Evaluation of Coarse-Grained Mapping Schemes for Polysaccharide Chains in Cellulose

Sergiy Markutsya
Iowa State University

Ajitha Devarajan
Iowa State University

John Y. Baluyut
Iowa State University

Theresa Lynn Windus
Iowa State University, twindus@iastate.edu

Mark S. Gordon
Iowa State University, mgordon@iastate.edu

See next page for additional authors

Follow this and additional works at: http://lib.dr.iastate.edu/ameslab_pubs

Part of the Chemistry Commons

The complete bibliographic information for this item can be found at http://lib.dr.iastate.edu/ameslab_pubs/353. For information on how to cite this item, please visit http://lib.dr.iastate.edu/howtocite.html.
Evaluation of Coarse-Grained Mapping Schemes for Polysaccharide Chains in Cellulose

Abstract
A fundamental understanding of the intermolecular forces that bind polysaccharide chains together in cellulose is crucial for designing efficient methods to overcome the recalcitrance of lignocellulosic biomass to hydrolysis. Because the characteristic time and length scales for the degradation of cellulose by enzymatic hydrolysis or chemical pretreatment span orders of magnitude, it is important to closely integrate the molecular models used at each scale so that, ultimately, one may switch seamlessly between quantum, atomistic, and coarse-grained descriptions of the system. As a step towards that goal, four multiscale coarse-grained models for polysaccharide chains in a cellulose-1α microfiber are considered. Using the force matching method, effective coarse-grained forces are derived from all-atom trajectories. Performance of the coarse-grained models is evaluated by comparing the intrachain radial distribution functions with those obtained using the all-atom reference data. The all-atom simulation reveals a double peak in the radial distribution function for sites within each glucose residue that arises from the distinct conformations sampled by the primary alcohol group in the glucose residues. The three-site and four-site coarse-grained models have sufficient degrees of freedom to predict this double peak while the one-site and two-site models do not. This is the first time that coarse-grained models have been shown to reproduce such subtle, yet important, molecular features in a polysaccharide chain. The relative orientations between glucose residues along the polysaccharide chain are evaluated and it is found that the four-site coarse-grained model is best at reproducing the glucose-glucose conformations observed in the all-atom simulation. The success of the four-site coarse-grained model underscores the importance of decoupling the pyranose ring from the oxygen atom in the glycosidic bond when developing all-atom to coarse-grained mapping schemes for polysaccharides.

Keywords
Crystal structure, Particle distribution functions, Trajectory models, Molecular dynamics, Multiscale methods

Disciplines
Chemistry

Comments
The following article appeared in *Journal of Chemical Physics* 138 (2013): 214108, and may be found at doi:10.1063/1.4808025.

Rights
Copyright 2013 American Institute of Physics. This article may be downloaded for personal use only. Any other use requires prior permission of the author and the American Institute of Physics.

Authors
Sergiy Markutsya, Ajitha Devarajan, John Y. Baluyut, Theresa Lynn Windus, Mark S. Gordon, and Monica H. Lamm

This article is available at Iowa State University Digital Repository: http://lib.dr.iastate.edu/ameslab_pubs/353
Evaluation of coarse-grained mapping schemes for polysaccharide chains in cellulose
Sergiy Markutsya, Ajitha Devarajan, John Y. Baluyut, Theresa L. Windus, Mark S. Gordon, and Monica H. Lamm

Citation: The Journal of Chemical Physics 138, 214108 (2013); doi: 10.1063/1.4808025
View online: http://dx.doi.org/10.1063/1.4808025
View Table of Contents: http://scitation.aip.org/content/aip/journal/jcp/138/21?ver=pdfcov
Published by the AIP Publishing

Articles you may be interested in
A coarse-grained model for polyethylene glycol polymer

Gay-Berne and electrostatic multipole based coarse-grain potential in implicit solvent

Statistical temperature molecular dynamics: Application to coarse-grained β-barrel-forming protein models

Structure and stability of chiral β-tapes: A computational coarse-grained approach

A coarse-grained, “realistic” model for protein folding
Evaluation of coarse-grained mapping schemes for polysaccharide chains in cellulose

Sergiy Markutsya,1,2, a) Ajitha Devarajan,1,3, a) John Y. Baluyut,1 Theresa L. Windus,1,3 Mark S. Gordon,1,3 and Monica H. Lamm1,2, b)

1Ames Laboratory, Iowa State University, Ames, Iowa 50011, USA
2Department of Chemical and Biological Engineering, Iowa State University, Ames, Iowa 50011, USA
3Department of Chemistry, Iowa State University, Ames, Iowa 50011, USA

(Received 4 February 2013; accepted 10 May 2013; published online 6 June 2013)

A fundamental understanding of the intermolecular forces that bind polysaccharide chains together in cellulose is crucial for designing efficient methods to overcome the recalcitrance of lignocellulosic biomass to hydrolysis. Because the characteristic time and length scales for the degradation of cellulose by enzymatic hydrolysis or chemical pretreatment span orders of magnitude, it is important to closely integrate the molecular models used at each scale so that, ultimately, one may switch seamlessly between quantum, atomistic, and coarse-grained descriptions of the system. As a step towards that goal, four multiscale coarse-grained models for polysaccharide chains in a cellulose microfiber are considered. Using the force matching method, effective coarse-grained forces are derived from all-atom trajectories. Performance of the coarse-grained models is evaluated by comparing the intrachain radial distribution functions with those obtained using the all-atom reference data. The all-atom simulation reveals a double peak in the radial distribution function for sites within each glucose residue that arises from the distinct conformations sampled by the primary alcohol group in the glucose residues. The three-site and four-site coarse-grained models have sufficient degrees of freedom to predict this double peak while the one-site and two-site models do not. This is the first time that coarse-grained models have been shown to reproduce such subtle, yet important, molecular features in a polysaccharide chain. The relative orientations between glucose residues along the polysaccharide chain are evaluated and it is found that the four-site coarse-grained model is best at reproducing the glucose-glucose conformations observed in the all-atom simulation. The success of the four-site coarse-grained model underscores the importance of decoupling the pyranose ring from the oxygen atom in the glycosidic bond when developing all-atom to coarse-grained mapping schemes for polysaccharides. © 2013 AIP Publishing LLC. [http://dx.doi.org/10.1063/1.4808025]

I. INTRODUCTION

Cellulose, composed of linear polysaccharide chains with glucose repeat units connected by 1,4 beta glycosidic bonds, is a primary component in plant biomass.1 Crystalline cellulose microfibers, present in the core of plant cell walls, resist degradation by natural enzymatic hydrolysis processes.2 To overcome this recalcitrance, costly physical or chemical pretreatment steps are used to obtain high yields of fermentable sugars from biomass in the biofuel refining process.3-5 There is a great need for fundamental, molecular-based (microscopic) models that can accurately predict crystalline and amorphous cellulose structures and describe the forces that bind polysaccharide chains together in a microfiber so that efficient routes to deconstruct cellulose biomass can be rationally designed.4

Enzymatic hydrolysis of cellulose is a complex problem.2,6 The length and time scales of interest span orders of magnitude, from reaction mechanisms (10^-10 m; 10^12 s) to phase transitions in crystalline cellulose (10^-8 m; 10^-6 s) to cellulosomal interactions with plant cell walls (10^-6 m; 10^-3 s).7 Because no single computational method can address all of the relevant spatial and temporal scales, much recent effort in this field has been devoted to bridging molecular models and methods derived from quantum, atom, and, coarse-grained (CG) descriptions. For example, the development of a robust, coarse-grained molecular model for polysaccharide chains in a microfiber is viewed as an important step toward the goal of using multiscale simulation to study the enzymatic hydrolysis of cellulose biomass.8,9

Most biomass contains two forms of crystalline cellulose, cellulose-Iα (triclinic unit cell) and cellulose-Iβ (monoclinic unit cell).10,11 The relative distribution of each crystal structure in a given sample depends on the biomass source, with cellulose-Iα being predominant in algal and bacterial biomass and cellulose-Iβ being predominant in higher-order plant biomass.12 Cellulose-Iα is known to convert to cellulose-Iβ and these two structures comprise a model system for studying solid-state cellulose phase transformations.12,13 Hence, the capability to model both crystalline phases at the molecular level will help advance fundamental knowledge of cellulose stability and structure. The intent of this paper is to systematically derive coarse-grained models for polysaccharide chains and examine their suitability for future use in a multiscale simulation scheme in which the all-atom and

a)S. Markutsya and A. Devarajan contributed equally to this work.
b)Author to whom correspondence should be addressed. Electronic mail: mhlamm@iastate.edu.
coarse-grained descriptions of polysaccharide chains in cellulose are modeled concurrently.\textsuperscript{14,15} A starting structure of polysaccharide chains in cellulose-\textsubscript{Iα} is chosen to complement the recent developments in coarse-grained models for polysaccharides in cellulose-\textsubscript{Iβ}.\textsuperscript{8,16,17}

Several investigators have developed coarse-grained models for glucose in aqueous solution\textsuperscript{9,18–20} since glucose is the fundamental building block in a polysaccharide chain. Molinero and Goddard\textsuperscript{18} introduced the first coarse-grained model for \(\alpha\)-D-glucose and used a mapping scheme of three coarse-grained sites for glucose and one coarse-grained site for water. The non-bonded and bonded parameters for this coarse-grained model were fit to two-body Morse potentials and harmonic potentials, respectively. The use of the Morse potential allowed for a relatively simple implementation of the coarse-grained model in the molecular dynamics code; however, there is no clear evidence to suggest that this analytical potential best describes the non-bonded interactions between coarse-grained sites of glucose. Despite this uncertainty about the functional form of the Molinero and Goddard model, the mapping scheme itself is chemically sensible and it has been adopted for use in subsequent coarse-grained modeling studies on glucose.\textsuperscript{9,19,20} In these later studies, the force matching method\textsuperscript{21–23} has been used to parameterize the coarse-grained potentials for the three-site glucose model because this technique systematically optimizes coarse-grained forces from all-atom simulation trajectories without fitting them to a predetermined analytical function. In addition to glucose, the force matching approach has been used to obtain coarse-grained force fields for aqueous solutions of celllobiose,\textsuperscript{9} cellotetraose,\textsuperscript{9} and a polysaccharide containing 14 glucose units.\textsuperscript{19} These studies on glucose and polysaccharides have shown good agreement in the radial distribution functions obtained from the coarse-grained models and the reference all-atoms simulations. This evidence for the accuracy of the coarse-grained models with respect to the all-atom simulations is very promising for future developments aimed at deriving coarse-grained models for crystalline cellulose structures based on force matching.

A few studies have developed coarse-grained models for crystalline cellulose using alternative coarse-graining methods to force matching. Bu \textit{et al.}\textsuperscript{16} derived a coarse-grained model, for the surface of cellulose-\textsubscript{Iβ}, where the non-bonded and bonded parameters were fit to 6–12 Lennard-Jones potentials and harmonic potentials, respectively. The coarse-grained mapping follows the Molinero and Goddard scheme. Distinct rescaling factors were applied to the non-bonded interactions to introduce directionality, an attribute necessary for the coarse-grained model to produce a stable crystalline structure. Wohlerl and Berglund\textsuperscript{17} reparameterized the non-bonded interactions in the MARTINI force field for carbohydrates\textsuperscript{24} and added extra cellulose-cellulose interactions to generate a stable, coarse-grained model for an ordered cellulose-\textsubscript{Iβ} microfibril in water. The MARTINI model uses three sites per glucose residue with a coarse-grained mapping scheme that necessarily follows MARTINI mapping conventions rather than the Molinero and Goddard scheme. This includes a coarse-grained site for water that is equivalent to four water molecules. The models that were used in Refs.\textsuperscript{16 and 17} required constraints to maintain stable ordered structures. A constraint-free coarse-grained model for crystalline and amorphous cellulose fibrils in aqueous solution was developed by Srinivas \textit{et al.}\textsuperscript{5} In this model, each glucose residue was defined by a single coarse-grained site. Two sets of parameters were derived so that chains in alternating planes (termed origin and center) could be treated distinctly. The non-bonded and bonded interactions for this one-site model were parameterized using an iterative Boltzmann inversion approach.\textsuperscript{25} Although the model in Ref.\textsuperscript{8} is efficient and reliable for the large-scale, long-time simulations that are required to model phase transitions in cellulose, the mapping scheme of one coarse-grained site per glucose residue makes restoring the all-atom degrees of freedom very difficult. For example, specific details about the conformation of individual glucose residues in polysaccharide chains, and the relative orientation of glucose residues along polysaccharide chains and between polysaccharide chains in a microfiber are missing, yet this information is important for modeling the forces that bind cellulose together and result in its recalcitrance to hydrolysis. A five-site, coarse-grained model for glucose was introduced by Bellesia and co-workers and used to predict the structural phase transition between cellulose-\textsubscript{Iβ} and cellulose-\textsubscript{IIIα}.\textsuperscript{26} The non-bonded interactions were modeled with 6–12 Lennard-Jones potentials and the bonded interactions were modeled with harmonic potentials. The parameters for this model were assigned by comparisons with crystallographic data. This model was used in molecular dynamics simulations of an infinite crystal \textit{in vacuo}; hence, no additional constraints were applied to maintain crystallinity. The simulations using this model provide physical insight about the transition from cellulose-\textsubscript{Iβ} to cellulose-\textsubscript{IIIα}, specifically that any successful chemical pretreatment must disrupt the intersheet spacing in cellulose crystals.

In this paper, four coarse-grained models for polysaccharide chains are evaluated. The force matching method is used to derive the coarse-grained force fields for both bonded and non-bonded interactions, and no additional constraints are applied. To our knowledge, this is the first systematic comparison of coarse-grained mapping schemes for polysaccharide chains in cellulose. The remainder of the paper is organized as follows. Section II describes the force matching method used to derive the coarse-grained models for polysaccharides and gives the details of how the all-atom and coarse-grained simulations were performed. In Sec. III, the simulation results are presented and discussed. A brief summary and conclusions are given in Sec. IV.

II. METHODS

A. Multiscale coarse-graining by force-matching

There are several strategies available for deriving and parameterizing coarse-grained models for complex biomolecular systems; these strategies have been categorized according to how quantitatively they are coupled to atomistic scale properties.\textsuperscript{14,15} One method that has been shown to bridge between the atomistic and the coarse-grained scales with high quantitative accuracy is the multiscale coarse-graining
method\textsuperscript{23,27,28} based on force-matching.\textsuperscript{21,22} This technique and the resulting coarse-grained potentials are termed “multiscale” due to the fact that the atomistic degrees of freedom directly inform the effective coarse-grained forces. This quantitative coupling between scales is a prerequisite for the eventual development of a concurrent multiscale molecular simulation scheme. For these reasons, the multiscale coarse-graining method (called force matching for the remainder of the paper) was used to obtain the coarse-grained force fields for polysaccharides in cellulose that are evaluated here.

A detailed description of the force matching method can be found elsewhere\textsuperscript{23,27,28} Briefly, for a given configuration from the reference all-atom molecular dynamics simulation, the positions of $N$ coarse-grained sites and the net forces $F_{i}^{\text{ref}}$ acting on them are computed. Recognizing that the net force on a particular coarse-grained site $i$ is due to the sum over all the effective forces between pairs of coarse-grained sites leads to the construction of an objective function

$$\sum_{j=1}^{N} f_{ij}(r_{i}, r_{j}, p_{1}, p_{2}, \ldots, p_{m}) = F_{i}^{\text{ref}}, \quad i = 1, 2, 3, \ldots, N,$$

where $f_{ij}(r_{i}, r_{j}, p_{1}, p_{2}, \ldots, p_{m})$ is the force on the $i$th coarse-grained site due to the $j$th coarse-grained site, and $r_{i}$ and $r_{j}$ are the position vectors for the $i$th and $j$th coarse-grained site, respectively. The analytical function for $f_{ij}$ is not known a priori so cubic splines are chosen to systematically write the pairwise force as a linear function of $m$ unknown parameters $(p_{1}, p_{2}, \ldots, p_{m})$. Hence, a system of $N$ linear equations with $m$ unknowns is obtained. This system of equations is over-determined ($N > m$) and can be solved using the singular value decomposition method.\textsuperscript{29} The parameters obtained from each all-atom configuration are averaged over the total number of configurations sampled from the reference all-atom trajectory. Both non-bonded and bonded interactions are derived directly from force matching. The tabulated forces for the coarse-grained models considered here are given in the supplementary material.\textsuperscript{45}

\section*{B. Multiscale coarse-grained models for polysaccharides}

Four coarse-grained mapping schemes for polysaccharide chains composed of $\beta$-D-glucopyranose repeat units connected by 1,4 beta glycosidic bonds were evaluated. For simplicity, the $\beta$-D-glucopyranose repeat units will be called glucose residues for the remainder of this paper. Figure 1 shows the four mapping schemes, and the atom labeling scheme for the glucose residue is shown in Figure 2. From left to right, the columns illustrate the all-atom structure of a single glucose residue in a polysaccharide chain (end units not shown), the number of coarse-grained sites in the glucose residue, and the coarse-grained structure of polysaccharide chains in a crystalline cellulose-lar structure.

The one-site CG model, in which each glucose residue (circled) is defined by a single coarse-grained bead ($A_1$) placed at the center of mass of the atoms in the residue, is shown in Figure 1(a). This model is similar to the model used by Srinivas et al.,\textsuperscript{8} however, there are differences between the one-site model used previously and the one-site model presented here. In Ref. 8, the coarse-grained site definition includes the oxygen in the 1,4 glycosidic bond attached to the fourth carbon (C4) of each glucose residue, while the model in Figure 1(a) has this oxygen bonded to the first carbon (C1). Figure 2 shows the positions of C1 and C4 on the glucose residue. Another difference is that Ref. 8 uses two distinct parameter sets for the coarse-grained sites, depending on whether the polysaccharide chain is an origin or center chain in the microfiber. The model in Figure 1(a) assigns unique parameters to each coarse-grained site along the polysaccharide chain, but makes no distinction based on chain positions in the crystalline structure.

Figure 1(b) shows the two-site CG model. The first site is defined by a bead ($A_2$) placed at the center of mass of the atoms in the glucose residue (excluding the oxygen atoms in the glycosidic bonds); the second site is defined by a bead placed at the position of the glycosidic oxygen ($B_2$). This coarse-grained mapping scheme allows for independent rotation of the pyranose rings about the glycosidic bond.

The three-site CG model shown in Figure 1(c) is consistent with previous three-site models for glucose\textsuperscript{9,18–20} and polysaccharides.\textsuperscript{9,19} In this model, each glucose residue is defined by three beads ($A_3$, $B_3$, $C_3$), placed at the center of mass of the corresponding group of atoms. The glycosidic oxygen atom is included in site $A_3$. In the four-site CG model shown in Figure 1(d), each glucose residue is defined with four beads where $A_4$, $B_4$, and $C_4$ are placed at the center of mass of the corresponding atoms and $D_4$ is placed at the position of the glycosidic oxygen. This model is similar to the three-site CG model except that the $A_4$ bead no longer contains the glycosidic oxygen; this scheme again allows for independent rotation of the pyranose rings as in the two-site CG model.

\section*{C. Molecular dynamics simulations}

The all-atom molecular dynamics simulation was run using GROMACS.\textsuperscript{30} The polysaccharides were modeled with the CHARMM36\textsuperscript{31} force field with a cutoff distance of 1.6 nm for long-range van der Waals interactions. The electrostatic interactions were calculated using the reaction field method\textsuperscript{32} in which the electrostatic force is calculated explicitly for a separation smaller than the cutoff radius. The cutoff radius for electrostatic interactions was 1.2 nm. For a separation outside the cutoff radius, the system is treated as a dielectric continuum. The “Reaction-field-zero” option, which sets the potential to zero beyond the electrostatic cutoff radius and uses an infinite dielectric constant, was used to ensure good energy conservation.\textsuperscript{30} This approach gives better performance on massively parallelized computers while delivering accurate results when compared with the particle mesh Ewald method.\textsuperscript{33}

An initial configuration for the cellulose-$\text{\textit{la}}$ microfiber was based on the unit cell structure data reported by Nishiyama et al.\textsuperscript{10} A structure containing 12 polysaccharide chains, each having a length of 12 glucose residues, was built with an in-house script. Twelve is the minimum number of chains that form a supercell with a structure that has...
FIG. 1. Coarse-grained mapping schemes for the glucose residues in a polysaccharide chain. The columns show the progression from the all-atom structure of a glucose residue in a polysaccharide chain (left), to the coarse-grained model of a glucose residue (center), to the coarse-grained representation of 12 polysaccharide chains in a cellulose-I\textsubscript{\alpha} microfiber (right). The rows show: (a) 1-site CG model (A\textsubscript{1}); (b) 2-site CG model (A\textsubscript{2}B\textsubscript{2}); (c) 3-site CG model (B\textsubscript{3}C\textsubscript{3}A\textsubscript{3}); and (d) 4-site CG model (B\textsubscript{4}C\textsubscript{4}A\textsubscript{4}D\textsubscript{4}). The lines drawn around the atoms in the first column are shown to guide the eye and are not meant to imply the shape of the coarse-grained sites.

Both inner and outer chains. Although this structure has fewer polysaccharide chains than most other simulation studies of microfibers reported in the literature,\textsuperscript{8,16,34–37} there are no significant differences in the structural properties of interest when compared to a larger, 36-chain structure. (See Sec. IV in the supplementary material.)\textsuperscript{45} The chain length considered here (12 residues) is necessarily much shorter than the degree of polymerization typically found in cellulose fibers (10\textsuperscript{2}–10\textsuperscript{4} residues)\textsuperscript{36} because all-atom simulations at the longer chain length would be prohibitively expensive. Furthermore, a smaller chain length makes it feasible to evaluate the multiple coarse-graining schemes; the insight gained from this evaluation will be used in future coarse-grained molecular dynamics computations on larger cellulose fiber structures.

The crystalline structure was simulated \textit{in vacuo} with periodic boundary conditions. The \textit{in vacuo} condition is sufficient for examining interactions between glucose residues. There is good agreement between this structure simulated \textit{in vacuo} and simulated while embedded in a 36-chain structure (See Sec. IV in the supplementary material\textsuperscript{45}), which indicates that no artifacts were introduced to the glucose-glucose interactions. Clearly, the presence of water at the surface of a cellulose microfiber will affect the surface structure of the polysaccharide chains; this will be addressed in future work.
After energy minimization, the microfiber had the dimensions 7.1 nm × 2.1 nm × 2.3 nm and it was placed in a simulation box with dimensions 10.4 nm × 9.1 nm × 8.9 nm. The terminal polysaccharide chains were capped to treat them as chains of finite length. The system was equilibrated with NVT simulation at 300 K for 8 ns. After equilibration, an NVT production simulation was run at 300 K for 12 ns. A Nosé-Hoover thermostat was used to control temperature. The time step in the all-atom molecular dynamics simulations was 0.5 fs. Configurations containing positions, velocities, and forces were collected every 1 ps, for a total of 12000 configurations.

The coarse-grained molecular dynamics simulations were run using LAMMPS. For these simulations, the initial topology, the dimensions of the microfiber, the thermostat, the time step, and the total simulation time were identical to those used in the all-atom molecular dynamics simulation. The LAMMPS code was modified to use tabulated forces for bond length, bond angle, dihedral, and non-bonded interactions. In the multiscale coarse-grained models, the bonded interactions can occur between beads within the glucose residue (for two-, three-, and four-site CG models) and between beads on neighboring glucose residues in the same chain. The distances of separation between beads on neighboring glucose residues in the same chain are large enough that the non-bonded forces between them are effectively zero. For computational efficiency these non-bonded interactions between neighboring glucose residues in the same chain are prevented by setting the forces between these sites to zero.

### III. RESULTS AND DISCUSSION

This section describes the results from molecular dynamics simulations of polysaccharide chains using the multiscale coarse-grained models discussed in Sec. II. The primary evaluation of the CG models is conducted by comparing radial distribution functions obtained from the CG and all-atom simulations, as this is a direct test of the accuracy of the CG model with respect to the all-atom MD system. Complete agreement in radial distribution functions is not guaranteed by force matching, as it would be in a structure matching approach where the radial distribution function of the CG system is fit to the all-atom radial distribution function.

The intrachain radial distribution functions from the molecular dynamics simulation using the one-, two-, three-, and four-site CG models are compared with molecular dynamics simulation results obtained using an all-atom model. For the three-site and four-site CG models, this is followed by an analysis of the glucose residue conformations and the relative orientation between neighboring glucose residues. This paper aims to evaluate how well coarse-grained models reproduce polysaccharide chain conformations. Due to the minimal supercell size used (12 chains), no interchain radial distributions are computed; increasing the supercell size to study interchain properties is the topic of future work.

To make direct comparisons between the coarse-grained models and the all-atom model, the configurations from the all-atom molecular dynamics simulation are first reduced by the appropriate coarse-grained mapping scheme before calculating the distribution function of interest. All of the simulations presented here are for polysaccharide chains of finite length (12 glucose residues), which means that the terminal glucose residues were capped. To avoid chain-end effects, these 24 terminal glucose residues have been excluded from the distribution function analyses.

### A. Evaluation of the one-site and two-site coarse-grained models

The one-site and two-site CG models shown in Figures 1(a) and 1(b) are minimal models for simulating the structure and dynamics of polysaccharide chains in cellulose. These models are useful for studying phenomena, such as structural phase transitions in cellulose, that require tracking chain-scale structure and motion for long simulation times (>1 μs).

Figures 3 and 4 show results from molecular dynamics simulations of the polysaccharide chains and compare the one-site and two-site CG models, respectively, with the all-atom reference model. In these figures and for the remainder of the paper, single and double prime notation is used to distinguish between pairs from neighboring glucose residues (e.g., A1A1′) and from non-neighboring glucose residues (e.g., A1A1″), respectively. Figure 3 shows the intrachain radial distribution function, g(r), for the one-site CG model. The peak positions at 5.5 Å, 10.5 Å, and 15.5 Å in the intrachain g(r) are consistent with those obtained from the reference all-atom data. The first peak at 5.5 Å agrees well with the neighboring residue bond distance of 5.3 Å reported by Srinivas et al. for a similarly defined one-site CG model derived from all-atom trajectories by Boltzmann inversion.

The two-site CG model is considered to be a minimal model that allows the pyranose rings to rotate independently about the glycosidic bond in the polysaccharide chain. Figure 4 shows intrachain radial distribution functions for the two-site CG model. In Figure 4(a), the intrachain distributions for A2A2′ and A2A2″ from the coarse-grained model are broader (hence, the peak heights are lower since the area under each distribution must be the same) and the peak positions slightly overestimate those obtained from the reference all-atom data. Figure 4(b) shows that the intrachain
distribution for the B₂B₂′ distance is consistent with the all-atom reference model, as one would expect, since this is the distance between oxygen atoms in consecutive glycosidic bonds and, hence, no degrees of freedom are removed from this site in the coarse-grained model.

B. Evaluation of the three-site and four-site coarse-grained models

The three-site and four-site CG models shown in Figures 1(c) and 1(d) are minimal models that permit the investigation of conformational changes within the glucose residue and the structure and dynamics of the polysaccharide chains in cellulose. These models have an advantage over the one-site and two-site models, in that mapping all-atom details to coarse-grained configurations should be more reliable.

Figure 5 shows the intraresidue radial distribution functions for the three-site (Fig. 5(a)) and four-site (Fig. 5(b)) CG models. As stated above, direct comparisons between radial distributions from the coarse-grained and the all-atom simulations are made by first reducing the configurations from the all-atom molecular dynamics simulation to the appropriate coarse-grained scheme before calculating the reference all-atom radial distribution functions. The radial distribution functions for the A₃B₃ (A₄B₄) and A₃C₃ (A₄C₄) pairs agree with those obtained with the all-atom reference model. The radial distribution function for B₃C₃ (B₄C₄) displays two peaks that are positioned at the same distances of separation, 3.3 Å and 3.7 Å, as the all-atom reference model, but the peak intensities differ. The conformations contributing to the double B₃C₃ (B₄C₄) peak arise due to the distinct conformations adopted by the primary alcohol in the glucose residue. There are four possible primary alcohol conformations: gauche-trans (GT), trans-gauche (TG), gauche-gauche (GG), and trans-trans (TT), where the first letter corresponds to the relative O₆−O₅ orientation and the second letter corresponds to the relative O₆−C₄ orientation (see Figure 2 for atom numbering system). In this context, a G is assigned to

FIG. 4. Intrachain radial distribution function for the two-site CG model of cellulose-Iα. (a) A₂A₂′ and A₂A₂′′, (b) B₂B₂′. The coarse-grained sites are defined as shown in Fig. 1(b). See text for explanation of single and double prime notation. For comparison, the intrachain radial distribution functions taken directly from the all-atom molecular dynamics simulation are shown.

FIG. 5. Intraresidue radial distribution functions for the three-site and four-site CG models of cellulose-Iα. (a) A₃B₃ (dotted blue line); A₃C₃ (dashed green line); B₃C₃ (dashed-dotted blue line). (b) A₄B₄ (dotted blue line); A₄C₄ (dashed green line); B₄C₄ (dashed-dotted blue line). The coarse-grained sites are defined as shown in Figs. 1(c) and 1(d). For comparison, the intraresidue radial distribution functions taken directly from the all-atom molecular dynamics simulation are shown (solid red line) for each pair of coarse-grained sites.
FIG. 6. Intrachain radial distribution function for coarse-grained sites in neighboring glucose residues for the three-site and four-site CG models of cellulose-Iα.

(a) A3A3′ (dotted blue line), B3-B3′ (dashed green line); (b) A4A4′ (dotted blue line), B4-B4′ (dashed green line); (c) C3-C3′ (solid blue line); and (d) C4-C4′ (solid blue line). The coarse-grained sites are defined as shown in Figs. 1(c) and 1(d). For comparison, the intrachain radial distribution functions taken directly from the all-atom molecular dynamics simulation are shown (solid red line) for each pair of coarse-grained sites.

angles with absolute values less than 90°; a T assigned to angles with absolute values greater than or equal to 90°. Upon analyzing the all-atom primary alcohol conformations corresponding to the double peak for B3C3 (B4C4), it is seen that the first peak has contributions from the GG (67%), TG (29%), and TT (4%) conformations, while the second peak is due to only GT conformations. This is the first time that coarse-grained models for glucose residues have been shown to predict the occurrence of distinct primary alcohol conformations along a polysaccharide chain.

FIG. 7. Distribution of the angles $\phi$, $\psi$ for the three-site and four-site CG models of cellulose-Iα. (a) Three-site CG model (blue squares) and (b) four-site CG model (blue squares). For comparison, the distribution of angles computed using the all-atom model is also shown (red squares). The black dashed lines are drawn as boundaries around the all-atom data to assist the identification of unphysical angles predicted by the coarse-grained models.
Fourier transform of the intrachain radial distribution functions for the three-site and four-site CG models for neighboring glucose residues along the polysaccharide chain. The peak positions for A3-A3′ slightly underestimate the all-atom data, while the peak positions for A4-A4′ slightly overestimate the all-atom data. The peak positions for B1-B1′ (B2-B2′) agree with the all-atom reference model. The distributions for C3-C3′ (C4-C4′) are significantly different from the all-atom reference model. Figure 6(c) shows that the three-site CG model predicts an additional peak at 2 Å in addition to the main peak located at 6.8 Å. Figure 6(d) shows that the four-site CG model correctly predicts the peak position at 6.5 Å but the distribution has a broad shoulder from 3 Å to 5.5 Å that is not observed with the all-atom reference model. Clearly, the neighboring coarse-grained sites C3 (C4) and C3′ (C4′) are permitted to interact at closer distances than predicted by the reference all-atom data. This suggests that there may be unphysical conformations between neighboring glucose residues in the coarse-grained models, and that these unphysical conformations are more prevalent in the three-site CG model.

To test for unphysical conformations, the relative orientation between glucose residues is defined by two angles. The angle $\phi$ (shown in the inset in Figure 7) determines the glucose-glucose conformation and is defined as the angle between $n_1$, a normal vector to the plane $A_3B_3C_3$ ($A_4B_4C_4$), and $n_2$, a normal vector to the plane $A_3′B_3′C_3′$ ($A_4′B_4′C_4′$). The angle $\psi$ is defined as the angle formed by $A_3B_3C_3′$ ($A_4B_4C_4′$). Figure 7 shows the $\phi, \psi$ distribution for the three-site and four-site CG models compared to the all-atom reference data. Based on the all-atom data, values of $|\phi| < 100^\circ$ are deemed unphysical conformations; these conformations appear between the black dashed lines drawn in Figure 7. The percentage of unphysical conformations identified for the three-site CG model (24.9%) is greater than that identified for the four-site CG model (16.7%).

Based on this analysis, all unphysical conformations ($|\phi| < 100^\circ$) were removed from the data set and the radial distribution function for coarse-grained C3C3′ (C4C4′) sites was recalculated. Figure 8 shows the intrachain radial distribution functions with the unphysical conformations removed. The peak positions predicted by the three and four-site CG models are now in better agreement with the all-atom reference model. The four-site model, which allows for independent rotation of pyranose rings around the glycosidic bonds that connect glucose residues, is best at reproducing the glucose-glucose conformations observed with the reference all-atom model.

IV. CONCLUSIONS

In this work, multiscale coarse-grained models for polysaccharides were derived using the force matching method. Four different coarse-grained mapping schemes (one-, two-, three-, and four-site) for glucose residues in polysaccharide chains were investigated. The performance of the coarse-grained models was evaluated by comparing the intraresidue and intrachain properties of polysaccharides predicted by coarse-grained molecular dynamics simulations with those predicted by a reference all-atom simulation.

The one-site and two-site CG models are best suited to modeling chain-scale properties at microsecond time scales, such as the transition from crystalline to amorphous structure in a cellulose microfiber. As expected, the one-site CG model for glucose residues leads to intrachain radial distribution functions that are in excellent agreement with those predicted by the all-atom reference simulation. The two-site CG model, which is defined to allow the pyranose rings to rotate independently, also performs well. Based on the distribution functions presented here there is no obvious reason to choose the two-site CG model over the one-site CG model. However, future studies of the structural phase transitions in cellulose with the two-site CG model may be warranted since the independent rotation of the pyranose rings may have a significant impact on interchain and intersheet interactions in the crystalline cellulose structure. Indeed, results from ab initio studies show that the orientation of the hydroxyl groups, as well as the relative residue conformations, has a significant effect on the hydrogen bonding in intrachain, interchain and intersheet interactions. For example, GT-TG interactions are the strongest interchain interactions on a per hydrogen bond...
basis and the large percent of these types of interactions will certainly affect the phase transitions.

The three-site and four-site CG models have sufficient degrees of freedom to model conformational changes in the glucose residues along the polysaccharide chains. In general, both the three-site and the four-site CG models can reproduce the intrachain radial distributions from the all-atom model. In particular, these models adequately sample the unique primary alcohol conformations that are observed in the all-atom trajectory. To the best of our knowledge this is the first time that coarse-grained models have been shown to predict glucose residue conformations in a cellulose microfibril. Further analysis of the three-site and four-site CG models shows that the four-site CG model samples fewer unphysical relative orientations between glucose residues. Thus, it appears that a coarse-grained mapping scheme that decouples the rotation of neighboring pyranose rings about the glycosidic bond in polysaccharide chains, such as the four-site CG model presented here, is essential for reliably modeling chain conformation without introducing constraints.

A long-term goal is to develop multiscale potentials for cellulose so that one may switch seamlessly from coarse-grained to all-atom descriptions of a system during a simulation. Based on the results presented here, the outlook is promising for developing multiscale models of crystalline cellulose microfibers. It is envisioned that coarse-grained molecular dynamics simulations, using models similar to the four-site CG model described here, can be used to sample a cellulose microfibril structure over large time and length scales with enough fidelity that all-atom configurations can be reliably reverse-mapped and used for subsequent computations of local reaction mechanisms. Given that any coarse-grained model can only perform as well as the reference (input) all-atom trajectory, future research will involve studying the behavior of cellulose in solvated environments using coarse-grained models derived from quantum mechanical fragmentation approaches43 as an alternative to all-atom potentials based on classical molecular mechanics.

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 USDOE under Contract No. DE-AC02-07CH11358. The authors thank Professor G. A. Voth for providing the multiscale coarse-graining software developed by his research group and Professor J. C. Smith for providing the parameters for the CHARMM36 potential used in the all-atom simulations.

44 The $\phi$, $\psi$ angles defined here are not dihedral angles and the plots in Figure 7 should not be confused with the well-known Ramachandran diagram for dihedral angles in peptides and carbohydrates.
45 See supplementary material at http://dx.doi.org/10.1063/1.4808025 for the tabulated forces used for the coarse-grained simulations, supporting results for establishment of the minimum supercell size for the all-atom simulations, and supporting results for the all-atom molecular dynamics simulations.