Chapter 3

Atomistic Simulation of Lignocellulosic Biomass and Associated Cellulosomal Protein Complexes

Loukas Petridis,[1,*] Jiancong Xu,[1] Michael F. Crowley,[2] Jeremy C. Smith,[1] and Xiaolin Cheng[1]

[1]University of Tennessee / ORNL Center for Molecular Biophysics, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6164, USA
[2]Chemical and Biosciences Center, National Renewable Energy Laboratory, Golden, CO 80401-3393, USA
[*]Petridisl@ornl.gov

Computer simulations have been performed to obtain an atomic-level understanding of lignocellulose structure and the assembly of its associated cellulosomal protein complexes. First, a CHARMM molecular mechanics force field for lignin is derived and validated by performing a molecular dynamics simulation of a crystal of a lignin fragment molecule and comparing simulation-derived structural features with experimental results. Together with the existing force field for polysaccharides, this work provides the basis for full simulations of lignocellulose. Second, the underlying molecular mechanism governing the assembly of various cellulosomal modules is investigated by performing a novel free-energy calculation of the cohesin-dockerin dissociation. Our calculation indicates a free-energy barrier of ~17 kcal/mol and further reveals a stepwise dissociation pathway involving both the central β-sheet interface and its adjacent solvent-exposed loop/turn regions clustered at both ends of the β-barrel structure.

Introduction

Plant cell wall structure has come under renewed interest in the context of producing bioethanol from the enzymatic hydrolysis of lignocellulosic biomass

(*1–5*). The plant cell wall is made of cellulose microfibrils that are embedded in a matrix of polysaccharides (hemicelluloses and pectins), lignins, and proteins (*6*). Cellulosic ethanol production is a multi-stage process often involving, first, the pretreatment of biomass, then the hydrolysis of cellulose (and hemicelluloses) by enzymes to smaller oligosaccharides, and, finally, the fermentation of sugars to ethanol. The hydrolysis step is the bottleneck in the process because of the natural resistance, or "recalcitrance," of plant cell walls to degradation (*2*).

Given the complex and heterogeneous nature of biomass materials, a better understanding of their structure, dynamics, and degradation pathways becomes a necessary first step toward overcoming their recalcitrance to hydrolysis. Through years of extensive biochemical and biophysical studies, it has been established that although biomass recalcitrance is a very complex phenotype, with many factors contributing to it, lignin plays an important role (*7*). There is evidence of an inverse correlation between the rate of biomass hydrolysis and the lignin content (the amount of lignins present in the cell wall) (*8*). Lignin acts as a physical barrier, preventing enzymes from reaching the cellulose substrate. There is also evidence that lignin-enzyme interactions significantly contribute to the decline of rate observed during hydrolysis of lignocellulose substrates (*8*). Lignin poses an additional challenge in that, unlike hemicellulose and pectins, it is not readily removed with economically sustainable pretreatment. It has been suggested that, although lignin is initially released during pretreatment, it precipitates back on the cellulose surface at the end of the process (*9*). Another factor contributing to biomass recalcitrance is the crystallinity of cellulose. Cellulose can be found in crystalline fibrils, the compact structure of which impedes enzymatic access. In comparison, amorphous cellulose is readily digested by enzymes (*10*). Lignin content and the degree of crystallinity of cellulose had the greatest impact on biomass digestibility of Poplar wood (*11*). A more recent study of *alfalfa* lines found that the efficiency of enzymatic hydrolysis and the amount of total sugars released is proportional to the plant's lignin content (*12*).

A second promising avenue for altering biomass recalcitrance is designing more efficient enzyme systems to degrade the plant cell wall. For this, we need to more completely understand the structure, mechanism, and function of these enzyme systems. Generally, two classes of enzyme systems have been observed in microorganisms (*13–15*). One class consists of several individual endoglucanases, exoglucanases, and ancillary enzymes that can act synergistically to deconstruct plant cell walls. These enzymes are usually found in aerobic fungi and bacteria, of which the glycosyl hydrolases from *Trichoderma reesei (T. reesei)* is the best studied. The other system class, which is usually found in anaerobic microorganisms, involves the formation of a large, extracellular enzyme complex called the cellulosome, which consists of a scaffolding protein and many associated enzymes. Lignocellulosic biomass is structurally heterogeneous and includes many components in addition to cellulose, so efficient decomposition requires a variety of enzymes with a wide range of specificities and activities. To this end, the multienzyme cellulosome system seems particularly advantageous and has become a paradigm for designing more efficient enzyme complexes and biomimetics. During the past few years, an increasing number of cellulosome systems have been identified (*14*). Information is also becoming available

regarding the structural principles governing the interactions among various cellulosomal domains (*16*, *17*). A cellulosome consists of a fibrillar protein (called the scaffolding protein) that contains binding sites (called cohesins) for the cellulosomal enzyme modules positioned periodically along the fibrils. In addition to their catalytic domains, all cellulosomal enzymes contain a cohesin-binding site called a dockerin. The cohesin–dockerin interaction is an important factor in cellulosome assembly. For example, the *Clostridium thermocellum* cellulosome assembles through the interaction of a type I dockerin with one of several type I cohesin modules. Although cohesins and dockerins exhibit relatively high sequence homology, the interaction between cohesins and dockerins is generally species specific (i.e., cohesins from one species do not recognize and interact with dockerins present in other species) (*16*, *18*).

Although computational studies have proven useful in providing detailed insight into diverse biochemical/biophysical processes otherwise inaccessible from experiment alone, atomistic simulation of lignocellulosic models has so far been limited. With the help of high-performance computing, the foundations for accurate simulation of these materials have been laid recently (*19*, *20*); and various simulations are starting to emerge that can be employed to derive physical properties of lignocellulosic biomass, thus serving as a reference for interpreting an array of biophysical experiments. On another front, atomic-level structural information is now being accumulated for individual cellulosomal modules (*17*, *21*), although the structure of the entire cellulosome complex is still difficult to obtain. The availability of this partially complete data from different sources, however, offers great opportunity for using computational approaches to study the structure, dynamics, and assembly process of cellulosome complexes. In this chapter, we will focus on two lines of our research as the initial efforts toward our long-term goal. One is on the parameterization of a potential energy function for simulating lignocellulosic biomass. The other is on modeling cohesin-dockerin interaction in cellulosome.

Toward More Realistic Simulation of Lignocellulosic Biomass

The chemical composition and structure of lignins are highly heterogeneous, varying significantly between different plant species and even within different parts of the same plant wall. Although the exact chemical formula of lignins is not known, abundant information is available on its composition. Lignins are composed primarily of three units: *p*-hydroxyphenyl (H), guaiacyl (G), and syringyl (S), derived by oxidation of three alcohol monolignols: *p*-coumaryl, coniferyl, and sinapyl, respectively (*22*) (Figure 1a). There are various linkages that connect the units, leading to the formation of the branched lignin biopolymer. The most common linkages are β-O-4', 5-5', α-O-4', and β-5' in guaiacyl and syringyl (Figure 1b). There is an ongoing heated debate on how monolignols couple to form the lignin macromolecule. One theory suggests that lignin monomers are oxidized and then coupled in a combinatorial fashion (*54*). The second theory suggests that lignin primary structure is controlled at the proteinaceous level (*55*). To the best of our knowledge, there are no currently

published reports on the exact primary structure of lignins. For this reason, our studies are based on the assumption that this primary structure is combinatorically derived. We stress that our work does not attempt to validate either of the previously mentioned theories.

Although there is a large volume of simulation work on cellulose (*23–28*), there are relatively few computational studies of lignin. Previous computational studies (*29–32*) employed the CHARMM27 empirical force field (*33*), which was developed to model proteins rather than lignin. In recent work (*20*), we presented the first essential step towards the accurate computer simulation of lignin: the derivation of an empirical molecular mechanics (MM) force field. Together with the existing force field for polysaccharides, this force field will enable full simulations of lignocellulose.

A Molecular Mechanics Force Field for Lignin

Parameterization Strategy

In this section, we outline the general strategy employed to obtain the lignin force field. The CHARMM potential energy function of a molecule is as follows (*33*):

$$E = \sum_{bonds} K_b (b - b_o)^2 + \sum_{angles} K_\theta (\theta - \theta_o)^2 + \sum_{U-B} K_{ub} (s - s_o)^2 + \sum_{dihedrals} K_\phi [1 + \cos(n\phi - \delta)]$$
$$+ \sum_{impropers} K_\psi (\psi - \psi_o)^2 + \sum_{non-bonded} \left\{ \varepsilon_{ij} \left[\left(\frac{R_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4 \varepsilon_0 \pi r_{ij}} \right\}, \qquad (1)$$

where contributions to the energy include bonded (bond, angle, Urey-Bradley, dihedral, and improper dihedral) and non-bonded (the Lennard-Jones 6-12 potential for the van der Waals interactions and Coulomb interactions) terms. The force constants K and partial atomic charges q are molecule-dependent and must be optimized to model any specific molecule prior to performing the simulation.

This parameterization of lignin follows the main procedure of parameterization of proteins (*33*) and ethers (*34*) for the CHARMM force field. Lignin also has a linear ether bond, but it is different from those examined in (*34*) in that the oxygen is bonded to a phenyl ring and a tertiary carbon. For this reason, it was necessary to create a new atom type, OET, to represent the lignin's ether oxygen. Parameters were optimized by considering two factors. First, the target data were reproduced as closely as possible. Second, compatibility with the existing CHARMM force field was ensured by restricting optimization to the new parameters that do not already exist.

Two model compounds were used. The first model system, methoxybenzene, also known as anisole (see Figure 1c), incorporates the basic features of the β-O-4' link, an ether oxygen bonded to a tertiary and an aromatic carbon. Anisole was used to obtain all parameters involving the ether oxygen atom. The second compound (see Figure 1d) is *p*-hydroxyphenyl (PHP), the simplest lignin unit. PHP was used to obtain all lignin parameters not involving the ether oxygen.

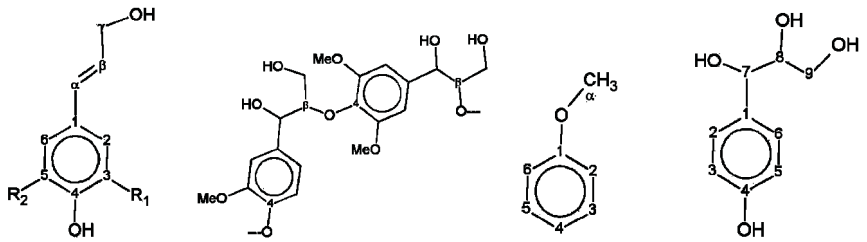

Figure 1. (a) The tree monolignols: p-coumaryl (R_1 = R_2 = H), coniferyl (R_1 = H, R_2 = OMe), and sinapyl (R_1 = R_2 = OMe). (b) A guaiacyl unit connected with a β-O-4' linkage to a syringyl unit. (c) Model compound anisole. (d) Model compound PHP.

The optimization strategy for the new parameters is summarized below (Figure 2). Equilibrium values for bonds, angles, and dihedrals were taken from MP2/6-31G* QM-optimized geometries and were not further revised. The van der Waals parameters were taken unaltered from the CHARMM force fields (*33*), including those for the new atom type, OET. Initial values for the partial atomic charges of O_1, C_1, and Cα were deduced from a restricted fit to the quantum mechanics (QM) electrostatic potential (RESP) on selected grid points (*35*), while all other partial charges were fixed to their original CHARMM values. An iterative procedure, described in the next paragraphs, was followed until convergence was reached.

Optimization of Partial Atomic Charges

Charges were further optimized with respect to the QM interaction energies using a supramolecular approach with a model compound (anisole) interacting with one water molecule. The partial charges were adjusted to reproduce minimum distances and interaction energies between anisole and a TIP3P water molecule (*36*). Two geometries were considered in this supramolecular approach: the first d_0 with water lying on the phenyl plane, and the second d_{120} with the water hydrogen pointing at the position of the ether oxygen lone pair. A list of all final atomic charges is shown in Table 1. Only three charges (O_1, C_1, and Cα) were optimized, with the rest being kept to their CHARMM values.

To mimic the effect of electronic polarizability, which is not explicitly taken into account in additive force fields, atomic charges were purposely overestimated. This leads to an enhanced molecular dipole moment, with the QM gas-phase dipole moment being 1.42 Debyes, whereas the MM value is 1.66 Debyes. Table 2 compares the MM and QM interaction energies and distances, which were used to optimize the anisole charges. The empirical calculations successfully reproduced the scaled QM interaction energies, with the error being less than 3%. The empirical model gives distances about 0.3Å shorter than the QM values, a result of intentionally overestimating the gas-phase charges to

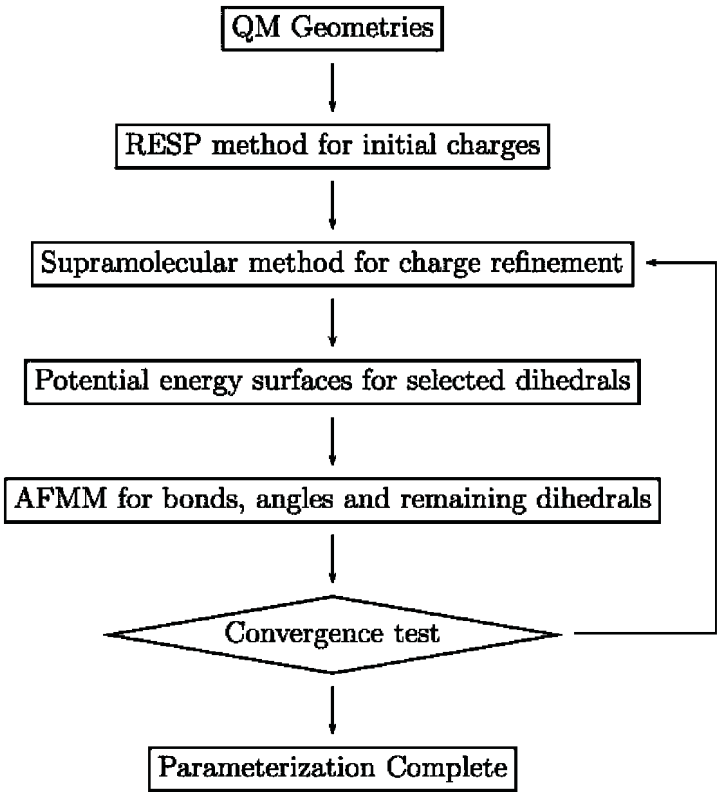

Figure 2. Schematic representation of parameterization strategy. (Reproduced from reference (20). Copyright 2008 Wiley Periodicals, Inc.)

Table 1. A list of the anisole atoms with their respective charges[a]

Atom name	*Atom type*	*Charge*
C_α	CT3	-0.060
$H_{\alpha1}$, $H_{\alpha2}$, $H_{\alpha3}$	HA	0.090
O	OET	-0.280
C1	CA	0.070
C_2, C_3, C_4, C_5, C_6	CA	-0.115
H_2, H_3, H_4, H_5, H_6	HP	0.115

[a] Atom names refer to Figure 1c and atoms types follow the CHARMM27 force field with the new atom type labeled as OET.

obtain good condensed-phase properties. In the previous general force field for ethers, a similar behavior was observed with a 0.3Å difference between QM and MM (*34*). Finally, the electronic charge density was examined by Mulliken analysis (using the NWChem software), and the charge transfer was found to be insignificant.

After completing parameterization, we performed a further calculation to ensure that the partial atomic charges of Table 1, derived using a model compound, can be transferred to lignin. The minimum interaction energies and distances between a lignin dimer (G and S units connected with a β-O-4' linkage shown in Figure 1b) and a water molecule were obtained without further modifying the charge parameters. The excellent agreement between the QM and MM interaction energies justifies using these charges for the β-O-4' lignin linkage.

Dihedral Parameters

After completing the non-bonded terms, parameters for dihedral rotations were deduced from the QM potential energy surfaces. Six dihedral rotations were considered. The two rotations around the β-O-4' linkage (ω_1= X-C_1-O-C_α and ω_2= C_1-O-C_α-H, where X refers to any atom types) were obtained using the anisole model compound. The remaining four dihedrals that do not involve the ether oxygen (ω_3=C_2-C_1-C_7-X, ω_4=C_1-C_7-O_7-HO_7, ω_5=C_1-C_7-C_8-X, and ω_6=X-C_8-C_9-X) were deduced from the more complex rotational potential energy profiles of the second model compound, PHP. The optimization was based on reproducing the adiabatic QM energy surfaces. As an example, two plots are shown in Figure 3. In Figure 3a, the MM surface closely follows the target QM data, whereas in Figure 3b, although the agreement between the QM and MM data is not perfect, the rather complex shape is reproduced satisfactorily.

Bond and Angle Vibrations

The remaining bonded parameters (bonds and angles) were optimized to reproduce vibrational frequencies and eigenvector projections derived from QM calculations. For this, we used the Automated Frequency Matching Method (AFMM) (*37*), which optimizes the MM parameter set until the best fit with the QM reference set is obtained. AFMM requires both the eigen frequencies and eigenvectors of the MM set to match the QM data. This is an important aspect of the method, because it avoids incorrect mode matching and misleading reproduction of vibrational frequencies. The resulting plots of the vibrational frequencies obtained with QM and the MM for anisole and PHP are shown in Figures 4a and 4b, respectively. In both model compounds, the MM and QM frequencies match very well, with root mean square deviation of 51.6 cm^{-1} for anisole and 55.6 cm^{-1} for PHP, indicating that the bond and angle parameters are well-optimized.

Table 2. Minimum interaction energies (kcal/mol) and distances (Å) between water:anisole and water:lignin-dimer[a]

	Interaction energies		*Interaction distances*	
Orientations	*QM*	*MM*	*QM*	*MM*
d_0	-4.01	-3.96	2.15	1.82
d_{120}	-3.18	-3.09	2.16	1.87
dimer	-3.93	-4.02	2.10	1.81

[a] QM interaction energies were scaled by 1.16 as described in the text. Orientation geometries considered have the dihedral between the water molecule and the phenyl ring being 0, and 120 degrees, respectively and "dimer" refers to a G and S unit connected with a β-O-4' linkage.

Force Field Validation

In the final part of this work, the parameter set was tested without further adjustment against a condensed-phase experimental property of lignin that was not used during the parameterization. Because of the highly heterogeneous structure of lignin, the most appropriate experimental data to use is the crystal structure of a lignin subunit dimer, erythro-2-(2,6-dimethoxy-4-methylphenoxy)-1-(4-hydroxy-3,5-dimethoxyphenyl) propane-1,3-diol (EPD) (*38*). The chosen compound is very similar to two syringyl units connected with a β-O-4' linkage, but with a methyl group replacing the hydroxyl group of one of the phenol rings. The single crystal X-ray diffraction study revealed a triclinic $P\overline{1}$ structure whose unit cell dimensions are listed in Table 3.

To mimic as closely as possible the conditions under which the experiment was run, the MD simulation was performed for a 4×4×4 unit cell (128 dimers) using periodic boundary conditions while maintaining the temperature and pressure at their experimental values. The unit cell dimensions were allowed to vary during the simulation, and their time averages are listed in Table 4. The MD unit cell dimensions were close to the experimental values, and the system remained triclinic. The unit cell underwent a moderate expansion, with a 5% increase in volume. After aligning the MD coordinates with the experimental structure, the root mean square deviation (RMSD) between the experimental and calculated structure was 0.173±0.033 Å.

In particular, we should also note that the current force field models the β-O-4' linkage that is essential to the conformation of the lignin macromolecule very well. The time averages of the two dihedrals (d_1 and d_2) that define the β-O-4' linkage were compared with the experimental crystal values. The two dihedrals are (numbering scheme in Fig. 1b): $d_1 = C_5\text{-}C_4\text{-}O\text{-}C_8' = 77.9 \pm 6.3°$, compared to the experimental value of 80.0° and $d_2 = C_4\text{-}O\text{-}C_8'\text{-}C_7' = -148.5 \pm 5.5°$, compared to the experimental value of –152.8°. As with previous results, the simulation results agree with the experimental ones.

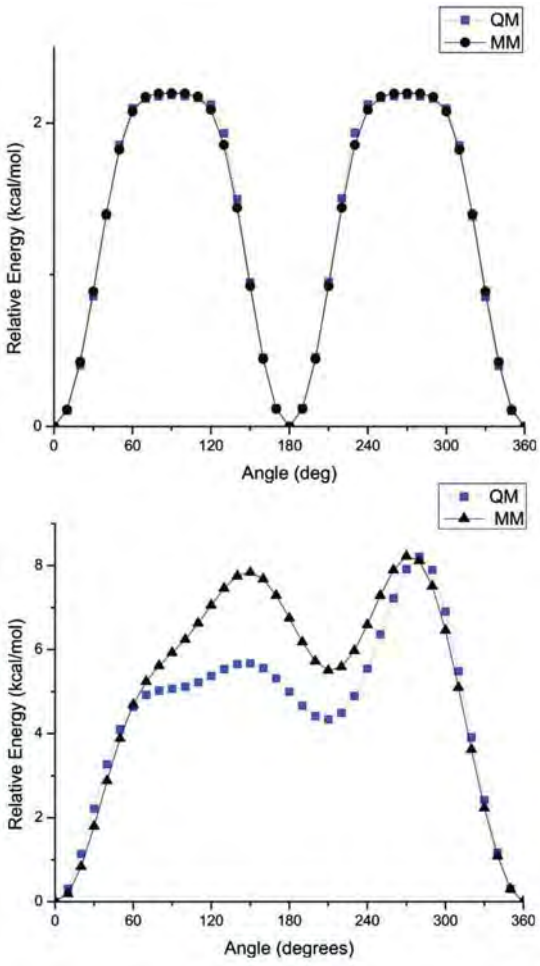

Figure 3. Potential energy profiles for rotation around the (a) $\omega_1 = C_2\text{-}C_1\text{-}O\text{-}C_\alpha$ dihedral of anisole and (b) $\omega_4 = C_1\text{-}C_7\text{-}O_7\text{-}HO_7$ dihedral of anisole. (Reproduced from reference (20). Copyright 2008 Wiley Periodicals, Inc.)

Building Lignocellulose Models

The accurate computer simulation of lignin presents significant challenges. Unlike many biological macromolecules that have been studied with molecular simulation, both the primary and three-dimensional structures of lignins are not known. Hence, a logical strategy is to build random multiple lignin units that have ensemble composition (and linkage) properties consistent with experimentally derived average chemical composition. In particular, emphasis was placed on ensuring that the models accurately represent the lignins found in the cell walls of softwoods. The following paragraphs describe how the atomistic lignin models were built.

We built 26 lignin molecules altogether, each with a distinct primary and tertiary structure. The initial structural models were generated by first deriving the

topology of the molecules and then generating the tertiary structure. To generate molecular topologies, we used a variety of experimental data on lignin composition in softwoods. Softwoods are composed mainly of guaiacyl (G) units (*3*, *22*), so only G units are considered to be present in the model. A typical linkage composition of softwoods is: β-O-4' 50%, 5-5' 30%, α-O-4' 10% and β-5' 10%. Linkages β-O-4', α-O-4', and β-5' contain chiral centers at the β and α-carbons. However, lignins are not found to be optically active (*39*). Hence, the constructed lignin molecules contain equal numbers of left- and right-hand linkages. The molecular weight of lignins is on the order of 10,000 or greater (*40*), and the models have a molecular weight of 13,000. Crosslinks are formed when one unit participates in more than one linkage. Twenty-six lignins were built with varying degrees of crosslinking, but the average crosslink density was chosen to be consistent with the experimental value of 0.052 obtained from spruce wood (*41*). With these experimental data as a guide, random lignin configurations were created using a script written in the program language Python. This method produced 26 molecules that were different to each other, but were all consistent with the experimentally determined properties of softwoods. For example, although all lignins had the same linkage composition, the order of the linkages was different. Furthermore, the number of crosslinks and their positions in individual lignin molecule were also different.

Once the topologies were derived, the 3D structures for lignin molecules were constructed using a step-wise approach. Each new unit was added to the existing structure using the appropriate linkage. As mentioned above, the geometries of all the units and linkages were obtained using high-level quantum chemical calculations. Subsequently, the entire new molecule was minimized using a molecular mechanics force field. The procedure was repeated until the maximum molecular weight of 13,000 was reached. As indicated earlier, our approach, while consistent with the average chemical properties of lignin, is limited by the lack of primary and tertiary structures of these molecules.

In contrast to lignins, the chemical structure of cellulose is known. It is a straightforward process to build cellulose microfibrils using the molecular structure of cellulose Iα (*42*) and Iβ (*43*), obtained with a combination of X-ray and neutron diffraction. In the present model, as in other studies, cellulose is in the Iβ form; and the MD simulation starts with the fibril being a perfect crystal. A preliminary model of cellulose surrounded by lignin molecules in solution is shown in Figure 5. Such initial models can probe the interactions between lignin and cellulose at the atomic level, as well as provide a way to parameterize coarse-grained mesoscale models.

Modeling Cellulosomes: Cohesin-Dockerin Interaction

Insight of Type I Cohesin-Dockerin Recognition from the Crystallographic Structure

The first 1.9-Å crystal structure of the type I cohesin-dockerin complex from the cellulosome of *C. thermocellum* has been determined (*17*) (Figure 6), providing insight into the structure and mechanism by which the cellulosome assembles. The

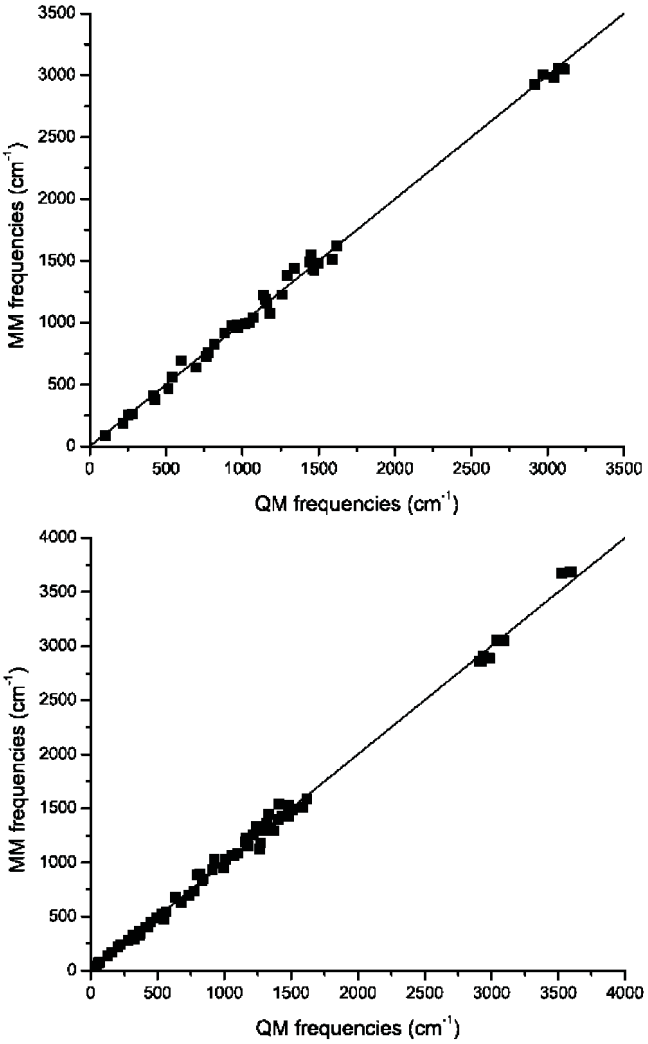

Figure 4. Vibrational frequencies of model compounds anisole (left) and PHP (right). The plotted line shows an ideal fit between QM and MM data. (Reproduced from reference (20). Copyright 2008 Wiley Periodicals, Inc.)

cohesin module forms an elongated, nine-stranded β-barrel in a classical jelly-roll topology with a tightly-packed aromatic/hydrophobic core. The two faces of the β-barrel are composed of strands 5, 6, 3, and 8 on the contact face with the dockerin, and strands 4, 7, 2, 1, and 9 on the opposite face. The dockerin partner of the cohesin-dockerin complex contains a duplicated 22-amino-acid sequence that comprises α-helix 1 and 3 in conformation, respectively. The dockerin structure is organized into two calcium-binding loop-helix motifs separated by a short linker region. Indeed, it has been found that Ca^{2+} plays a key role in maintaining the structural integrity of the cohesin-dockerin complex (*44*, *45*).

Table 3. Unit cell properties of small-molecule-dimer for experimental crystal structure and from molecular dynamics simulation

Cell dimension	*X-ray*	*MD*
A (Å)	8.69	8.73 ± 0.02
B (Å)	8.90	8.93 ± 0.01
C (Å)	13.11	13.68 ± 0.03
α (deg)	73.85	74.48 ± 0.05
β (deg)	86.15	86.30 ± 0.01
γ (deg)	83.06	83.06 ± 0.02
Cell volume (Å^3)	966	1020

Table 4. Dihedrals defining the β-O-4' linkage d_1 = C_5-C_4-O-C_8 and d_2 = C_4-O-C_8-C_7, see Figure 2d

Dihedral	*X-ray*	*MD*
d_1 (deg)	80.0	77.9 ± 6.3
d_2 (deg)	–152.8	–148.5 ± 5.5

The cohesin structure's compact nature, together with the fact that the contact surface features no obvious binding pocket or cleft, suggests that binding between cohesins and dockerins occurs through the exposed surface residues. The cohesin in the type I complex comes into contact with the entire length of α-helix 3, but is only in contact with the C-terminal end of helix 1 from the type I dockerin. The N terminus of helix 1 is diverted away from the cohesin surface. Given the orientation of the dockerin on the cohesin surface and the two-fold structural symmetry within the dockerin domain, Carvalho et al. provided evidence for a dual binding mode of dockerin modules to cohesins (*21*).

The available crystal structures suggests that the cohesin-dockerin association is maintained mainly by hydrophobic interactions, consistent with the negative heat capacity associated with the binding event (*17*, *46*, *47*). The proteins also interact through an extensive hydrogen-bonding network between one face of the cohesin and the corresponding dockerin domain. Several hydrophilic residues play an essential role in recognizing and forming the complex: Arg77, Tyr74, Asp39, Glu86, and Gly89 of the cohesin domain, and Leu22, Arg23, Ser45, Thr46, and Arg53 from α-helices 1 and 3 of the dockerin domain (Figure 6). Biochemical mutagenesis studies have provided complementary clues to the mode of cohesin-dockerin interaction. One of the striking mutations, known to cause recognition failure, is D39N. Asp39 of the cohesin, one of the most conserved residues, is located at the protein-protein interface of the complex. This residue forms direct hydrogen bonds with Ser45 of the dockerin, the most critical residue for domain recognition (*16*, *21*, *48*), and forms water-mediated hydrogen bonds with Val21 and Ile43. It has been shown that the single substitution of Asp39 by a neutrally charged Asn reduces the affinity of the interaction by more than three

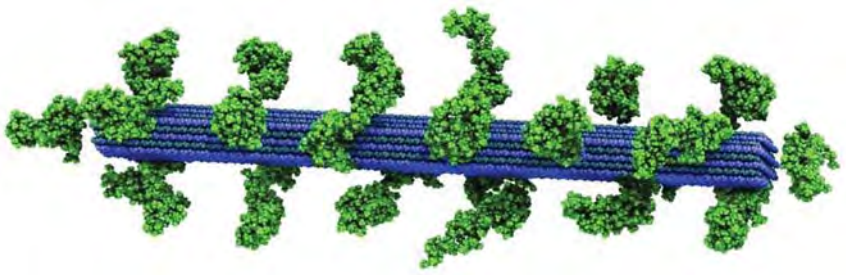

Figure 5. Atomistic model of the plant cell wall components cellulose (blue) and lignin (green).

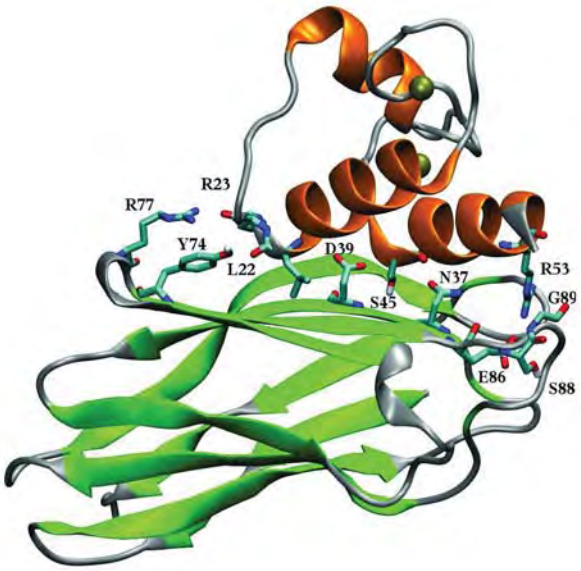

Figure 6. Crystal structure of the cohesin-dockerin complex in cartoon representation with β-sheets (cohesin) in green, α-helices (dockerin) in orange and loop regions in silver. Key residues involved in inter-domain interaction are highlighted in licorice mode, and colored by atom names.

orders of magnitude and disrupts the normal recognition of the dockerin (*49*). Thus, this residue is a hot spot for the cohesin-dockerin interaction. In addition, more recent biophysical and dockerin-mutagenesis experiments have revealed an association constant (K_a) of 8×10^7 M^{-1} for the wild-type cohesin-dockerin complex and the importance of highly conserved Ser45-Thr46 in the Ca^{2+}-binding loop for recognition of type I dockerin (*21*). It has been demonstrated that an alternative binding mode can be achieved by substituting the helix-3 Ser45/Thr46 pair with alanines; and the resultant crystal structure at 2 Å resolution shows that the dockerin module interacts with its cognate cohesin module through the helix-1, in which Ser11 and Thr12 play an equivalent role in binding.

Free Energy Landscape of Cohesin-Dockerin Dissociation

Recognizing Type I cohesins by dockerins is the essential event in assembling individual enzymatic subunits into the cellulosome complex. Even though the crystallographic structure and experimental measurements have provided essential information about the association of cohesins and dockerins, the underlying microscopic dynamic and energetic processes are not directly accessible to experiments. Consequently, aspects of the mechanism governing the assembly of cohesins and dockerins remain uncertain. It is therefore particularly informative to elucidate at the atomistic level the detailed molecular principles upon which the cohesin-dockerin interaction is based.

Understanding the underlying molecular association/dissociation mechanism in terms of structure and dynamic events is facilitated by the knowledge of the free-energy profile for the cohesin-dockerin dissociation. The effective free energy (or the potential of mean force, PMF) of cohesin-dockerin dissociation was estimated from a total of 100 ns MD simulation in bulk solution, using the adaptive biasing force (ABF) method (*50*, *51*) implemented in NAMD (*52*). This method relies upon integrating the average forces acting along a reaction coordinate (ξ) that was constructed from endpoints corresponding to the average, or most probable, configurations from unconstrained MD simulations of initial and final states. The reaction coordinate for the dissociation process was defined by the separation distance between the cohesion and dockerin center-of-masses. The results are shown in Figure 7. Although this simple, low-dimensional, reaction coordinate has not been refined, if properly converged, the PMF from this reaction coordinate gives an upper bound on the barrier and, again if converged, will give a proper free-energy change between the states specified.

The free-energy profile's overall shape along the reaction coordinate shows a general uphill trend, illustrating quantitatively that the cohesin-dockerin complex exhibits a resistance against external forces, and that there is a high affinity for the two domains to remain bound. The global free-energy minimum in the profile appears at a distance separating the centers of mass equal to 22.5 Å, corresponding to the stable bound state with the key residues directly in contact. As the two domains move away from each other, the cohesin-dockerin interactions are progressively disrupted. Initially, this leads to a steep increase of the free energy before reaching the first shoulder at ~ 24 Å, at which point the hydrogen bond Asp39 (OD)-Ser45 (HG) has broken; and residues Asp39 and Ser45 at the interface of the protein complex are no longer in contact (Figure 7b). Another characteristic of the initial dissociation is the flow of water molecules into the binding area, substituting protein residues and forming new hydrogen bonds. The first dissociation step, therefore, corresponds to disrupting the hydrophobic core and overcoming the resistance imposed by the Asp39-Ser45 hydrogen bond. As the two domains move further apart, the free-energy profile reaches the second slight shoulder at ~ 26 Å. Inspection of the simulation trajectory indicates that the second shoulder corresponds to the disruption of the recognition strip interaction with the C-terminal region of α-helix 3, accompanied by the rupture of hydrogen bonds/salt bridges between Arg53 and Glu86 (Figure 7c). In contrast, at this point of the dissociation, the C-terminal of the first α-helix of the dockerin,

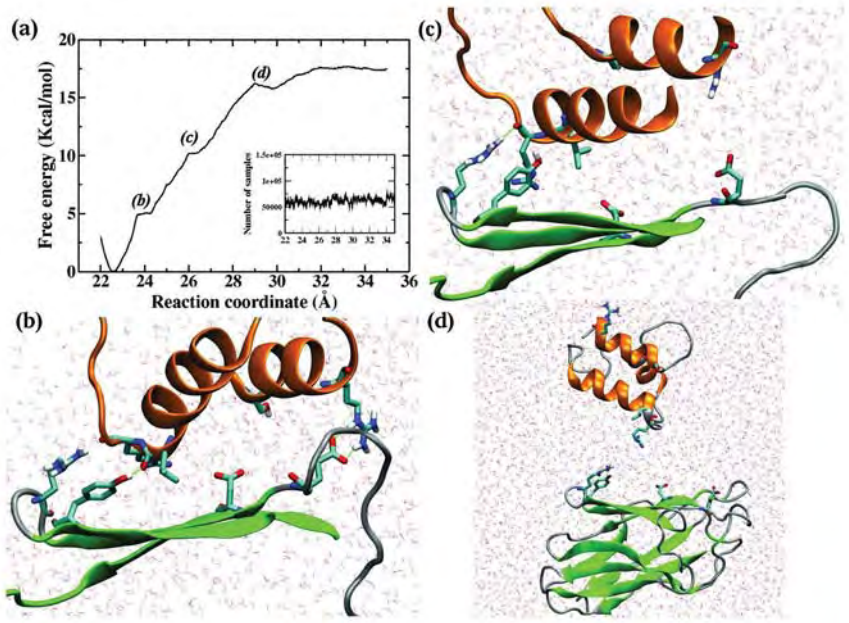

Figure 7. (a) Free-energy profile for the dissociation of cohesin and dockerin domains. The sampling distribution is included in the inset. (b) Snapshot of the cohesin-dockerin complex at ξ = 24 Å; (c) Snapshot at ξ = 27 Å; (d) Snapshot of cohesin-dockerin complex in the dissociated state, i.e., ξ > 30 Å. The two α-helices, β-strands 3, 5, 6, and loop/turn regions are represented in cartoon mode, colored orange, green, and gray, respectively. The rest of the protein structure was omitted for clarity.

and especially the backbone carbon atom of residue Arg23, is still repeatedly in contact with the side chains of the solvent-exposed Arg74 and Tyr77 in the β–strand 5/6 loop at the other end of the β-barrel, with large fluctuations of interatomic distances. The final dissociation of the two subunits corresponds to the shallow well at ~ 30 Å before the PMF eventually becomes nearly flat at > 35 Å (Figure 7d).

The experimental estimate of the overall equilibrium binding constant for the present cohesin-dockerin complex is 8×10^7, corresponding to a free-energy change of about 12 kcal/mol ($\Delta G = -RT\ln K_a$, where R is the gas constant and T is temperature). In the simulations, the overall difference in the calculated free energy between the minimum of the bound state and the barrier is ~ 17 kcal/mol. This agreement is reasonable, given that directly comparing the dissociation free energy with the experimentally determined absolute binding energy requires a knowledge of the contributions which were not considered in this study. The free energy change in the translational and rotational degrees of freedom on complexation was not included. Implementations of free energy algorithms have inherent errors. The sampling errors that may arise from the conformational flexibility of the unbound dockerin domain in solution were also not considered. Some significant extension to the present computational

methodologies is needed to tackle the complex situation in the cohesin-dockerin protein complex. Furthermore, the present study focuses on a detailed view of the underlying mechanism of association and interaction in the cohesin-dockerin complex rather than calculating the absolute binding free energy.

Summary and Outlook

The accurate computer simulation of lignocellulosic biomass materials presents significant challenges. An important first step is the parameterization of a potential energy function for the system. Here, we derived an MM force field for lignin that is compatible with the CHARMM potential energy function. The parameterization was based on reproducing quantum-mechanically derived target data. Special care was taken to correctly describe the most common lignin linkage: the β-O-4' bond. The partial atomic charges of the oxygen and carbon atoms participating in the linkage were derived by examining interactions between a lignin fragment model compound and a water molecule. Dihedral parameters were obtained by reproducing QM potential energy profiles, with emphasis placed on accurately reproducing the thermally sampled low-energy regions. The remaining bond and angle parameters were derived using the AFMM method. To test the validity of the force field, we performed a simulation of a lignin-dimer crystal. The overall good agreement between the structural properties of the simulation and the experiment provide confidence that the force field can be used to simulate biomass. Furthermore, using a large body of experimental data on the average chemical composition of lignin as references, we have also constructed preliminary atomic-detail models of lignin.

Another important area of concentration is unraveling the assembly mechanism of the cellulosome complex using computer simulations. We have calculated the PMF profile for the wild-type cohesion-dockerin dissociation. The PMF reveals a high free-energy barrier and a stepwise pattern for the dissociation process. The sequential dissociation events revealed by the free-energy profile provides evidence that a set of residues lying on the flattened β-sheet surface and in the peripheral loop regions is the main obstacle to dockerin unbinding. Although examination of the crystal structure alone suggests that the formation of the cohesin-dockerin complex involves relatively large surface areas on both partners, our simulation results indicate that specific surface regions play more critical roles than others in forming and maintaining the integrity of the cellulosome complex. In turn, the insight gained from the present simulation can be used to guide protein engineering modifications to alter cohesin-dockerin binding. Efforts are underway to design engineered cellulosomal modules that can conduct more efficient biomass degradation than the corresponding wild-type protein complexes. Both atomic-detail and coarse-grained computer simulations are expected, in conjunction with appropriate biochemical and biophysical experiments (e.g., Hammel et al. 2005) (*53*), to provide a foundation for understanding the principles of domain synergy and cellulosomal activity, thus allowing the rational, structure-based design of improved cellulosomal assemblies for cellulosic ethanol production.

Acknowledgments

The authors acknowledge funding from the BioEnergy Science Center. The BioEnergy Science Center is a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science. Loukas Petridis is supported by DOE Field Work Proposal ERKP704. Xiaolin Cheng is also partly funded by DOE SciDAC. This research was supported in part by the National Science Foundation through TeraGrid resources provided by NICS. Additional support to Jeremy C. Smith is provided by ORNL LDRD fund and DOE INCITE award.

References

1. Lynd, L. R.; Laser, M. S.; Brandsby, D.; Dale, B. E.; Davison, B.; Hamilton, R.; Himmel, M.; Keller, M.; McMillan, J. D.; Sheehan, J.; Wyman, C. E. *Nat. Biotechnol.* **2008**, *26*, 169–172.
2. Himmel, M. E.; Ding, S. Y.; Johnson, D. K.; Adney, W. S.; Nimlos, M. R.; Brady, J. W.; Foust, T. D. *Science* **2007**, *315*, 804–807.
3. Ragauskas, A. J.; Williams, C. K.; Davison, B. H.; Britovsek, G.; Cairney, J.; Eckert, C.A.; Frederick, W. J.; Hallett, J. P.; Leak, D. J.; Liotta, C. L.; Mielenz, J. R.; Murphy, R.; Templer, R.; Tschaplinski, T. *Science* **2006**, *311*, 484–489.
4. Gray, K. A.; Zhao, L. S.; Emptage, M. *Curr. Opin. Chem. Biol.* **2006**, *10*, 141–146.
5. Zhang, Y. H. P.; Ding, S. Y.; Mielenz, J. R.; Cui, J. B.; Elander, R. T.; Laser, M.; Himmel, M. E.; McMillan, J. R.; Lynd, L. R. *Biotechnol. Bioeng.* **2007**, *97*, 214.
6. Cosgrove, D. J. *Nat. Rev. Mol. Cell Biol.* **2005**, *6*, 850–861.
7. Grabber, J. H. *Crop Sci.* **2005**, *45*, 820–831.
8. Chen, F.; Dixon, R. A. *Nat. Biotech*. **2007**, *25*, 759. Reddy, M. S. S.; Chen, F.; Shadle, G.; Jackson, L.; Aljoe, H.; Dixon, R. A. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 16573.
9. Liu, C. G.; Wyman, C. E. *Ind. Eng. Chem. Res.* **2003**, *42*, 5409–5416.
10. Fan, L. T.; Lee, Y. H.; Beardmore, D. R. *Biotechnol. Bioeng.* **1981**, *23*, 419.
11. Chang, V. S.; Holtzapple, M. T. *Fundamental factors affecting biomass enzymatic reactivity*; Humana Press Inc.: Totowa, NJ, 2000.
12. Chen, F.; Dixon, R. A. *Nat. Biotechnol.* **2007**, *25*, 759–761.
13. Doi, R. H.; Kosugi, A. *Nat. Rev. Microbiol.* **2004**, *2*, 541–551.
14. Bayer, E.A.; Chanzy, H.; Lamed, R.; Shoham, Y. *Curr. Opin. Struct. Biol.* **1998**, *8*, 548–557.
15. Gilbert, H. J. *Mol. Microbiol.* **2007**, *63*, 1568–1576.
16. Mechaly, A.; Yaron, S.; Lamed, R.; Fierobe, H. P.; Belaich, A.; Belaich, J. P.; Shoham, Y.; Bayer, E. A. *Proteins* **2000**, *39*, 170–177.
17. Carvalho, A. L.; Dias, F. M.; Prates, J. A.; Nagy, T.; Gilbert, H. J.; Davies, G. J.; Ferreira, L. M.; Romao, M. J.; Fontes, C. M. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13809–13814.

18. Pages, S.; Belaich, A.; Belaich, J. P.; Morag, E.; Lamed, R.; Shoham, Y.; Bayer, E. A. *Proteins* **1997**, *29*, 517–527.
19. Kuttel, M.; Brady, J. W.; Naidoo, K. J. *J. Comput. Chem.* **2002**, *23*, 1236–1243.
20. Petridis, L.; Smith, J. C. *J. Comput. Chem.* **2009**, *30*, 457–467.
21. Carvalho, A. L.; Dias, F. M.; Nagy, T.; Prates, J. A.; Proctor, M. R.; Smith, N.; Bayer, E. A.; Davies, G. J.; Ferreira, L. M.; Romao, M. J.; Fontes, C. M.; Gilbert, H. J. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 3089–3094.
22. Boerjan, W.; Ralph, J.; Baucher, M. *Annu. Rev. Plant Biol.* **2003**, *54*, 519–546.
23. Matthews, J. F.; Skopec, C. E.; Mason, P. E.; Zuccato, P.; Torget, R. W.; Sugiyama, J.; Himmel, M. E.; Brady, J. W. *Carbohydr. Res.* **2006**, *341*, 138–152.
24. Nimlos, M. R.; Matthews, J. F.; Crowley, M. F.; Walker, R. C.; Chukkapalli, G.; Brady, J. V.; Adney, W. S.; Clearyl, J. M.; Zhong, L. H.; Himmel, M. E. *Protein Eng., Des. Sel.* **2007**, *20*, 179–187.
25. Yui, T.; Hayashi, S. *Biomacromolecules* **2007**, *8*, 817–824.
26. Yui, T.; Nishimura, S.; Akiba, S.; Hayashi, S. *Carbohydr. Res.* **2006**, *341*, 2521–2530.
27. Vietor, R. J.; Mazeau, K.; Lakin, M.; Perez, S. *Biopolymers* **2000**, *54*, 342–354.
28. Mazeau, K; Heux, L. *J. Phys. Chem. B* **2003**, *107*, 2394–2403.
29. Besombes, S.; Mazeau, K. *Biopolymers* **2004**, *73*, 301–315.
30. Besombes, S.; Mazeau, K. *Plant Physiology and Biochemistry* **2005**, *43*, 299–308.
31. Besombes, S.; Mazeau, K. *Plant Physiol. Biochem.* **2005**, *43*, 277–286.
32. Besombes, S.; Robert, D.; Utille, J. P.; Taravel, F. R.; Mazeau, K. *Holzforschung* **2003**, *57*, 266–274.
33. MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D.T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B.* **1998**, *102*, 3586–3616.
34. Vorobyov, I.; Anisimov, V. M.; Greene, S.; Venable, R. M.; Moser, A.; Pastor, R. W.; MacKerell, A. D. *J. Chem. Theory Comput.* **2007**, *3*, 1120–1133.
35. Breneman, C. N.; Wiberg, K. B. *J. Comput. Chem.* **1990**, *11*, 361.
36. Chen, I. J.; Yin, D. X.; MacKerell, A. D. *J. Comput. Chem.* **2002**, *23*, 199–213.
37. Vaiana, A. C.; Cournia, Z.; Costescu, I. B.; Smith, J. C. *Comput. Phys. Commun.* **2005**, *167*, 34–42.
38. Langer, V.; Lundquist, K.; Miksche, G. E. *Acta Crystallogr., Sect. E: Struct. Rep. Online* **2005**, *61*, O1001–O1003.
39. Ralph, J.; Peng, J. P.; Lu, F. C.; Hatfield, R. D.; Helm, R. F. *J. Agric. Food Chem.* **1999**, *47*, 2991–2996.

40. Brunow, G.; Kilpelainen, I.; Lapierre, C.; Lundquist, K.; Simola, L. K.; Lemmetyinen, J. *Phytochemistry* **1993**, *32*, 845–850.
41. Yan, J. F.; Pla, F.; Kondo, R.; Dolk, M.; McCarthy, J. L. *Macromolecules* **1984**, *17*, 2137–2142.
42. Nishiyama, Y.; Sugiyama, J.; Chanzy, H.; Langan, P. *J. Am. Chem. Soc.* **2003**, *125*, 14300–14306.
43. Nishiyama, Y.; Langan, P.; Chanzy, H. *J. Am. Chem. Soc.* **2002**, *125*, 9074–9082.
44. Chauvaux, S.; Beguin, P.; Aubert, J. P.; Bhat, K. M.; Gow, L. A.; Wood, T. M.; Bairoch, A. *Biochem. J.* **1990**, *265*, 261–5.
45. Lytle, B. L.; Volkman, B. F.; Westler, W. M.; Wu, J. H. *Arch. Biochem. Biophys.* **2000**, *379*, 237–44.
46. Lytle, B. L.; Volkman, B. F.; Westler, W. M.; Heckman, M. P.; Wu, J. H. *J. Mol. Biol.* **2001**, *307*, 745–53.
47. Spinelli, S.; Fierobe, H. P.; Belaich, A.; Belaich, J. P.; Henrissat, B.; Cambillau, C. *J. Mol. Biol.* **2000**, *304*, 189–200.
48. Schaeffer, F.; Matuschek, M.; Guglielmi, G.; Miras, I.; Alzari, P. M.; Beguin, P. *Biochemistry* **2002**, *41*, 2106–2114.
49. Handelsman, T.; Barak, Y.; Nakar, D.; Mechaly, A.; Lamed, R.; Shoham, Y.; Bayer, E. A. *FEBS Lett.* **2004**, *572*, 195–200.
50. Henin, J.; Chipot, C. *J. Chem. Phys.* **2004**, *121*, 2904–14.
51. Darve, E.; Pohorille, A. *J. Chem. Phys.* **2001**, *115*, 9169–83.
52. Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781–1802.
53. Hammel, M.; Fierober, H. P.; Czjzek, M.; Kurkal, V.; Smith, J. C.; Bayer, E. A.; Finet, S.; Receveur-Brechot, V. *J. Biol. Chem.* **2005**, *280*, 38562–38568.
54. Ralph, J.; Brunow, G.; Harris, P. J.; Dixon, R. A.; Schatz, P. F.; Boerjan, W. *Recent Adv. Polyethenol Res.* **2008**, *1*, 36–66.
55. Davin, L. B.; Lewis, N. G. *Curr. Opin. Biotechnol.* **2005**, *16*, 407–415.