Cellulosic ethanol: progress towards a simulation model of lignocellulosic biomass

Loukas Petridis and Jeremy C Smith

Center for Molecular Biophysics , Oak Ridge National Laboratory, Oak Ridge, TN, USA

E-mail: smithjc@ornl.gov

Abstract. A CHARMM molecular mechanics force field for lignin is derived. Parameterization is based on reproducing quantum mechanical data of model compounds. Partial atomic charges are derived by the examination of methoxybenzene:water interactions. Dihedral parameters are optimized by fitting to critical rotational potentials, and bonded parameters are obtained by optimizing vibrational frequencies and normal modes. The force field is 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 will enable full simulations of lignocellulose.

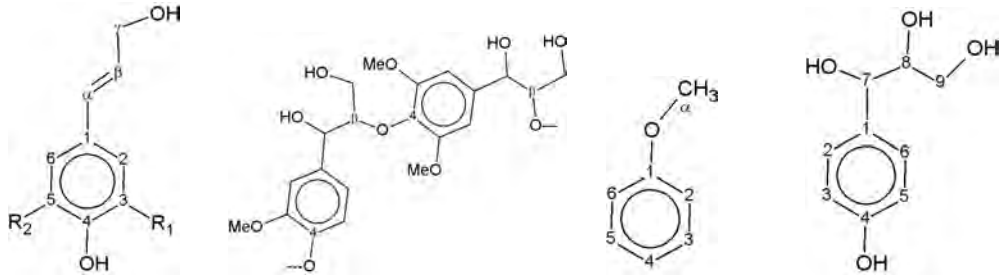
1. Introduction

The plant cell wall is made of cellulose microfibrils embedded in a matrix of polysaccharides (hemicellulose and pectins) and lignin [1]. Plant cell wall structure has come under renewed interest recently in the context of the production of bioethanol from the enzymatic hydrolysis of lignocellulosic biomass [2]. Cellulosic ethanol production is a two-stage process involving first the hydrolysis of cellulose by cellulases to smaller oligosaccharides and then fermentation of the sugars to ethanol. The hydrolysis step is the bottleneck of the process because of the natural resistance, or "recalcitrance," of plant cell walls to degradation [2].

Two main physical factors contribute to biomass recalcitrance [3]. First, cellulose is found in crystalline fibrils the compact structure of which impedes enzymatic access. In comparison, amorphous cellulose is readily digested by enzymes [4]. Second, matrix polysaccharides and lignin coat the cellulose fibril and act as a physical barrier preventing enzymes from reaching the cellulose. Confirmation of the contribution of lignin to biomass recalcitrance is provided by the finding that removing lignin from biomass increases the cellulose-hydrolysis yield from $\sim 20\%$ to 98% [5].

Lignin is a heterogeneous biopolymer found in the plant cell wall; for a review see [6]. It is formed by radical polymerization of three alcohol monolignols: p-coumaryl, coniferyl, and sinapyl, shown in figure 1a. The chemical composition and structure of lignin is highly heterogeneous, varying significantly between plant species and even within different parts of the same plant cell wall. Although complex, lignin is composed primarily of three units p-hydroxyphenyl (H), guaiacyl (G), and syringyl (S), derived by oxidation of p-coumaryl, coniferyl and sinapyl respectively. An illustration of one G and one S units is shown in figure 1b.

Five types of linkages connect the units, leading to the formation of the long lignin biopolymer. The most common linkage (50–80% probability) is β-O-4′, connecting the oxygen of the hydroxyl



(a) monolignols (b) G and S units with β-O-4′ link (c) anisole (d) PHP

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.

on the phenyl ring in one unit with the second tertiary carbon of the other; see figure 1b. Other common linkages are β-β', α-O-4′ and β-5′.

Computer simulation is a powerful tool for complementing experiment in obtaining an understanding of the molecular-level structure and dynamics of lignocellulose. The computational studies of lignin [7, 8] employed the CHARMM27 empirical force field, which was developed to model proteins rather than lignin. In this paper we present the first essential step toward the accurate computer simulation of lignin: the derivation of a molecular mechanics force field, complementing the CHARMM force field. Together with the existing force field for polysaccharides [9], this work will enable full simulations of lignocellulose.

2. Parameterization strategy

In this section we outline the general strategy employed to obtain the force field of lignin. The CHARMM potential energy function [10] of a molecule is approximated by a sum over bonded interactions (bond and angle vibrations and dihedral rotations) and nonbonded interactions (Coulomb electrostatic interactions and the Lennard-Jones potential for the van der Waals interactions). The strength of these interactions is determined by a set of parameters, such as bond, angle and dihedral force constants, partial atomic charges and Lennard-Jones parameters. These empirical parameters are molecule dependent and must be optimized to model the specific molecule prior to performing simulations. This optimization step is referred to as parameterization and its application to the case of lignin is the main task of this work.

This parameterization of lignin follows the main procedure of parameterization of proteins [11] and ethers [12] for the CHARMM force field. Lignin also has a linear ether bond, but different from those examined in [12] in that the oxygen is bonded to a phenyl ring and a tertiary carbon. For this reason it was found necessary to create a new atom type, OET, to represent the lignin's ether oxygen. Parameters were optimized by considering two factors. Firstly, the target data was reproduced as closely as possible. Second, compatibility with the existing CHARMM force field was ensured by restricting optimization to parameters that did not already exist.

Two model compounds were used, since extensive quantum mechanical (QM) calculations on a full lignin dimer are computationally expensive. The first model system, methoxybenzene, also known as anisole (see figure 1c), incorporates the basic features of a 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.

The optimization strategy for the new parameters is summarized below. 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 existing protein [11] and ether [12] CHARMM force fields. Parameterization of the van der Waals parameters for the new atom type, OET, was not deemed necessary since all three ether oxygen types in reference [12] have the same Lennard-Jones parameters. This is a strong indication that the new atom-type in the present work, which is also an ether oxygen, will have the same van der Waals parameters.

Partial atomic charges were optimized by using a supramolecular approach with a model compound (anisole) interacting with one water molecule. The charges on the new atom type OET and its adjacent carbons were optimized by examining the minimum interaction energies and distances between anisole and a water molecule. The convergence criterion for charges was a $< 5\%$ difference between QM and molecular mechanics (MM) energies. The HF/6-31G* energy was computed varying only the inter-molecular distance so as to locate the minimum-energy separation and interaction energy. Anisole and water were fixed at their MP2 optimized and TIP3P [13] geometries, respectively. For comparison between the QM and MM data, the QM energies were scaled by a factor of 1.16 so as to correct for the absence of polarization and attractive Lennard-Jones contributions in the Hartree-Fock calculations, but no correction was made for basis set superposition error [12, 14].

After completing the nonbonded interactions, parameters for dihedral rotations were deduced from QM (MP2/6-31G*) potential energy surfaces. In cases where QM surfaces were well reproduced, the value of the force constant that gave the smallest error was chosen. In the cases where the shapes were not as well reproduced emphasis was placed on the low-energy region < 3 kcal/mol.

The remaining bond and angle parameters were derived by reproducing QM vibrational frequencies and normal modes with the automated frequency matching method (AFMM) [15]. The DFT/B3LYP/SBKJC/6-31G* level of theory was used and a frequency scaling factor of 0.9614 was used to compensate for the use of the harmonic approximation to the potential energy surface [16]. Good starting values of force constants were provided for AFMM by considering similar existing parameters in ethers, phenol and alcohols. Bond and angle optimization was repeated until the merit function (equation 1) reached a value of less than 60 cm^{-1}. The iterative procedure was repeated until all convergence criteria were satisfied.

3. Parameter optimization

3.1. Partial atomic charges

The partial charges were adjusted so as to reproduce minimum distances and interaction energies between anisole and a water molecule (see, for example, [17]). The use of this supramolecular approach ensures that the derived charges take into account effects of polarization. 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 lone pair of the ether oxygen. A list of all final atomic charges is shown in table 1.

In order 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 shows the comparison of the MM and QM interaction energies and distances, which were used to optimize the anisole charges. The empirical calculations reproduce well 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 obtain good condensed phase properties. In the previous general force field for ethers a similar behavior was observed with a

Table 1: A list of the anisole atoms with their respective charges. Atom names refer to figure 1c and atoms types follow the CHARMM27 force field, with the new atom type labeled as OET.

Atom Name	Atom Type	Charge
C_α	CT3	-0.060
$H_{\alpha 1}$, $H_{\alpha 2}$, $H_{\alpha 3}$	HA	0.090
O	OET	-0.280
C_1	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

~ 0.3Å difference between QM and MM [12].

A further calculation was performed to ensure that the partial atomic charges of table 1, which were derived using a model compound, can be transferred to lignin. The minimum interaction energies and distances between a whole lignin dimer (G and S units connected with a β-O-4′ linkage), shown in figure 1b, and a TIP3P water molecule were obtained without further refinement of the parameters. This supramolecular complex is labeled as "dimer" in table 2. As in the case of anisole and water, the agreement between the QM and MM interaction energies was excellent, further justifying the use of the charges in Table 1 for the β-O-4′ lignin linkage. Furthermore, the lignin dimer had two methoxy groups at position 3 and 5 of the phenol ring, indicating that the presence of the methoxy group does not alter the charge of the ether oxygen.

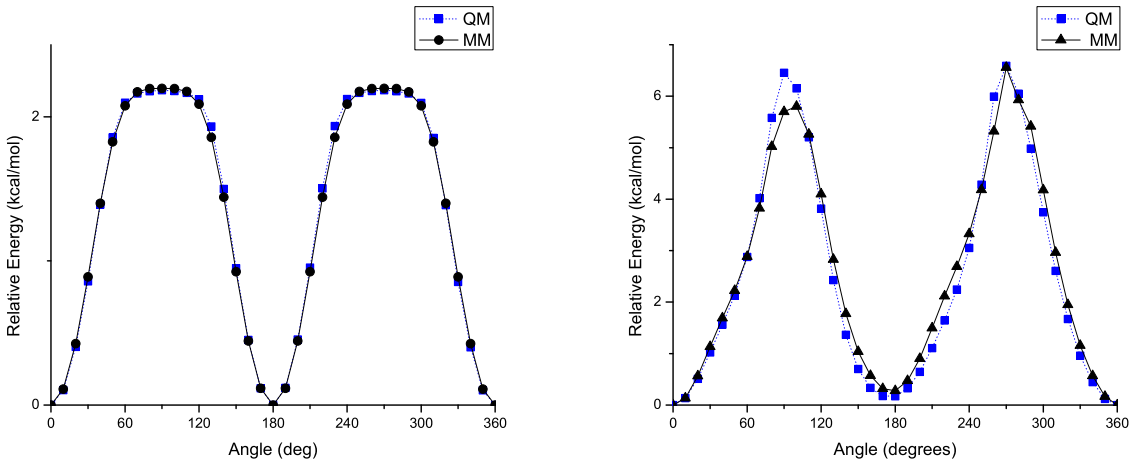
3.2. Bonded parameters

3.2.1. Dihedral rotations Dihedral rotations around the β-O-4′ linkage play a significant role in determining the configuration of the lignin macromolecule, and therefore special care was taken to obtain good parameters for the equivalent dihedrals of the model system. These are ω_1: X-C_1-O-C_α, where X refers to wild atom types, and ω_2: C_1-O-C_α-H. The optimization was based on reproducing quantum-chemically obtained adiabatic energy surfaces, where the selected dihedral (ω_1 or ω_2) is held constant while the remaining degrees of freedom are allowed to relax to a constrained energy minimum. An example of such potential energy surfaces is shown in figure 2a, where it can be seen that the MM surfaces closely follow the target QM data.

The remaining dihedral parameters of lignin that do not involve the ether oxygen were deduced from the more complex rotational potential energy profiles of the second model compound, PHP. Four dihedrals were examined: ω_3=C_2-C_1-C_7-X, ω_4=C_1-C_7-O_7-H_{O7}, ω_5=C_1-C_7-C_8-X, and ω_6=X-C_8-C_9-X. All other dihedral parameters were obtained from the existing CHARMM force field. To accurately model a specific dihedral potential (e.g., ω_3), it was found

Table 2: Minimum scaled interaction energies (kcal/mol) and distances (Å) between water:anisole (d_0 and d_{120}) and water:lignin-dimer (dimer).

Geometry	Interaction Energy		Interaction Distance	
	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) ω_1 (b) ω_3

Figure 2: Potential energy for rotation around the (*a*) ω_1= C_2-C_1-O-C_α dihedral of anisole (*b*) ω_3=C_2-C_1-C_7-X dihedral of PHP.

to be necessary to constrain the remaining three (ω_4, ω_5 and ω_6) to their equilibrium values.

An example of the resulting energy surfaces is shown in figure 2b. Although the agreement between the QM and MM data is not perfect, the rather complex shapes are reproduced satisfactorily. During parameterization, special care was taken to reproduce as accurately as possible the low energy ($\leq$ 3 kcal/mol) regions since it is these regions that will be thermally most-frequently sampled.

3.2.2. 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 the automated frequency matching method [15] (AFMM) was employed, which optimizes the MM parameter set until the best fit with the QM reference set is obtained. AFMM requires both the eigenfrequencies and eigenvectors of the MM set to match with the QM data. This is an important aspect of the method, since it avoids incorrect mode matching and thus misleading reproduction of vibrational frequencies. The merit function is given by

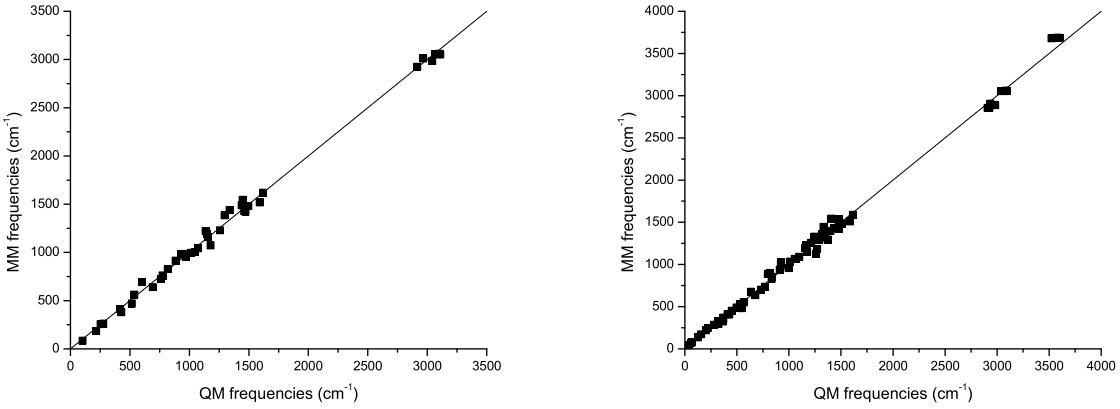
$$\sigma^2 = \frac{\sum_i^{3N-6} (\omega_i^{qm} - \omega_i^{mm})^2}{3N-6}, \quad (1)$$

where ω^{qm} and ω^{mm} refer to the vibrational frequencies obtained with QM and the MM methods respectively and N is the number of atoms of the molecule.

The resulting plots of ω^{mm} against ω^{qm} for ansisole and PHP are shown in figure 3 and indicate satisfactory matching between MM and QM modes. After optimization the root mean square deviation from the reference set was found to be $\sigma = 51.6\ cm^{-1}$ for anisole and $\sigma = 55.6\ cm^{-1}$ for PHP, similar in range to previous parameterizations involving AFMM, which quote values of $\sigma = 47 - 94\ cm^{-1}$ [19, 20]. A list of the bonded parameters of lignin can be found in [18].

3.3. 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.



(a) anisole (b) PHP

Figure 3: Vibrational frequencies of model compounds anisole and PHP. The plotted line shows the ideal fit between QM and MM data

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) [21]. The single crystal -ray diffraction study revealed a triclinic $P\bar{1}$ structure whose unit cell dimensions are listed in table 3. In order to mimic as closely as possible the conditions under which the experiment was run, the MD simulation was performed for 64 unit cells using periodic boundary conditions while keeping the temperature and pressure fixed at their experimental values.

The MD unit cell dimensions were allowed to vary during the simulation and their time averages are shown in table 3. The cell dimensions were found to be 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 between the experimental and calculated structure was found to be 0.173 ± 0.033 Å.

It is also of particular importance that the current force field models well the β-O-4′ linkage that plays an important role in the conformation of lignin. For this reason the time average of the two dihedrals, d_1 and d_2, that define the β-O-4′ linkage was compared with the experimental crystal values. The two dihedrals are (numbering scheme in Fig. 1d): d_1=C_5-

Table 3: Unit cell properties of small-molecule-dimer from crystal structure and MD simulation.

Cell Dimension	Experiment	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

C_4-O-C_8' $= 77.9^o \pm 6.3$, compared to the experimental value of 80.0^o and d_2= C_4-O-C_8'-C_7' $= -148.5^o \pm 5.5$, compared to the experimental value of -152.8^o. As with previous results, the simulation results are in accord with experiment.

4. Discussion

This work presents a molecular mechanics 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 charge 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 reproducing accurately the thermally sampled low energy regions. The remaining bond and angle parameters were derived using the AFMM method. In order to test the validity of the force field a simulation of a lignin-dimer crystal was performed. The overall good agreement between the structural properties of the MD run and the experiment provide confidence that the force field can be used in simulation of biomass.

The accurate computer simulation of lignin in lignocellulose will present significant challenges. Unlike many biological macromolecules that have been studied with molecular simulation, both the chemical and three-dimensional structures of lignin are relatively poorly researched. However, the present force field provides a basis for constructing molecular models of lignin systems, and, in combination with a range of biophysical measurements, significant progress in determining structures of lignocellulosic biomass can be expected in the near future.

Acknowledgments

We acknowledge funding from the U.S.A. Department of Energy (DOE) Bioenergy Science Center and DOE Field Work Proposal ERKP704.

References

[1] Cosgrove D J 2005 *Nat. Rev. Mol. Cell Biol.* **6** 850–861
[2] Himmel M E, Ding S Y, Johnson D K, Adney W S, Nimlos M R, Brady J W and Foust T D 2007 *Science* **315** 804–807
[3] Zhang Y H P, Ding S Y, Mielenz J R, Cui J B, Elander R T, Laser M, Himmel M E, McMillan J R and Lynd L R 2007 *Biotechnol. Bioeng.* **97** 214–223
[4] Fan L T, Lee Y H and Beardmore D R 1981 *Biotechnol. Bioeng.* **23** 419–424
[5] Mooney C A, Mansfield S D, Touhy M G and Saddler J N 1998 *Bioresour. Technol.* **64** 113–119
[6] Davin L B and Lewis N G 2005 *Curr. Opin. Biotechnol.* **16** 407–415
[7] Besombes S and Mazeau K 2005 *Plant Physiol. Biochem* **43** 299–308
[8] Besombes S and Mazeau K 2005 *Plant Physiol. Biochem* **43** 277–286
[9] Kuttel M, Brady J W and Naidoo K J 2002 *J. Comp. Chem.* **23** 1236–1243
[10] Brooks B R, Bruccoleri R E, Olafson B D, States D J, Swaminathan S and Karplus M 1983 *J. Comp. Chem.* **4** 187–217
[11] MacKerell A D *et al* 1998 *J. Phys. Chem. B* **102** 3586–3616
[12] Vorobyov I, Anisimov V M, Greene S, Venable R M, Moser A, Pastor R W and MacKerell A D 2007 *J. Chem. Theory Comput.* **3** 1120–1133
[13] Jorgensen W L, Chandrasekhar J, Madura J D, Impey R W and Klein M L 1983 *J. Chem. Phys.* **79** 926–935
[14] Mackerell A D and Karplus M 1991 *J. Phys. Chem.* **95** 10559–10560
[15] Vaiana A C, Cournia Z, Costescu I B and Smith J C 2005 *Comput. Phys. Commun.* **167** 34–42
[16] Scott A P and Radom L 1996 *J. Phys. Chem.* **100** 16502–16513
[17] Chen I J, Yin D X and MacKerell A D 2002 *J. Comp. Chem.* **23** 199–213
[18] Petridis L and Smith J 2008 *J. Comp. Chem.* (to appear)
[19] Cournia Z, Smith J C and Ullmann G M 2005 *J. Comp. Chem.* **26** 1383–1399
[20] Vaiana A C, Schulz A, Wolfrum J, Sauer M and Smith J C 2003 *J. Comp. Chem.* **24** 632–639
[21] Langer V, Lundquist K and Miksche G E 2005 *Acta Crystallogr. Sect.* **E***: Struct. Rep. Online* **61** O1001–O1003