High-performance computer simulation in the energy biosciences

Jiancong Xu,1,2 Mike F Crowley,2,3 Loukas Petridis,1,2 Hao-Bo Guo,1 Hong Guo,1,4 Marimuthu Krishnan,1 Benjamin Lindner,1,4 Anne O Summers,5 Susan M Miller,6 Liyuan Liang,7 Jerry M Parks,1 Roland Schulz,1,4 Moumita Saharay,1,2 Xiaolin Cheng1,2 and Jeremy C Smith2,4,*

1Center for Molecular Biophysics, Building 6011, Oak Ridge National Laboratory, Oak Ridge, TN, 37830, USA
2BioEnergy Science Center
3Chemical and Biosciences Center, National Renewable Energy Laboratory, 1617 Cole Blvd, Golden, CO, 80401-3393, USA
4Department of Biochemistry & Cellular & Molecular Biology, University of Tennessee, Knoxville, TN 37966, USA
5Department of Microbiology, University of Georgia, Athens, Georgia, 30602-2605
6Department of Pharmaceutical Chemistry, University of California San Francisco, 600 16th Street, San Francisco, California, 94158
7Environmental Science Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, 37831

Email: *smithjc@ornl.gov

Abstract. An overview is provided of the use of high-performance computing for molecular simulations in the energy biosciences. Examples are given of research in cellulosic ethanol production, involving the construction of models of lignocellulosic biomass and the characterization of hydrolytic enzyme complexes. The use of quantum mechanical methods in determining enzyme reaction mechanisms is discussed. The synergy between neutron scattering and computer simulation is examined. Finally, MD scaling results on the ORNL Jaguar Cray XT5 supercomputer are given.

1. Introduction

High-performance computer simulation has a significant role to play in the energy biosciences in obtaining an understanding of physical processes leading to biological function. Molecular mechanical and quantum mechanical techniques can provide atomic detailed insight into processes at the core of research into bioenergy, bioremediation, carbon capture, neutron scattering and other critical research missions.

Here we will provide a synopsis of some current projects at Oak Ridge. Work is in progress to derive computational simulation models of use in cellulosic ethanol production and models are being constructed of heterogeneous lignocellulosic biomass, an effort that will hopefully shed light on the phenomenon of ‘biomass recalcitrance,’ or resistance, to hydrolysis into sugars, which is a bottleneck in biofuel production. Further biofuel research involves simulation work on the functioning of microbial enzymes and enzyme complexes that hydrolyse cellulose chains. The fate of mercury in
streams contaminated from DOE sites is strongly influenced by bacterial enzymes, and we outline research aimed at understanding how these enzymes function. Neutron scattering will play a significant role in the energy-related materials and biosciences, and the synergy between neutron scattering and HPC is described. Finally, we present some scaling results for MD of large systems on a petascale supercomputer.

2. Bioenergy Research

2.1. Lignocellulosic Biomass

This section outlines the construction of an all-atom computational model of lignocellulose, a complex material found in plant cell walls. Lignocellulosic biomass is composed of crystalline cellulose microfibrils laminated with hemicellulose, pectin, and lignin polymers [1].

Initial work has focused on cellulose and lignin. Cellulose is the unbranched $\beta$-(1–4)-linked polymer of D-glucopyranose [2]. It exists as an array of many parallel, oriented chains, organized into crystalline microfibrils [3]. The glucan chain length varies from about 2000 to more than 15,000 glucose residues. The organization of cellulose can vary from elementary fibrils in plants, which contain approximately 36 chains, to large macrofibrils of cellulosic algae, which contain more than 1200 chains. Lignins are heterogeneous branched biopolymers [4]. 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. Unlike the other components of the plant cell wall, lignins do not have a well-defined primary structure.

In recent years there has been a revived interest in lignocellulose structure, due to its immense abundance in nature offering a potentially cheap sugar source for industrial bioethanol production [5]. However, the conversion efficiency of lignocellulose into ethanol in integrated enzyme-based industrial processes is low. The bottleneck in the process is the hydrolysis of biomass into sugars; the intermolecular interactions between the lignocellulosic components may prevent efficient decomposition [5–7]. Therefore, understanding the physical properties of this biomass is crucial for overcoming the major technological challenge in the development of viable cellulosic bioethanol.

Computer simulation is a powerful tool for complementing experiment in obtaining an understanding of the molecular-level structure and dynamics of lignocellulose. Molecular Dynamics (MD) simulation may provide valuable insights on the molecular-level structure, dynamics and energetics of this complex biomass system.

2.1.1. A molecular mechanics force field for lignin. In recent work [8] 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 [9], this force field will enable full simulations of lignocellulose.

We outline the general strategy employed to obtain the lignin force field. The CHARMm potential energy function [10] of a molecule is given by the following equation:

$$
E = \sum_{\text{bonds}} K_b (b - b_o)^2 + \sum_{\text{angles}} K_\theta (\theta - \theta_o)^2 + \sum_{\text{U-B}} K_{ub} (s - s_o)^2 + \sum_{\text{dihedrals}} K_\phi [1 + \cos(n\phi - \delta)] + \sum_{\text{impropers}} K_\psi (\psi - \psi_o)^2 + \sum_{\text{non-bonded}} \left\{ \epsilon_{ij} \left[ \frac{R_{ij}}{r_{ij}} \right]^{12} - \left[ \frac{R_{ij}}{r_{ij}} \right]^{6} \right\} + \frac{q_i q_j}{4\pi\epsilon_0}
$$

where the contributions to the energy include bonded (bond, angle, Urey-Bradley, dihedral and improper dihedrals) and non-bonded (Lennard-Jones 6–12 potential for the van der Waals interactions, and Coulomb electrostatic) terms. The parameters, including the force constants $K$ and partial atomic charges $q$, are molecule-dependent and must be derived prior to performing simulations. The parameterization of the force field for lignin was the main task of Ref. [8].
This parameterization of lignin followed the main procedure of parameterization of proteins [11]. Two model compounds were used to save computational time (see Figure 1): methoxybenzene, and p-hydroxyphenyl (PHP), the simplest lignin unit. Parameters were optimized by considering two factors. Firstly, the ‘target data’ were reproduced as closely as possible. Effectively this ensures that the force field describes accurately specific properties of lignin. Secondly, compatibility with the existing CHARMM force field was ensured by restricting optimization to parameters that did not already exist in the force field.

The optimization strategy for the new parameters is summarized in the diagram in Figure 2. Equilibrium values for bond lengths, angles and dihedrals were taken from high level (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. Charges were optimized with respect to the QM interaction energies using a supramolecular approach of methoxybenzene interacting with one water molecule. The partial charges were adjusted so as to reproduce minimum distances and interaction energies between anisole and water. 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. The empirical calculations were found to reproduce well the scaled QM interaction energies, with the error being less than 3%. After completing the non-bonded interactions, parameters for dihedral rotations were deduced from QM potential energy surfaces. The optimization was based on reproducing quantum-chemically obtained adiabatic energy surfaces, where the selected dihedral is held constant while the remaining degrees of freedom are allowed to relax to a constrained energy minimum. 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 [13] was employed, which optimizes the MM parameter set until the best fit with the QM reference set was obtained. Finally, the parameter set was tested without further adjustment against condensed phase experimental properties of lignin that were not used during the parameterization.

2.1.2. Lignocellulose models. The next step towards a realistic simulation of lignocellulosic biomass is the construction of an atomistic model. This is performed by first generating models of individual components, such as lignin and cellulose (hemicellulose and pectins will be incorporated in the model in the future) and then combining these components to form supramolecular assemblies.
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. However, there is abundance of information on lignin composition. Lignins are composed primarily of three units p-hydroxyphenyl (H), guaiacyl (G), and syringyl (S). Softwoods, in contrast, contain mainly only G units [4]. There are various linkages that connect the units, leading to the formation of the branched lignin biopolymer. The typical linkage composition of softwoods is β-O–4' 50%, 5–5' 30%, α-O–4' 10% and β–5' 10%. Crosslinks are formed when one unit participates in more than one linkage. The crosslink density of spruce wood lignins has been deduced via calculations on extractive delignification [14] and was found to be 0.052. Finally, lignins are known to be optically inactive [15].

The lignin models were constructed by first deriving the topology of the molecules and then generating their 3D structures (Figure 3). With the above experimental data as a guide, random configurations of lignins were created producing 26 molecules with different topologies, but all consistent with experimentally-determined properties of softwood. For example, although all lignins had the same linkage composition, the order of the linkages was different. Furthermore, the number of crosslinks varied between zero and six and their positions along the lignin molecule were also different. Once topologies were derived, the 3D structures for the lignin molecules were constructed. A step-wise approach was used, where each new unit is first added to the existing structure using the appropriate linkage and subsequently the entire new molecule was optimized using the molecular mechanics function.

In contrast to lignins, structures of cellulose microfibrils [16] or and Iβ qre known [3]. In the present model, as in other studies [2,17], cellulose is in the Iβ form and the MD simulation starts from a configuration in which the fibril is a perfect crystal. The final step in the generation of an initial model of lignocellulose is to surround the cellulose fibril with the lignin molecules.
2.2. Towards Structure-Based Cellulosome Design for Cellulosic Ethanol

One promising avenue for overcoming biomass recalcitrance is to understand and modify the properties of bacterial cellulosomes [18]. The cellulosome is a large extracellular multienzyme complexes produced by anaerobic bacteria for the efficient degradation of plant cell-wall polysaccharides [19]. The cellulosome complex consists of various kinds of enzymes arranged around a non-catalytic scaffolding protein that enables the complex to adhere to the biomass and bacterial cell surface. The most extensively studied cellulosome system belongs to *Clostridium thermocellum*, in which the incorporation of the cellulosomal enzymes into the complex is accomplished by the high affinity interaction between two complementary modules: the type I cohesin modules within the scaffolding protein and the type I dockerin domain carried by cellulosomal enzymes.

The crystal structure of the type I cohesin-dockerin complex from *C. thermocellum* has been determined [20] (Figure 4). The cohesin domain exhibits an elongated nine-stranded β-barrel in a classical jelly-roll topology, while the dockerin partner is organized into two symmetrically oriented calcium-binding loop-helix motifs. The cohesin-dockerin association is maintained mainly by hydrophobic interactions, augmented through an extensive hydrogen-bonding network between one face of the cohesin and the corresponding dockerin domain. A number of hydrophilic residues play an essential role in the recognition and formation of 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 4).

2.2.1. Atomistic molecular dynamics simulations of cohesin-dockerin complexation. Recognition of Type I cohesins by dockerins is the determining event in assembly of individual enzymatic subunits into the cellulosome complex. Although 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

---

**Figure 3.** Atomistic model of lignocellulose of softwoods. The cellulose elementary microfibril is shown in blue and the 26 lignin molecules are shown in green. The total system size in 3.3 million atoms (water molecules not shown for image clarity).

**Figure 4.** 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.
accessible to experiments. It is therefore particularly informative to elucidate computationally the detailed molecular principles upon which the cohesin-dockerin interaction is based at the atomic level.

The understanding of the underlying molecular association/dissociation mechanism in terms of structural and dynamical events is facilitated by the knowledge of the free energy profile for the WT cohesin-dockerin dissociation. The free energy of cohesin-dockerin dissociation was estimated from a total of 100 ns MD simulation in bulk solution, using the adaptive biasing force (ABF) method [21,22] implemented in NAMD [23]. This method relies upon the integration of the average force acting on the reaction coordinate (ξ) obtained from unconstrained MD simulations. The free energy profile was then obtained by allowing the two domains to diffuse reversibly along the relative center-of-mass reaction coordinate. The results are shown in Figure 5.

![Figure 5](image)

**Figure 5.** (a) Free energy profile for the dissociation of cohesin and dockerin domains, (b) ξ = 24 Å, and (c) ξ = 27 Å; (d) ξ > 30 Å.

The overall shape of the free energy profile along the reaction coordinate exhibits 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. As the two domains move away from the stable bound state (22.5 Å), the cohesin-dockerin interactions are progressively disrupted. Firstly, this leads to a steep increase of the free energy before reaching the first shoulder at ~24 Å, at which point residues Asp39 and Ser45 at the interface of the protein complex are no longer in contact (Figure 5b) and water flows into the binding area to fill up the space. As the two domains move further apart, the free energy profile reaches the second slight shoulder at ~26 Å. Inspection of the simulation trajectory indicated 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 5c). In contrast, at this point of the dissociation the C-terminal of the first α-helix of the dockerin is still repeatedly in contact with the side chains of the solvent-exposed residues in the β-strand 5/6 loop at the other end of the β-barrel. The ultimate dissociation of the interactions corresponds to the shallow well emerging at ~30 Å before the PMF eventually becomes nearly flat at >35 Å (Figure 5d).
The sequential events of interdomain hydrogen-bond rupture and the step-by-step pattern of the cohesin-dockerin dissociation revealed by the free energy calculations provide evidence for a mode of binding involving both α-helices in the dockerin and the corresponding surface region from its cohesin partner: the C-terminal end of dockerin helix 1 interacting with the β-strand 5/6 loop, while the N-terminus diverted away from the cohesin surface; the N-terminus of helix 2 covering the core fragment of the β-sheet interface and the C-terminus interacting with the recognition strips. The cohesin-dockerin binding may therefore take place in a cooperative manner. Although examination of the crystal structure alone suggests that the formation of the cohesin-dockerin complex involves relatively large surface areas on both partners, the present results show that specific surface regions play more critical roles than others in forming and maintaining the integrity of the cellulosome complex.

2.3. Cellulases
Cellulases are a family of enzymes that catalyze the hydrolysis of celluloses, the main component of the plant cell wall, thus converting the polysaccharides to sugars that can be further fermented to renewable energy sources such as ethanol and other chemicals [24–26]. Three important types of enzymes are included in this family, endoglucanases that cleave glycosidic bonds of cellulose by producing chain ends, exoglucanases that break a single fiber into smaller sugars, and β-glucosidases that hydrolyse simple sugars [27]. Depending on their catalytic mechanisms, cellulases can be further classified into two categories, i.e., the inverting enzymes that invert the chirality of the anomeric carbon, and the retaining enzymes that retain the anomeric configuration [28–30]. A number of cellulase crystals or solution structures are available in the Protein Data Bank [31]. The detailed catalytic mechanisms of cellulases, however, have yet to be elucidated. Computational methods are of importance in meeting the challenge of understanding the mechanistic details of the biodegradation of crystalline cellulose and plant biomass.

Quantum mechanical/molecular mechanical (QM/MM) simulation methods [32] are an effective way to characterize enzyme reactions. QM/MM can also be applied to study the dynamic properties of macromolecules and energetics along the reaction pathways. In QM/MM only the reaction center is considered quantum mechanically, the rest being treated with MM.

Figure 6 shows the Michaelis complex structure of a retaining cellulase *Humicola grisea* Cel12A with a cellulose tetramer (cellotetraose) bound to the –2 to +2 subsites (pdb code 1W2U, resolution 1.52 Å) [33]. The glucose unit at subsite –1 is believed to undergo conformational change upon binding to the enzyme and during the catalytic reaction [34,35]. The two steps in the retaining reaction mechanism involve glycosylation and deglycosylation. QM/MM MD and free energy simulations can provide information on both the conformational changes of the glucose rings along the reaction pathway and the reaction steps, together with the free energy as a function of the reaction coordinate. Moreover, the roles of active site residues can be explored through mutagenesis studies.

3. Bioremediation: Mechanisms of Bacterial Mercury Resistance
Mercury is a toxic and highly reactive element that has no known biological function. It is readily interconverted between its elemental Hg(0) form and more toxic Hg(II) species through various microbial and abiotic pathways [36]. Mercury-resistant bacteria are able to transform inorganic and organic Hg(II) to less toxic Hg(0) via proteins and enzymes of the *mer* operon. The narrow-spectrum *mer* operon includes a regulatory protein, MerR, a periplasmic transporter, MerP, one or more

Figure 6. Structure of H. grisea Cel12A (ribbons) in complex with a cellotetraose (spheres), occupying the –2 to +2 sites at the substrate binding cleft, pdb code 1W2U.
transmembrane Hg(II) transporters (MerT, MerC), and a mercuric reductase, MerA. Additionally, broad-spectrum mercury resistant bacteria include an organomercurial lyase, MerB.

3.1. Conformational Changes Upon Hg(II)-Binding in MerR
MerR, the first protein identified in a family of transcriptional activators known as the MerR family, is the key regulator of the mer operon [37,38]. MerR activates transcription of other mer proteins in the presence of Hg(II) in nanomolar concentrations, and represses transcription in its absence. The functional form of MerR is a homodimer that consists of a DNA binding domain, a helix-turn-helix motif, and a long, antiparallel coiled-coil. Although MerR contains two binding sites, only a single Hg(II) cation is bound at a time [38]. MerR is known to bind Hg(II) in a trigonal planar geometry in which three Cys residues coordinate to Hg(II) [39]. No conventional NMR structures have been determined for MerR thus far, due to its tendency to aggregate [40]. Although other MerR family proteins have been crystallized [41], no X-ray structures have been solved yet for MerR. At present, the detailed mechanism of allosteric transcriptional activation upon Hg(II) binding is not well understood.

Initially, MerR is bound to its operator DNA in complex with RNA polymerase (RNAP) such that the DNA is bent away from and inaccessible to RNAP [42]. Binding of Hg(II) induces an allosteric conformational change that enables transcription of all mer proteins. In an effort to determine the conformational changes of MerR upon Hg(II) binding, we are performing MD simulations on homology-modelled structures of Hg(II)-bound and Hg(II)-free MerR. The dynamic structural models are validated by comparison with small-angle X-ray scattering (SAXS) data in solution, which provides the molecular envelope shape, the pair distribution function, and the radius of gyration.

3.2. Enzymatic Mechanism of Hg-C Cleavage Catalyzed by the Organomercurial Lyase MerB
Methylmercury, \([\text{CH}_3\text{Hg(II)}]^+\), is a particularly toxic organomercurial species that bioaccumulates in living organisms [43]. Its extremely high affinity for thiols and its ability to cross the blood-brain barrier are major contributing factors to its toxicity. As a result, methylmercury degradation is an important topic in bioremediation. The organomercurial lyase, MerB, catalyzes the cleavage of Hg-C bonds in organomercurials to yield Hg(II) and a hydrocarbon (Scheme 1). A separate enzyme, the NADPH-dependent flavoenzyme, MerA, then catalyzes the reduction of Hg(II) to Hg(0).

\[
[R-\text{Hg(II)}]^+ \xrightarrow{\text{MerB}} [\text{Hg(II)}]^{2+} + R-H
\]

Scheme 1. Reaction catalyzed by MerB (R = alkyl, aryl).

MerB functions as a monomer, and no cofactors or other prosthetic groups are involved in the reaction, but free thiols such as cysteine or glutathione are required for enzymatic turnover [44]. The active site of MerB contains two essential Cys residues [45] and an Asp residue [46]. Various mechanisms for MerB have been proposed in the literature [45–50], but there is still debate over the precise mechanism involved [51].

In our laboratory, efforts are underway to determine the detailed reaction mechanism of Hg-C cleavage in MerB. Active site models of MerB have been constructed from available X-ray structures [46]. Density functional theory (DFT) methods [52] are used to distinguish between likely and unlikely reaction pathways and to determine the specific chemistry involved in achieving catalysis by MerB.

The characterization of the structure and internal dynamics of biomolecules such as proteins and biopolymers is essential to understanding the mechanisms of their biological functions [53]. A combination of molecular dynamics simulations and neutron scattering techniques has emerged as a promising synergistic pathway to elucidating atomistic details of the dynamics of various biological
systems. We review here one aspect of neutron scattering in biology: recent applications in characterizing protein physics, and, in particular, the protein glass transition and protein hydration dynamics [54–70].

In neutron scattering experiments, neutron beams of suitable wavelengths and energies are directed at a sample. Neutrons are scattered by atomic nuclei of the sample. The resulting change in the energy and momentum of incident neutrons provides information about the structure and internal dynamics of the molecules concerned [71]. In MD simulation, the time evolution of the positions and momenta of atoms of biomolecules are tracked by solving the Newtonian equation of motion, taking into account the anharmonic interactions between atoms. The space and time correlation functions calculated from MD trajectories can then be used to interpret experimental neutron data [53,71].

In neutron scattering experiments, the number of neutrons scattered within a solid angle between $\Omega$ and $\Omega + d\Omega$ with a change in energy $\omega$ and momentum $Q$ is measured. This number is proportional to the double-differential cross-section $\frac{\partial^2 \sigma}{\partial \Omega \partial \omega}$, which in turn is proportional to the dynamic structure factor, $S(Q, \omega)$, as shown below,

$$\frac{\partial^2 \sigma}{\partial \Omega \partial \omega} \propto S(Q, \omega)$$

The dynamic structure factor can be written in terms of van Hove function, $G(r, t)$, which characterizes the space-time correlation of individual atoms as well as between the motions of pair of atoms, as shown in Figure 7 [71]. It is evident from Figure 7 that Fourier transformation of physical quantities that quantifies the nature of atomic dynamics in the “Molecular Dynamics Space” leads to information determined in “Neutron scattering space” and vice versa.

**Figure 7.** 4D representation of the relationship between neutron scattering functions (determined from experiments) and space-time correlation functions (calculated using MD simulations) is shown here. Here, $R_i(t)$ denotes the atomic coordinates of $i^{th}$ atom at time $t$. 

$$I(Q, t) = \frac{1}{2\pi} \int G(r, t) e^{-iQ \cdot r} dr$$

$$S(Q, \omega) = \frac{1}{2\pi} \int G(r, t) e^{-iQ \cdot r - i\omega t} dr dt$$

$$G(r, t) = \frac{1}{N} \sum_i \langle \delta(r - R_i(t) + R_i(t)) \rangle$$

$$g(r, \omega) = \frac{1}{2\pi} \int G(r, t) e^{-i\omega t} dt$$
A number of neutron scattering and other experimental and computer simulation studies have revealed that proteins exhibit a temperature-dependent dynamical transition around 180–220 K [54–70]. Below this temperature, the dynamics is similar to that of a glassy material, while at temperatures above 220 K, protein atoms exhibit liquid-like dynamics. Many proteins exhibit this behavior, and correlations have been observed between the onset of protein function, such as ligand binding, proton pumping or electron transfer, and the onset of the anharmonic dynamics at around 220 K. In some cases proteins may function only when the temperature is above the dynamical transition temperature. The transition has also been shown to be sensitive to changes in solvent conditions. For example, proteins immersed in viscous solvents, such as trehalose, exhibit no transition.

The experimental observations led to some important questions concerning the microscopic dynamical details of the protein glass transition, including whether the dynamical transition in a solvated protein is controlled by the solvent and whether proteins exhibit intrinsic anharmonic dynamics much below the glass transition temperature.

Neutron scattering experiments and molecular dynamics simulations have shown signatures of anharmonic dynamics below the ~220 K dynamical transitions for several proteins. The dynamic processes associated with this low-temperature anharmonicity, and how these motions may be related to global dynamical changes at the dynamical transition, have yet to be fully understood.

The mean square displacement (MSD) of atoms of hydrated myoglobin as a function of temperature calculated from MD simulation is shown in Figure 8. The MSD increases linearly at low temperatures then exhibits two slope changes: one at ~150 K and the other at ~220 K. The significant change in gradient at ~220 K is the solvent-driven dynamical transition as observed in many biological systems. At 150 K, the rotational excitations of methyl groups are observed, and these jump-like motions of methyl protons will lead to quasi-elastic neutron scattering of the type that has been observed at around 150 K in proteins. In lysozyme, the low-temperature anharmonicity was observed at 100 K and was attributed to the onset of methyl dynamics [67,68]. It was also demonstrated that the

![Figure 8. Time-averaged mean square displacement of myoglobin as a function of temperature. The straight lines are fits to the data for different temperature ranges (solid line, 0 to 150 K; dashed line, 150 to 220 K; dotted line, above 220 K) and are shown as a guide to the eye. (Reproduced with permission from J. Phys. Chem. B 2008, 112, 5522–5533. Copyright 2008 Am. Chem. Soc.)](image-url)
anharmonic dynamics observed at ~100 K is independent of hydration level, while the dynamical transition at ~200–220 K is observed only at hydration levels greater than 0.2 g of water/g of protein. In incoherent neutron scattering experiments the main contribution to the scattered protein intensity arises from the nonexchangeable hydrogen atoms, and a significant fraction of nonexchangeable hydrogens in proteins reside on CH₃ groups: 26% in lysozyme, for example. Thus, the CH₃ groups would be expected to contribute significantly to the scattered intensity. It has been suggested that a dominant contribution of the relaxation observed in dry myoglobin neutron scattering is due to methyl dynamics. ¹H NMR relaxation studies have also investigated the reorientational dynamics of C-H bond vectors of methyl groups, and ¹H NMR experiments on dry lysozyme have shown that 70% of the total proton relaxation is due to methyl dynamics.

In conclusion, simulation models are tested by calculating neutron scattering structure factors and comparing the results directly with experiments. If the scattering profiles agree the simulations are used to provide a detailed decomposition and interpretation of the experiments, and if not, the models are rationally adjusted. Comparison with neutron experiment is at the level of the scattering functions and also of quantities derived from them. The end result is simulation models consistent with the neutron data that can be used to design a wide range of experiments. In the above work, neutron scattering experiments and MD simulations have demonstrated the non-negligible role of intrinsic anharmonicity of protein dynamics in protein glass transition behavior.

5. Extreme Scale Supercomputers

Current supercomputers, such as the Jaguar Cray XT5 at Oak Ridge National Laboratory, are beginning to assemble over 10⁵ cores and reach petaflop nominal speeds. A challenge for MD is to achieve efficient scaling up to O(10⁵–10⁶) cores. The simulations are limited by the parallel efficiency of the MD algorithms i.e., their ability to run in parallel in many thousand processors.

The computationally most demanding part of an MD simulation of biological systems is the treatment of the long range Coulomb interactions. A common method for this is to directly calculate the Coulomb interaction for any pair of atoms separated by less than another cut-off distance, R_coul, and outside this distance to calculate the electrostatic interactions with the Particle Mesh Ewald (PME) method [72,73]. However, full electrostatic treatment with the PME method limits the performance on massively parallel computers.

An alternative method, that avoids the electrostatics bottleneck, is the use of the Reaction Field (RF) [74]. In this method it is assumed that any given atom is surrounded by a sphere of radius, R_rf, within which the electrostatic interactions are calculated explicitly. Outside the sphere the system is treated as a dielectric continuum. Scaling benchmarks on a 3.3 million atom lignocellulosic biomass system show that the use of the RF drastically improves the parallel efficiency of the algorithm relative to PME, yielding 28ns/day (Figure 9) [75].

6. Conclusions

The work in progress involves both methodological development and applications to systems of prime interest in the energy biosciences. Simulations of the lignocellulosic systems described above are currently being performed on local clusters and DOE supercomputers, the latter with the help of a DOE INCITE allocation.

The enzyme simulation results have provided insight into the molecular principles that govern cohesin-dockerin domain recognition in cellulosome assembly. These principles could in turn be used to guide protein engineering modifications so as to alter cohesin-dockerin binding. Efforts are underway to design engineered cellulosomal modules which can conduct more efficient biomass degradation than the corresponding wild-type protein complexes. Both atomic-detail and coarse-grained computer simulations are expected 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, a key step to attaining more cost-effective cellulosic ethanol.
Figure 9. Strong scaling (i.e., increasing the number of cores with fixed system size) of the 3.3 million atom lignocellulose model shown in Figure 3.

With the QM/MM methodology and the high-performance computer simulations outlined above, an understanding of cellulase catalysis is being obtained that may be of practical use in designing new, efficient enzymes for industrial purposes. Similarly, an understanding of the catalytic mechanisms behind the activity of proteins in the mer operon could potentially lead to the synthesis of small-molecule mimics of use in remediation of mercury-contaminated sites.

Finally, the excitement generated by the coming on line of petascale supercomputers and next-generation neutron sources, such as the Jaguar Cray XT5 and the Spallation Neutron Source, both at ORNL, arises largely from the capabilities these facilities provide in extending the time- and length-scales accessible to detailed physical characterization. We look forward to participating in future advances made possible by the advent of these capabilities.

Acknowledgments
This work was supported in part by the BioEnergy Science Center (BESC). 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. The work was also supported in part by a DOE—Laboratory Directed Research and Development Fund: Systems Biology 003-00044, and a DOE INCITE award to JCS. Research described was also supported by DOE-OBEP grant ERKP704 “Dynamic Visualization of Lignocellulose Degradation” and by the DOE-BER Science Focus Area “Biogeochemical and Molecular Mechanisms Controlling Contaminant Transformation in the Environment” under the Environmental Remediation Science Program. This work was also supported in part by the National Science Foundation (Award No. 0817940 to H. G.).
References


Miller S M. Cleaving c-hg bonds: Two thiolates are better than one. *Nature Chemical Biology.* 2007 Sep, 3(9):537–8.


