–S14
3. Marrink SJ, Risselada HJ, Yefimov S, Tieleman DP, de Vries AH (2007) The MARTINI force field: coarse grained model for biomolecular simulations. J Phys Chem B 111:7812–7824
4. Monticelli L, Kandasamy SK, Periole X, Larson GR, Tieleman DP, Marrink SJ (2008) The MARTINI coarse-grained force field: extension to proteins. J Chem Theory Comput 4:819–834
5. Ashbaugh HS, Patel HA, Kumar SK, Garde S (2005) Mesoscale model of polymer melt structure: self-consistent mapping of molecular correlations to coarse-grained potentials. J Chem Phys 122:104908
6. Reith D, Pütz M, Müller-Plathe F (2003) Deriving effective mesoscale potentials from atomistic simulations. J Comput Chem 24:1624
7. Muller-Plathe F (2002) Coarse-graining in polymer simulation: from the atomistic to the mesoscopic scale and back. Chem Phys Chem 3:754–769
8. Peter C, Delle Site L, Kremer K (2008) Classical simulations from the atomistic to the mesoscale and back: coarse graining an azobenzene liquid crystal. Soft Matter 4:859–869
9. Lyubartsev AP, Laaksonen A (1995) Calculation of effective interaction potentials from radial distribution functions: a reverse Monte Carlo approach. Phys Rev E 52:3730–3737
10. Ercolessi F, Adams JB (1994) Interatomic potentials from 1st-principles calculations- the force-matching method. Europhys Lett 26:583–588
11. Izvekov S, Voth GA (2005) A multiscale coarse-graining method for biomolecular systems. J Phys Chem B 109:2469–2473
12. Izvekov S, Parrinello M, Burnham CJ, Voth GA (2004) Effective force fields for condensed phase systems from ab initio molecular dynamics simulation: a new method for force-matching. J Chem Phys 120:10896–10913
13. Baron R, de Vries AH, Hunenberger PH, van Gunsteren WF (2006) Comparison of atomic-level and coarse-grained models for liquid hydrocarbons from molecular dynamics configurational entropy estimates. J Phys Chem B 110:8464–8473
14. Kim SH, Lamm MH (2011) Reintroducing explicit solvent to a solvent-free coarse-grained model. Phys Rev E 84:025701
15. Molinero V, Goddard WA (2004) M3B: a coarse grain force field for molecular simulations of malto-oligosaccharides and their water mixtures. J Phys Chem B 108:1414–1427
16. Molinero V, Goddard WA (2005) Microscopic mechanism of water diffusion in glucose glasses. Phys Rev Lett 95:04701
17. Liu P, Izvekov S, Voth GA (2007) Multiscale coarse-graining of monosaccharides. J Phys Chem B 111:11566–11575
18. Hynninen AP, Matthews JF, Beckham GT, Crowley MF, Nimlos MR (2011) Coarse-grain model for glucose, cellobiose, and cellotetraose in water. J Chem Theory Comput 7:2137–2150
19. Srinivas G, Cheng X, Smith JC (2012) A solvent-free coarse grain model for crystalline and amorphous cellulose fibrils. J Chem Theory Comput 7(8):2539–2548
20. Hills Jr R, Lu L, Voth GA (2009) Multiscale coarse-graining of the protein energy landscape. J Phys Chem B 133:4443
21. Thorpe IF, Zhou J, Voth GA (2008) Peptide folding using multiscale coarse-grained model. J Phys Chem B 112:13079–13090
22. Noid WG, Chu JW, Ayton GS, Krishna V, Izvekov S, Voth GA, Das A, Andersen HC (2008) The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. J Chem Phys 128(24):244114
23. Noid WG, Chu JW, Ayton GS, Krishna V, Izvekov S, Voth GA, Das A, Andersen HC (2008) The multiscale coarse-graining method. II. A rigorous bridge between atomistic and coarse-grained models. J Chem Phys 128(24):244115
24. Lawson CL, Hanson RJ (1974) Solving least squares problems. Prentice-Hall, Englewood Cliffs
25. Izvekov S, Voth GA (2005) Multiscale coarse graining of liquid-state systems. J Chem Phys 123:134105
26. http://lammps.sandia.gov
27. Plimpton SJ (1995) Fast parallel algorithms for short-range molecular dynamics. J Comp Phys 117:1–19
28. Jorgensen WL, Chandrasekhar J, Madura JD (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79:926–935
29. Palma R, Zuccato P, Himmel M, Liang G, Brady JW (2001) In: Himmel ME (ed) Glycosyl hydrolases for biomass conversion, ACS symposium series. American Chemical Society, Washington, DC, pp 112–130
30. Darden T, York D, Pedersen LG (1993) Particle mesh Ewald: an Nlog(N) method for Ewald sums in large systems. J Chem Phys 98:10089–10092
31. Ryckaert JP, Ciccotti G, Berendsen HJC (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J Comput Phys 23:327–341
32. Nose S (1984) A molecular dynamics method for simulations in the canonical ensemble. Mol Phys 52:255
33. Franks F (2000) Water: 2nd edition a matrix of life. Royal Society of Chemistry, Cambridge