Supercomputers Tackle BIOFUEL Production Problems

If you have ever dealt with an uncooperative, fractious kid or a combative employee, you understand the meaning of "recalcitrance" — over-the-top stubbornness, disobedience, and noncompliance. But recalcitrance is not just a human trait — plants can be recalcitrant, too, and for them it is a matter of survival. Over millions of years, plants have evolved complex structural and chemical mechanisms to ward off assaults on their structural sugars by microbial and animal marauders. So it should be no surprise that when humans attempt to turn plant biomass into biofuels to meet our energy needs, we discover how stubborn and noncompliant our vegetative friends can be. Plant recalcitrance is one of the biggest stumbling blocks to realizing an energy-efficient future.

Cellulosic ethanol may ultimately replace as much as a third of the current U.S. demand for transportation fuels with a clean, homegrown, renewable energy source that does not impact food production.

Unlocking the Promise of Biofuels

The hidden promise of biofuels is locked up in lignocellulosic biomass — the cell wall material that makes up the bulk of plant mass and provides plants with their resilience and shape. Producing ethanol from cellulose will reduce global dependence on fossil fuels and increase national energy independence and security. Also, according to the U.S. Department of Energy (DOE), ethanol from cellulose may reduce greenhouse gas emissions by about 90% relative to gasoline, and production of other pollutants is also minimal. Cellulosic ethanol may ultimately replace as much as a third of the current U.S. demand for transportation fuels with a clean, homegrown, renewable energy source that does not impact food production.

Cellulosic ethanol or other biofuels can be created from a wide variety of lignocellulosic biomass "feedstocks," including plant wastes from agriculture and industry. Because much of the feedstock can be grown on marginal land, expanding the use of biofuels from stalks, trunks, and leaves of specific energy crops — notably switchgrass or poplar trees — may eliminate the need to choose between growing food or fuel crops on fertile land. Finally, because several sources of cellulose are perennial crops that demand little water and no fertilizer, their environmental impact is far less than that of annual crops like corn.

Plant Cell Wall Recalcitrance

Clearly, lignocellulosic biomass holds great promise as a major renewable energy source. All we need to do is process the cellulosic biomass to extract fermentable sugars from the feedstock. However, the plants are not cooperating.

The plant cell wall material in lignocellulosic biomass is composed mostly of the polymers cellulose, hemicellulose, and lignin (figure 1). Roughly two-thirds of the dry mass of cellulosic materials is present as cellulose and the gel-like matrix polysaccharides, hemicellulose and pectins. Another polymer, lignin, makes up the bulk of the remainder. Lignocellulose is a sturdy, microstructured laminate in which the cellulose fibers are surrounded by the heterogeneous hemicellulose and pectin and associated with lignin.

The sugars in cellulose and hemicellulose are locked in complex carbohydrates called polysaccharides (long chains of simple sugars). Cellulose itself is composed of linear glucose chains (figure 2, p36) assembled into insoluble fibers.

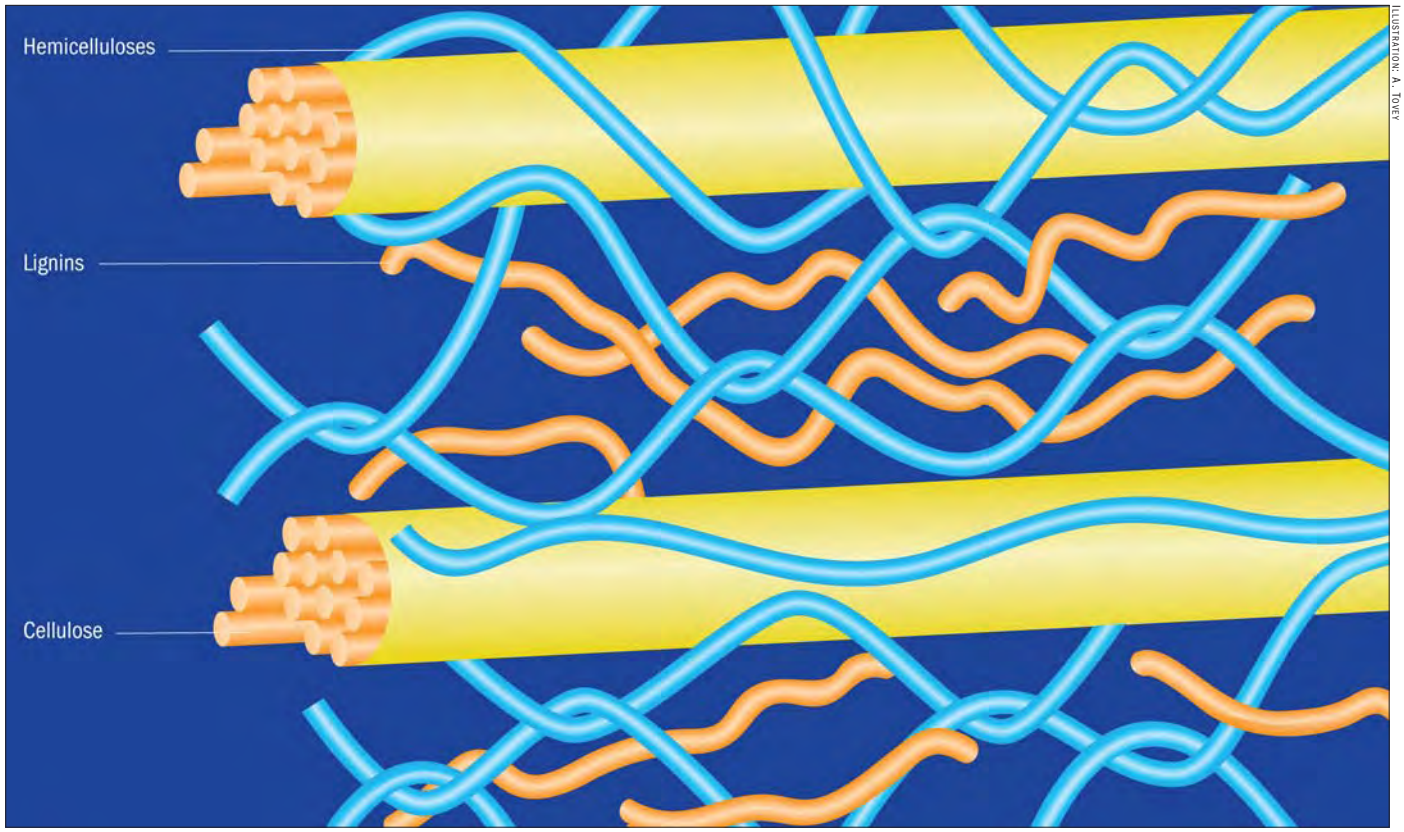


ILLUSTRATION: A. TOVEY

Figure 1. *A schematic of a plant cell wall showing cellulose fibrils (brown) laminated with hemicellulose (turquoise) and lignin (orange) polymers.*

The key to efficient, economical production of cellulosic ethanol is separating these complex structures into fermentable sugars using a chemical reaction, hydrolysis. Hydrolyzed cellulose is potentially a major feedstock for ethanol production by glucose fermentation (producing bioethanol), as well as many other fermentation products.

Although the fermentation itself is cost-effective and efficient, cellulose hydrolysis is difficult. Plants are recalcitrant — they resist being subjected to this kind of abuse. Naturally, the trees and grasses that are the most promising as potential sources of fuel are also the most oppositional when it comes to giving up their sugars to cellulous hydrolysis.

The recalcitrance originates from the complex lattice of cellulose, hemicelluloses, and lignin that provides strength, prevents breakage, and protects against the ravages of weather, insects, and disease. This structuring serves as a barrier to transforming the plants into simpler sugars that can be processed into ethanol or other types of fuels and chemicals. It is this recalcitrance that is largely responsible for the high cost of lignocellulose conversion.

Natural factors believed to contribute to the recalcitrance of lignocellulosic feedstock to chemicals or enzymes include:

- the degree of lignifications
- the structural heterogeneity and complexity of cell-wall constituents, such as the cellulose microfibrils and the matrix polymers and cross-linkages between these components
- the difficulty enzymes have in acting on an insoluble substrate
- crystallinity and restricted solvent accessibility

These chemical and structural features of biomass affect liquid penetration and/or enzyme accessibility and activity.

At the sub-molecular level, the crystalline cellulose core of cell-wall microfibrils is highly resistant to chemical and biological hydrolysis because of the precisely arranged chains that make up its structure. Biological hydrolysis can be accomplished by enzymes called cellulases. However, strong inter-chain hydrogen bonding between adjacent chains in a cellulose sheet make crystalline cellulose resistant to enzymatic hydrolysis. In contrast, hemicellulose and amorphous cellulose are more readily digestible.

Higher-order structures in plants also contribute to biomass recalcitrance. For example, a coating of amorphous cellulose and hemicellulose

The key to efficient, economical production of cellulosic ethanol is separating these complex structures into fermentable sugars using a chemical reaction, hydrolysis.

ILLUSTRATION: A. TOVEY

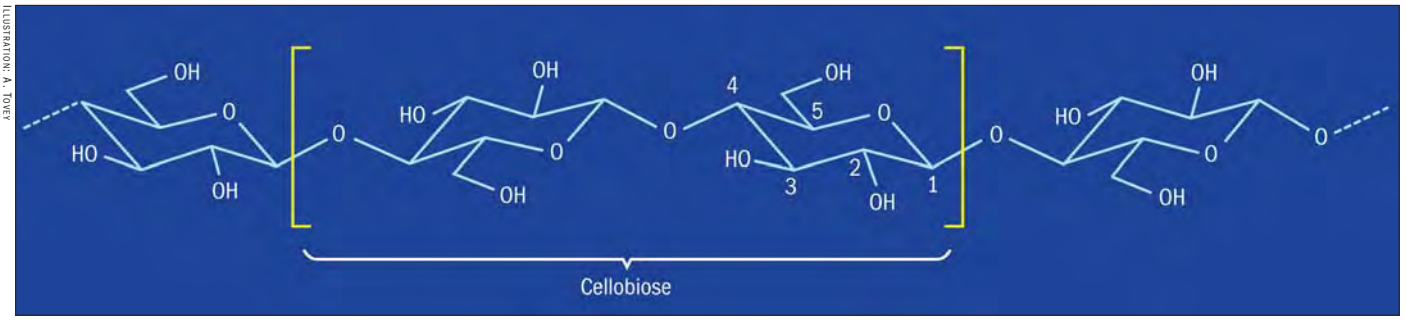


Figure 2. *Cellulose, a linear polysaccharide chain with cellobiose (two glucose units) as the repeating unit.*

restricts access to the crystalline cellulose cores of microfibrils. At the microscopic and a macroscopic scale, the complex heterogeneous nature of biomass makes it difficult to deliver chemical or biochemical catalysts to the reaction sites.

Chemical and Mechanical Pretreatment

Pretreatment of biomass is currently required for efficient hydrolysis of cellulose to glucose. Several types of mechanical and chemical pretreatments are used to increase the hydrolysis rate of biomass (without generating undesirable side products), such as enzymatic treatment and mechanical comminution, alkali swelling, acid hydrolysis, steam and other fiber explosion techniques, and exposure to supercritical fluids. These processes act by a variety of mechanisms to render the carbohydrate components of lignocellulosic materials more susceptible to enzymatic hydrolysis and microbial conversion. The pretreatments commonly increase the surface area and the accessibility of cellulose to cellulase enzymes by destruction of the lignocellulose microstructure.

However, these techniques also increase the cost of ethanol production because of their high energy requirements for mechanical size reduction, and the heating and costs associated with the responsible handling of caustic chemicals. Therefore, pretreatment is a rate-limiting, energy-intensive, and expensive processing step in the production of cellulosic ethanol. It is further complicated by the re-precipitation of the lignin residue on the cellulose surface, which results in inhibition of cellulose hydrolysis and other undesirable side effects.

Physical Principles of Recalcitrance

Biomass recalcitrance is largely related to the structure of the plant cell wall before and after the pretreatment steps. The accessibility of the cellulose strands to enzymic hydrolysis is of critical importance. Consequently, an understanding of biomass recalcitrance and its relationship to hydrolysis efficiency requires physical characterization of lignocellulosic biomass and the development and testing of ideas related to the physical properties. Overcoming plant cell wall recalcitrance requires transformational, fundamental and applied, interdisciplinary, experimental and computational research. As we will see, high-performance computer simulation can play a central and federating role in this endeavor, which helps to provide fundamental insight into plant cell wall biophysics.

Research is focused on understanding a variety of physical processes — for example, how microbes can act on biomass to transform it into sugars. In this case, we need to understand the mesoscopic microbe–biomass interface. The hydrolytic microbial enzyme complexes must be diffused to the biomass surface, past higher-order structures such as lignin, in order to come into contact with the hydrolyzable cellulose. The physical properties determining the extent of diffusion are likely to include pore sizes, solvent viscosity and structure, and lignin and hemicellulose structures and interactions. Of particular importance is the structure of the cellulose itself, which may be part crystalline and part noncrystalline (amorphous). Finally, the mechanical and chemical mechanisms by which the hydrolytic enzyme complexes bind and hydrolyze the cellulosic substrate need to be determined.

The secret to cracking the problem of efficient cellulosic ethanol production is to understand molecular mechanisms such as those above and to use that knowledge to rationally design new and less recalcitrant cell walls, as well as genetically manipulating the microbes and enzymes used in the process to enable them to work more efficiently.

Pretreatment is a rate-limiting, energy-intensive, and expensive processing step in the production of cellulosic ethanol.

Role of Computer Simulation

The complexity of biological molecular systems makes pencil-and-paper mathematical descriptions of their structure, dynamics, and function very difficult. Hence, these systems are classic candidates for computer simulation analysis. The simulation procedure includes entering as much relevant information as possible from experiment and high-level calculations (such as quantum chemistry) to create

Supercomputing and NCCS

Those recalcitrant plants do not have a chance — not when researchers like Jeremy C. Smith have access to the world's most powerful supercomputers for science. Jaguar, the Cray XT high-performance computing system at the National Center for Computational Science (NCCS) at Oak Ridge National Laboratory (ORNL), is allowing researchers to dig into the thorny problems associated with creating biofuels from sustainable biomass and a wide range of other strategically important scientific domains as well.

The NCCS mission is to advance the state of the art in high-performance computing and make available the resources of powerful parallel supercomputers to scientists investigating a wide variety of compute-intensive projects. NCCS works with industry, laboratories, government agencies, and academia to address challenges in such diverse areas as climate, fusion, astrophysics, materials science, nanoscience, chemistry, biology, combustion, accelerator physics, engineering, and other disciplines relevant to maintaining U.S. science leadership.

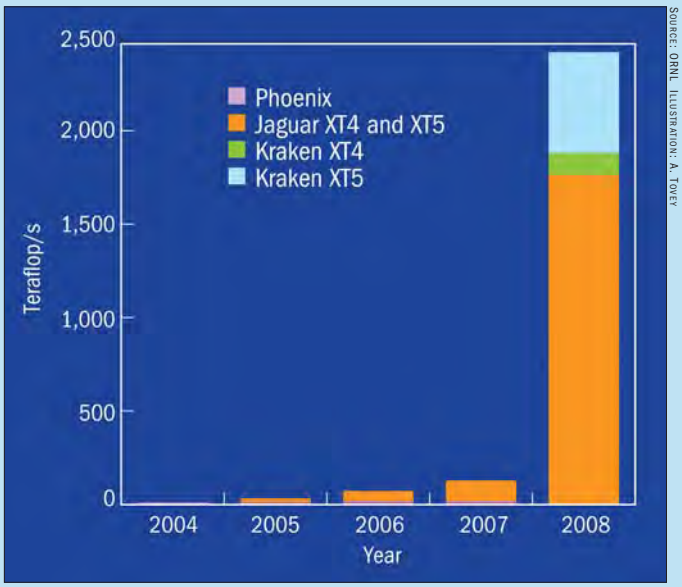
In 2008, NCCS threw the switch on Jaguar and changed the face of scientific computing. Jaguar has a peak performance of 1.64 petaflop/s, incorporating 1.382 petaflop/s XT5 and 263 teraflop/s XT4 systems. As shown in figure 3, this was a giant leap for supercomputing and cutting-edge scientific research.

Jaguar has already run scientific applications ranging from materials to combustion on the entire system, sustaining petaflop/s performance on multiple applications. A calculation that once took months can now be done in minutes. A 2008 report from the U.S. Department of Energy Office of Science — America's largest funder of basic physical science programs at universities and government laboratories — said six of the top ten recent scientific advancements in computational science used Jaguar to provide unprecedented insight into supernovas, combustion, fusion, superconductivity, dark matter, and mathematics.

To date, the computer simulations on Jaguar have focused largely on addressing new forms of energy and understanding the impact on climate resulting from energy use. For example, among the INCITE projects is the simulated enzymatic breakdown of cellulose to make production of biofuels commercially viable, reported on in this article. Other Jaguar-based projects include coal gasification processes to help industry design near-zero-emission coal-fired power plants; studies to determine how fuel burns, which is important for fuel-efficient, low-emission engines; and computer models that have helped physicists use radio waves to heat and control ionized fuel in a fusion reactor. Engineers have designed materials to recover energy escaping from vehicle tailpipes. Simulation insights have enabled biologists to design new drugs to thwart Alzheimer's disease fibrils and engineer the workings of cellular ion channels to detoxify industrial wastes.

Jaguar is not alone. Installed adjacent to the super supercomputer is another Cray XT5 system, dubbed Kraken after the mythical sea monster. Kraken is capable of 615 teraflop/s, making it the fastest computer in the world deployed exclusively for academic use. Kraken was funded by the National Science Foundation as part of a $65 million grant won by the University of Tennessee and its partners, including ORNL. The supercomputer will be upgraded with faster processors about a year from now, when it is expected to surpass the petaflop/s barrier — a quadrillion calculations per second.

The combination of these two high-performance machines and their associated network, file-handling, and storage infrastructures provides researchers with unprecedented computational resources to tackle some of today's most intractable problems.



SOURCE: ORNL ILLUSTRATION: A. TOVEY

Figure 3. *Supercomputing speeds shot skyward in 2008 as Oak Ridge National Laboratory's Cray XT5 Jaguar supercomputer attained a peak performance of 1.64 petaflop/s, a level that seemed astronomical only a few years ago.*

an atomic-detail model of the system and its interactions, and then using high-performance computer clusters and supercomputers to follow the evolution of the system when it is subjected to realistic forces. Simulations have provided much useful information on the functioning of proteins and have been successfully used to design drugs. The technique holds the promise of revealing the mechanisms of plant cell-wall deconstruction in exquisite detail and leading to an understanding of the structures, dynamics, and degradation pathways of lignocellulosic materials.

However, the structural complexity of lignocellulosic biomass means much work must be done before reliable simulation models of plant cell wall structures can be constructed. Some initial steps in this direction are being taken by members of the Center for Molecular Biophysics (CMB) at Oak Ridge National Laboratory (ORNL) in Tennessee, together with groups led by Mark Nimlos, Mike Himmel, and Mike Crowley at the National Renewable Energy Laboratory (NREL) in Colorado, and John Brady of Cornell University in New York.

SIMULATION: B. LINDNER, ORNL CMB

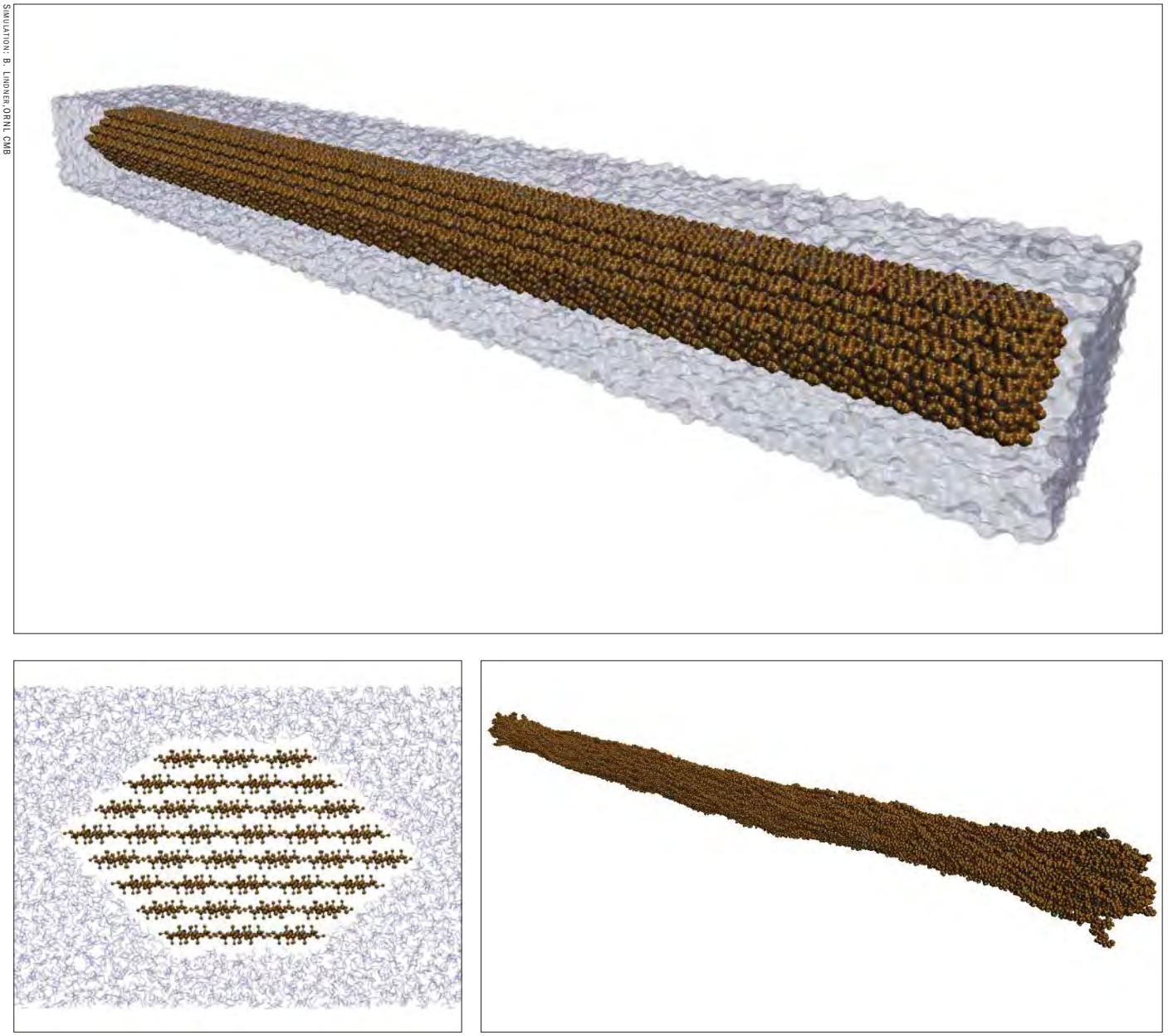
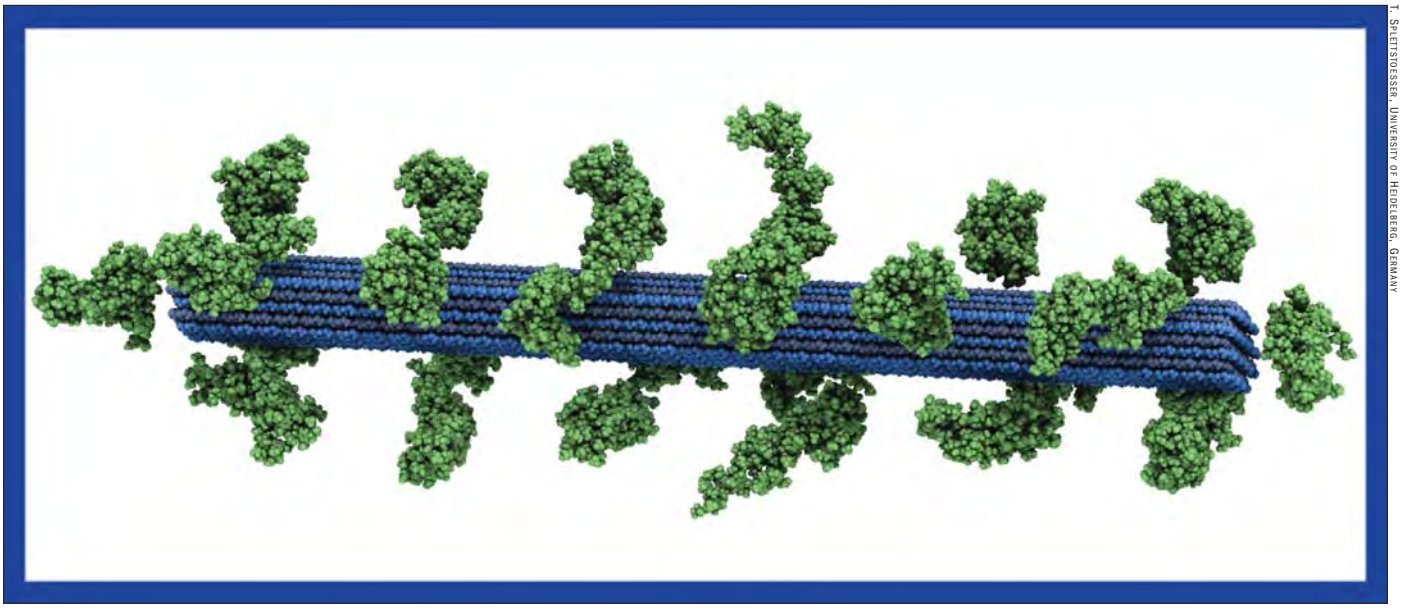


Figure 4. *An atomic-detail simulation model of a cellulose microfibril in aqueous solution, including an amorphous region generated by heating to 350 K. Top, the complete system; lower left, cross section of an amorphous region; lower right, close-up of an amorphous region.*

The increases in computer power available from various DOE will enable fine-grained simulation of complex heterogeneous lignocellulosic biomass systems.

Because biomass is a multicomponent and complex multiscale material, clarifying its structure is particularly challenging. Required are experimental imaging characterization techniques, such as atomic force microscopy and neutron scattering, which span a wide range of length scales (from angstroms to micrometers) and can differentiate between individual components and phases within the biomaterial, such as crystalline and amorphous cellulose, lignin, and hemi-cellulose.

In addition, the computer simulations will need to be performed at several levels of detail; that is, they will involve a multiscale approach. At the atomic level, input to simulations typically include the positions of the atoms and a formula yielding the energy of the system (the potential energy function). By solving Newton's equation, the motion trajectories of the atoms can be built up using the molecular dynamics (MD) method and the behavior of the system examined in detail. This approach is limited by computer power constraints and until now has been restricted to systems of approximately 100,000 particles over a 100 nanosecond timescale. However, the increases in computer power available from various DOE supercomputers (sidebar "Supercomputing and NCCS" p37) will allow access to microsecond and micrometer time and length scales for atomic-detail MD, and will enable fine-grained simulation of complex heterogeneous lignocellulosic biomass systems.



T. Saletsteosser, University of Heidelberg, Germany

Figure 5. *An atomic-detail model of the lignocellulose of softwoods. The model was built by Loukas Petridis of the ORNL CMB, based on experimental data on the structure of cellulose (blue) and lignin (green). This model is being subjected to molecular dynamics simulation on DOE supercomputers.*

Furthermore, progress in developing coarse-grained multiscale simulation techniques will permit even larger ensembles of heterogeneous biomass systems to be simulated, allowing us to understand the degradation process of the biopolymers in detail. To reach longer length and time scales by fully exploiting the potential of petascale (and beyond) leadership-class supercomputers, MD applications must be scaled up to 100,000 processors and beyond, while simultaneously and accurately coarse-graining the inter-atomic interactions.

Lignocellulosic Biomass Simulations

The simulation plan for understanding lignocellulosic biomass involves building up accurate models of cellulose, lignin, and lignocellulose based on first-principles calculation of inter-atomic interactions while integrating experimental input. This buildup will be followed by constructing and simulating models of heterogeneous biomass systems. At the same time, we need to obtain an understanding of the structure, dynamics, and function of the cellulase enzymes involved.

Potential Energy Function

The crucial first step in modeling biological systems is the derivation of a potential energy function, or force field. Force fields exist for carbohydrates such as cellulose thanks to pioneering work on simulations of simple carbohydrates by John Brady and his colleagues. However, a force field for lignin had been lacking, so Loukas Petridis of CMB set about obtaining the required fundamental information. Development of the force field was based on reproducing accurate data on small model compounds obtained by quantum-mechanical calculations. The force field Petridis derived was subsequently validated by performing an MD simulation of a crystal of a lignin fragment molecule and comparing simulation-derived structural features with the experimental results. Together with the existing force field for carbohydrates, this lignin force field will now allow us to perform full simulations of lignocellulosic biomass.

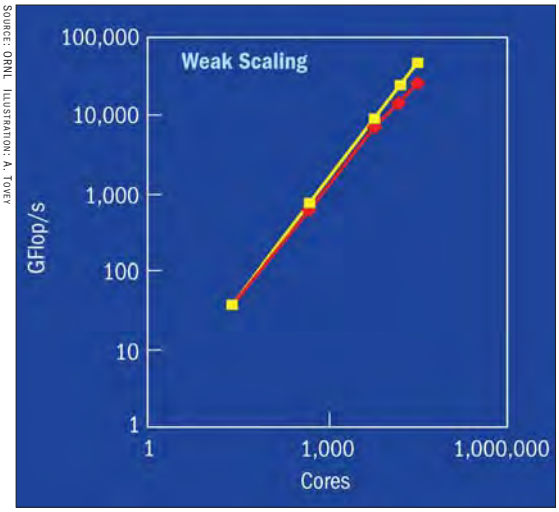
Crystalline Cellulose

Because the chemical structure of cellulose is known, it is a relatively uncomplicated task to build cellulose microfibrils using the molecular structure. Starting from these microfibril models, the mechanical properties of crystalline cellulose, such as the resistance of strands to being separated from each other, may be explored. These properties are likely to be important for recalcitrance to hydrolysis. This resistance can be explored using MD simulation and vibrational calculations.

In vibrational calculations, atoms are considered to be joined together by springs, and the resulting normal mode analysis describes the motions of the system as a set of superposed oscillating vibrational modes. To obtain a complete description of the vibrational properties of crystalline cellulose requires lattice dynamical calculations that incorporate the periodic properties of the crystal. These methods, already developed and programmed in our laboratory in collaboration with Franci Merzel of the National Institute of Chemistry in Ljubljana, Slovenia, are now being adapted for calculations on cellulose. Lattice dynamical calculations can determine the forms and frequencies of the soft dynam-

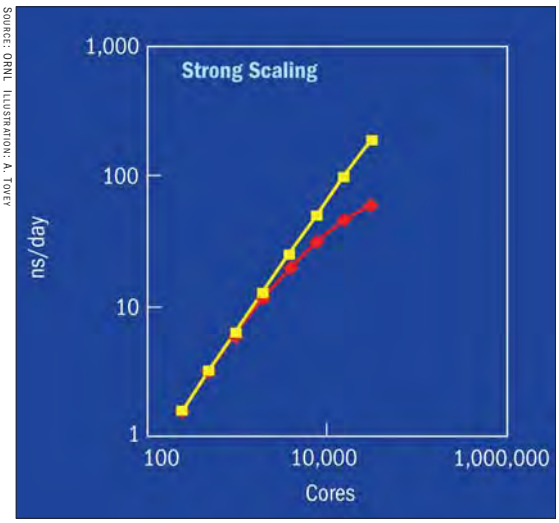
Progress in developing coarse-grained multiscale simulation techniques will permit even larger ensembles of heterogeneous biomass systems to be simulated, allowing us to understand the degradation process of the biopolymers in detail.

MD simulations at different temperatures and pressures of lignocellulosic subsystems can potentially provide a wealth of information useful for interpreting experiments.



SOURCE: ORNL ILLUSTRATION: A. TOVEY

Figure 6. *Single-temperature, weak-scaling data on Jaguar obtained by keeping the number of atoms simulated per core equal to 175 while increasing the number of cores. Data on 30,000 cores are projected.*



SOURCE: ORNL ILLUSTRATION: A. TOVEY

Figure 7. *Single-temperature, strong-scaling data obtained by keeping the size of the system fixed at 5.4 million atoms while increasing the number of cores on which the simulation was run on Jaguar. Data on 30,000 cores are projected.*

ical modes in the crystalline system, that is, those motions that determine the mechanical resistance of crystalline domains to fiber disruption and extrusion. MD simulation of crystalline cellulose can also probe additional physical properties such as the fibril resistance to temperature and pressure, as well as force-pulling MD, which examines the force required to pull strands apart.

Amorphous Cellulose

Part of the cellulose content of biomass is amorphous (noncrystalline), and the ratio of amorphous to crystalline cellulose is an important determinant of the recalcitrance to hydrolysis. Simulation models of cellulose containing both crystalline and amorphous regions are now being created at ORNL. See figure 4 (p38) for one example of these models.

Atomistic MD simulations of cellulose with both crystalline and amorphous regions can also be used to examine the amorphous–crystalline interface, including associated water structure and dynamics and hydrogen bonding interactions. The effect of the amorphous regions on the mechanical properties of the crystalline regions will also be able to be examined. Further, molecular simulation at this level is one of the few methods able to explore the influence of water structure on lignocellulose order and accessibility.

Atomistic Simulations of Lignocellulosic Biomass

Access to the crystalline cellulose cores of microfibrils in lignocellulosic biomass is restricted by lignin and a coating of amorphous cellulose and hemicellulose. Because of the complexity and variability of the systems involved, we expect that the generation of accurate simulation models for lignocellulosic biomass — including the hemicellulose, pectin, lignin, and amorphous/crystalline cellulosic components — will necessitate iterative adjustments as experimental data from the DOE Bioenergy Research Centers and other sources become available. A model of cellulose coated with lignin, presently being subjected to MD simulation on the DOE Jaguar and University of Tennessee (UT) Kraken supercomputers, is shown in figure 5 (p39).

To use computer simulation to investigate the effects of pretreatment, a number of types of biomass simulations will need to be performed. Heat pretreatment (170–250°C for a few seconds to a few minutes), which is widely used to increase the yield of cellulose hydrolysis, will be examined. Such simulations will be able to investigate, for example, how the cellulose crystallinity index and accessibility to enzymes varies with temperature, and whether application of pressure can lead to formation of free chain ends.

MD simulations at different temperatures and pressures of lignocellulosic subsystems can potentially provide a wealth of information useful for interpreting experiments. For example, we will be able to determine the temperature dependence of the interaction mode and strength of lignin with cellulose — information useful for describing the mechanism of delignification that follows heat pretreatment. In addition, we will examine the structural properties of the lignin macromolecules as a function of temperature. These may be critical in permitting enzymic attack.

Use of Supercomputers

The use of petascale supercomputers will allow us to conduct MD simulation of large, heterogeneous bio-

mass structures. An important task in supercomputing application is to optimally use the large number of processors (cores) in parallel. Graduate student Roland Schulz of the CMB has recently performed tests on ORNL's Jaguar system indicating that atomistic MD on a one million atom biomass system, run on 5,760 cores, produces 86 ns/day. Preliminary runs on a five million atom lignocellulosic system achieves very good efficiency at up to 30,000 cores, producing about 60 ns/day (figures 6 and 7).

Enzymes

Enzymes called cellulases act on cellulose to convert it into sugars. Some microorganisms produce a battery of enzymes that work synergistically to degrade crystalline cellulose (figure 8). Understanding how these enzymes access their cellulosic substrates, and the mechanisms by which they catalyze the ensuing chemical reactions, will furnish fundamental knowledge about the cellulosic ethanol design process.

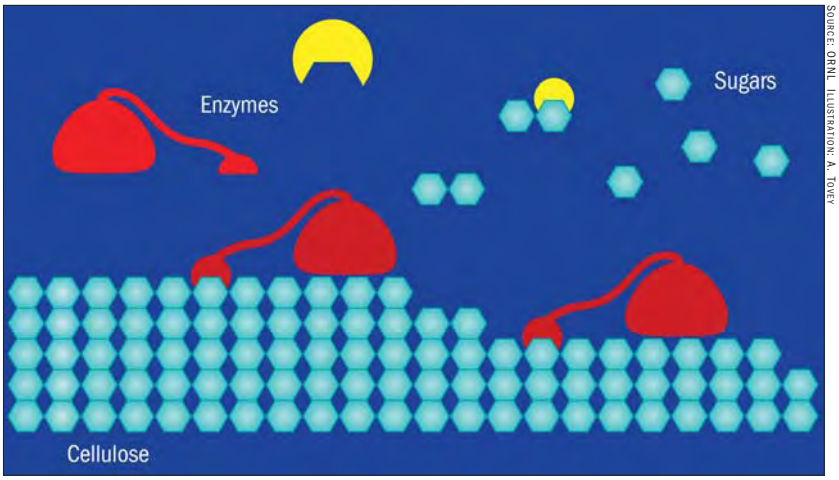
Enzymes such as the cellulase in figure 9 can act on crystalline cellulose. However, cellulase enzymes function much more slowly than many other enzymes. Unfortunately, the enzymatic decrystallization process is both critical and poorly understood. The mechanism of action of these enzymes in the context of the cellulose surface must be understood at the molecular level, and computer simulation is expected to furnish vital details in this regard.

Hydrolytic Reaction Mechanism

Enzymic catalysis of the hydrolytic reaction is at the core of cellulosic ethanol production. Computer simulation will play an important role in obtaining an understanding of how microbial enzymes catalyze this hydrolysis reaction. Catalysis corresponds to lowering the energy of the highest-energy structure (the transition state) along the reaction. Key questions include identifying the reactants, intermediates, products, and transition state structures, and then quantitatively understanding how the interactions in the enzyme active site achieve transition state stabilization.

Multiscale simulation techniques can provide the required information using the quantum mechanical/molecular mechanical (QM/MM) methodology. In this methodology, the enzyme is partitioned into two regions. One region contains the catalytically active atoms and is treated accurately with quantum mechanical methods that allow electrons to jump between atoms; that is, chemical bond breaking and forming. Surrounding this region are the environmental atoms, which are treated more empirically with quick-to-calculate, chemically intuitive molecular mechanics force fields.

Hao-Bo Guo, Moumita Saharay, and Hong Guo, of the UT and ORNL CMB, are using QM/MM to clarify the reaction mechanisms of various cellulases.



SOURCE: ORNL ILLUSTRATION: A. TOVEY

Figure 8. *An example of free enzyme systems. One strategy employed by nature and observed in oxygen-rich environments involves fungi and bacteria that secrete individual enzymes with complementary cell wall-degrading functions. These free enzymes work as a team to deconstruct plant cell wall carbohydrates into simple sugars.*

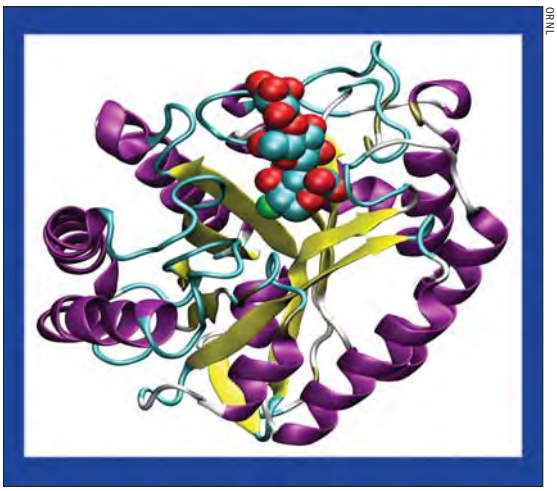

ORNL

Figure 9. *The structure of a cellulase enzyme. The enzyme is shown as ribbons and the substrate as spheres.*

In the example shown in figure 10 (p42), the QM/MM method was used to study a deglycosylation reaction catalyzed by a cellulase named Cel5A from *Bacillus agaradhaerens*. The energy barrier for the reaction was calculated to be 24.2 kcal/mol (figure 10a). The snapshots obtained from the QM/MM free energy simulations representing the intermediate state (IS), the structure near the transition state (TS2), and the product state (PS) are shown in figure 10. Results such as these are being used to identify important interactions in the enzyme active sites that determine the mechanism of cellulase catalysis.

Cellulosomes

Bacteria living in oxygen-free environments produce cellulosomes — large complexes that degrade the cell wall by assembling several different enzymes into a single protein structure. Cellulosomes protrude from bacterial surfaces, latch

The mechanism of action of these enzymes in the context of the cellulose surface must be understood at the molecular level, and computer simulation is expected to furnish vital details in this regard.

SOURCE: ORNL ILLUSTRATION: A. TOVEY

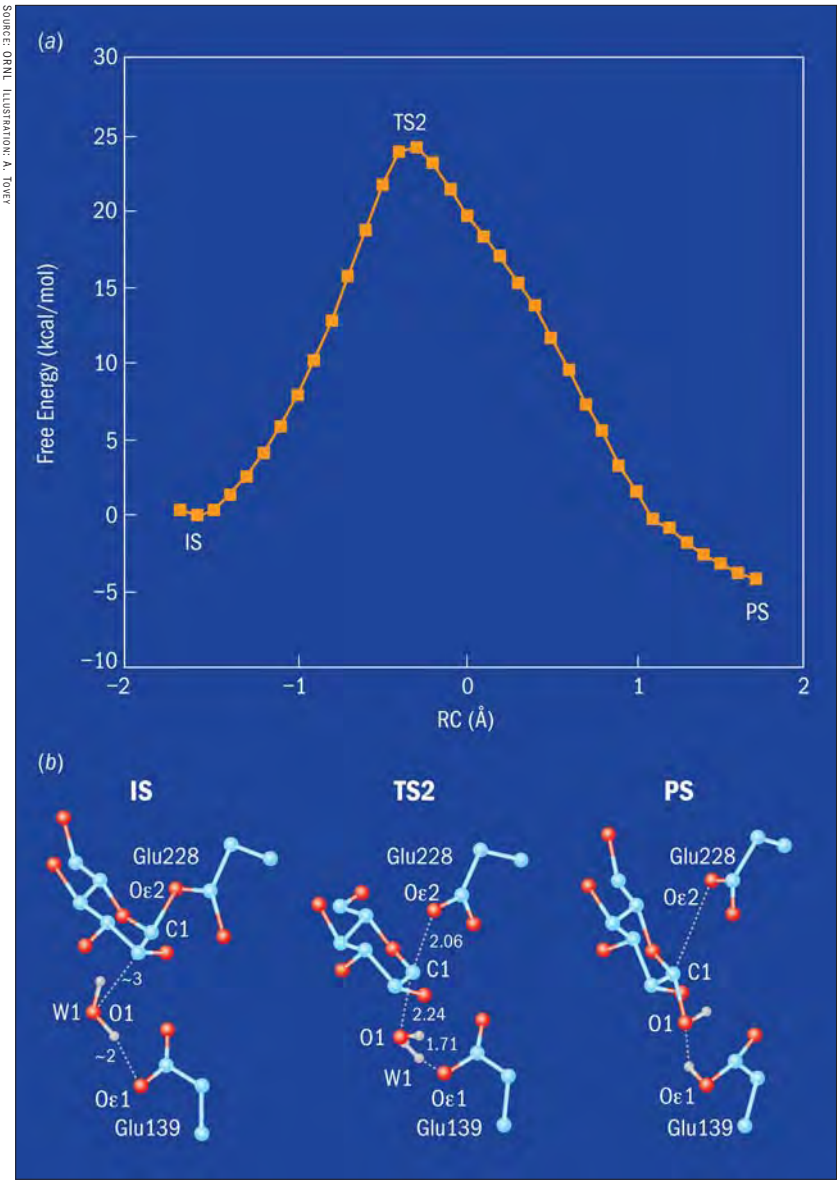


Figure 10. *(a) The free energy profile of the Cel5a cellulase reaction. (b) Snapshots from the QM/MM free energy simulation. Left, the intermediate state (IS); center, the structure near the transition state (TS2); right, the product state (PS).*

onto plant cell walls, and hydrolyse carbohydrates into simple sugars.

The cellulosome complex consists of various kinds of enzymes arranged around a scaffolding protein that does not exhibit catalytic activity but enables the complex to adhere to cellulose (figure 11). *Clostridium thermocellum*, a model bacterium for studying cellulosomes, can produce more than 25 different cell wall-degrading enzymes that it can plug and play into its large protein scaffold. By sensing the surrounding environment, *C. thermocellum* can modify the functionality of its cellulosomes on demand by assembling different combinations of enzymes to attack various compounds in the plant cell wall. The LEGO®-like arrangement of enzymes in cellulosomes offers a unique opportunity to engineer "designer" multienzyme complexes targeted to specific biomass types or for use at different stages of biomass deconstruction.

The organization of the cellulosome depends on high-affinity protein–protein interactions between cohesin domains within the scaffolding proteins and complementary dockerin domains carried by cellulosomal enzymes (figure 12, p44). To deepen our understanding of this interaction and gain further insight into the assembly mechanism, in recent work, Jiancong Xu of CMB collaborated with Mike Crowley of NREL to perform MD simulations of the cohesin–dockerin complex on the UT/ORNL Kraken supercomputer. The motions of the atoms in complexes and the detailed structure formed by several highly evolutionarily-conserved amino acid residues in the proteins were determined, and the simulations also determined the work profile for the dissociation of the cohesin and dockerin domains (figure 13, p45). The simulation results provided us with insight into the molecular principles that govern cohesin–dockerin recognition.

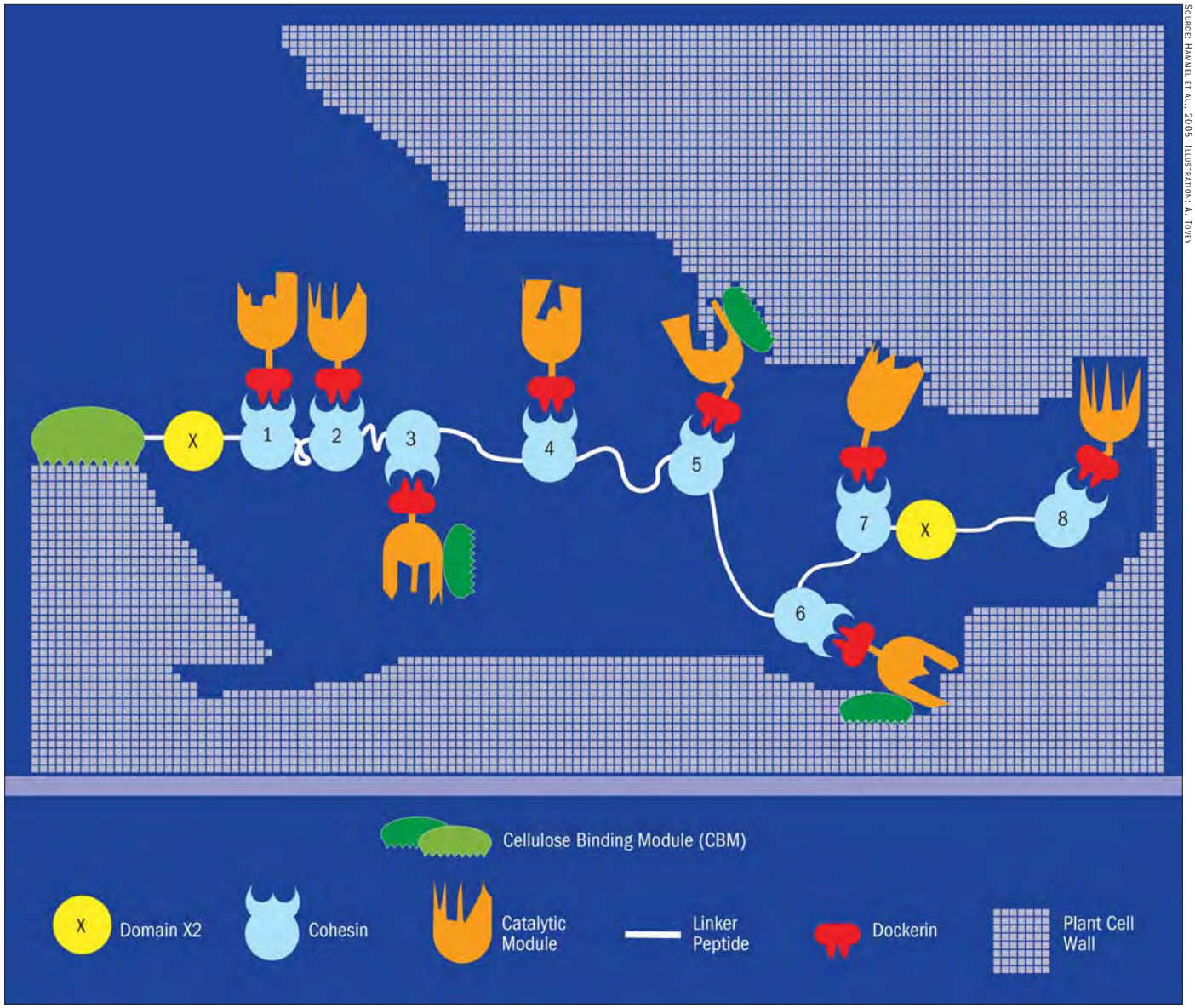
Multiscale: Toward Systems-Level Simulation

The scaling results on the Jaguar system described above indicate that petascale supercomputers can produce MD simulations of lengths approaching a microsecond for million-atom systems. This atomistic simulation work should allow a detailed description of the workings of essential protein machines and allow us to obtain the molecular structure and mechanics of biomass processes.

However, an important aim of our high-performance simulation work is to link atomistic models into systems-level models of biomass deconstruction. In this systems-level simulation modeling, we foresee gaining an explicit representation of the microbial–biomass interface at millimeter–micrometer length scales. Also, we should be able to simulate diffusive and dynamical processes on millisecond–microsecond timescales or longer. Information from multiscale simulation on the structure, dynamics, and mechanics of lignocellulosic biomass as well as the fate of the cellulose, pectin, hemi-cellulose, and lignin components during pretreatment will, when closely integrated with experiment, provide the fundamental understanding needed to overcome biomass recalcitrance to hydrolysis.

To reach the above goals, work is under way at ORNL and NREL to establish a multiscale software framework that will use petascale capability supercomputing to enable the performance of large-scale dynamical simulations of lignocellulosic biomass deconstruction. The methodology is exploring different, complementary, and potentially additive avenues for speeding up computations on leadership-class supercomputers. To

This atomistic simulation work should allow a detailed description of the workings of essential protein machines and allow us to obtain the molecular structure and mechanics of biomass processes.



SOURCE: HAMMEL ET AL., 2005 ILLUSTRATION: A. TOVEY

Figure 11. *A schematic representation depicting a functional model of a cellulosome and the interaction of its component parts with the cellulose substrate.*

do this, we are working on coarse-graining atomistic MD simulation by first replacing an explicit with an implicit solvent and then simplifying the description of the solute. Success in these proposed endeavors may lead to increases in processing speeds of up to five orders of magnitude. In a few years, this speedup would make accessible, for example, a 10 ms timescale for systems that today can only be simulated at 10 ns.

In modeling biological solutes, such as biomass or enzymes, it is important but computationally costly to take into account long-range electrostatic interactions involving the surrounding solvent. Explicit inclusion of solvent molecules in MD simulation is the most accurate way of doing this. However, this strategy not only adds significant extra cost to the calculation but also leads to scalability problems on petascale supercomputers.

An alternative is to model the solvent effect implicitly, by using approaches in which the electrostatic dielectric screening effect of the solvent on the interactions between solute atoms is included by modifying the effective interactions between the solute atoms themselves. Staff scientist Xiaolin Cheng at ORNL is working on technical aspects of these methods as they pertain to supercomputer implementation.

Further coarse-graining is possible by reducing the number of interaction centers within the solute (for example, enzymes, biomass components). As biomass recalcitrance is an essentially mechanical property, it is important to retain the mechanical properties of the solute in any simplification. In collaboration with Kei Moritsugu of the RIKEN National Laboratory in Japan, we have recently developed a multiscale approach aimed at transferring the

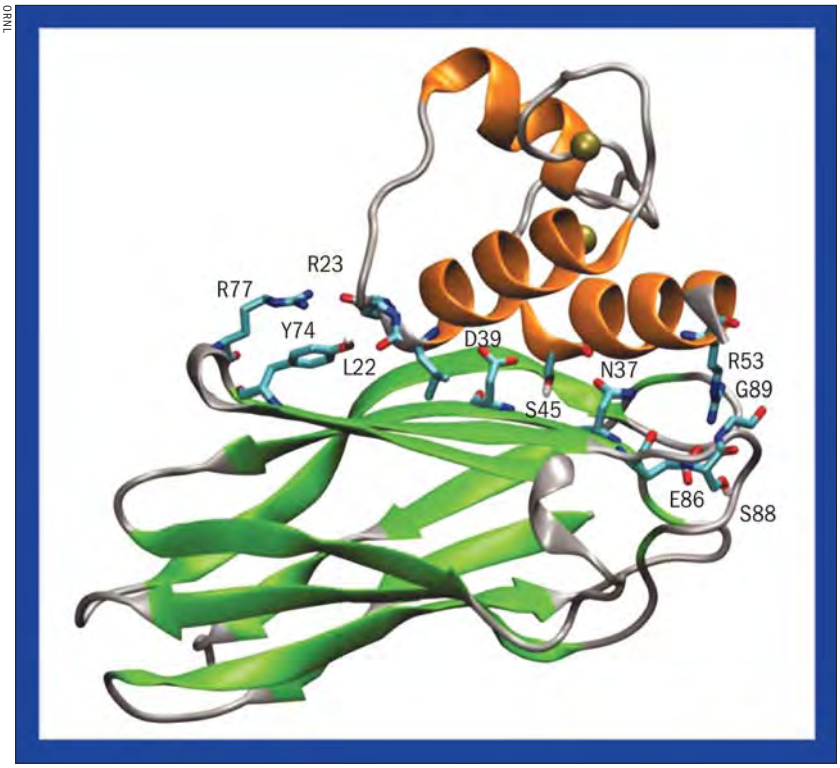


Figure 12. *The crystal structure of a cellulosomal cohesin–dockerin complex in cartoon representation with beta-sheets (cohesin) in green, alpha-helices (dockerin) in orange, and loop regions in silver. Key amino acid residues involved in inter-domain interaction are highlighted in licorice mode and colored by atom names.*

mechanical properties of biomass derived from atomistic simulations onto mesoscale models. The method is called REACH (Realistic Extension Algorithm via Covariance Hessian). REACH's underlying principle is to obtain from atomistic simulations information on correlated motions that is used to construct the coarse-grained simulation model. This model is built using interaction centers that, for example, may correspond to lignocellulosic biomass monomers. We are now further developing the methodology with the specific aims of applying REACH force fields to the simulation of lignocellulosic biomass and parallelizing the simulation code for petaflop/s supercomputers.

Progress Toward Overcoming Biomass Recalcitrance

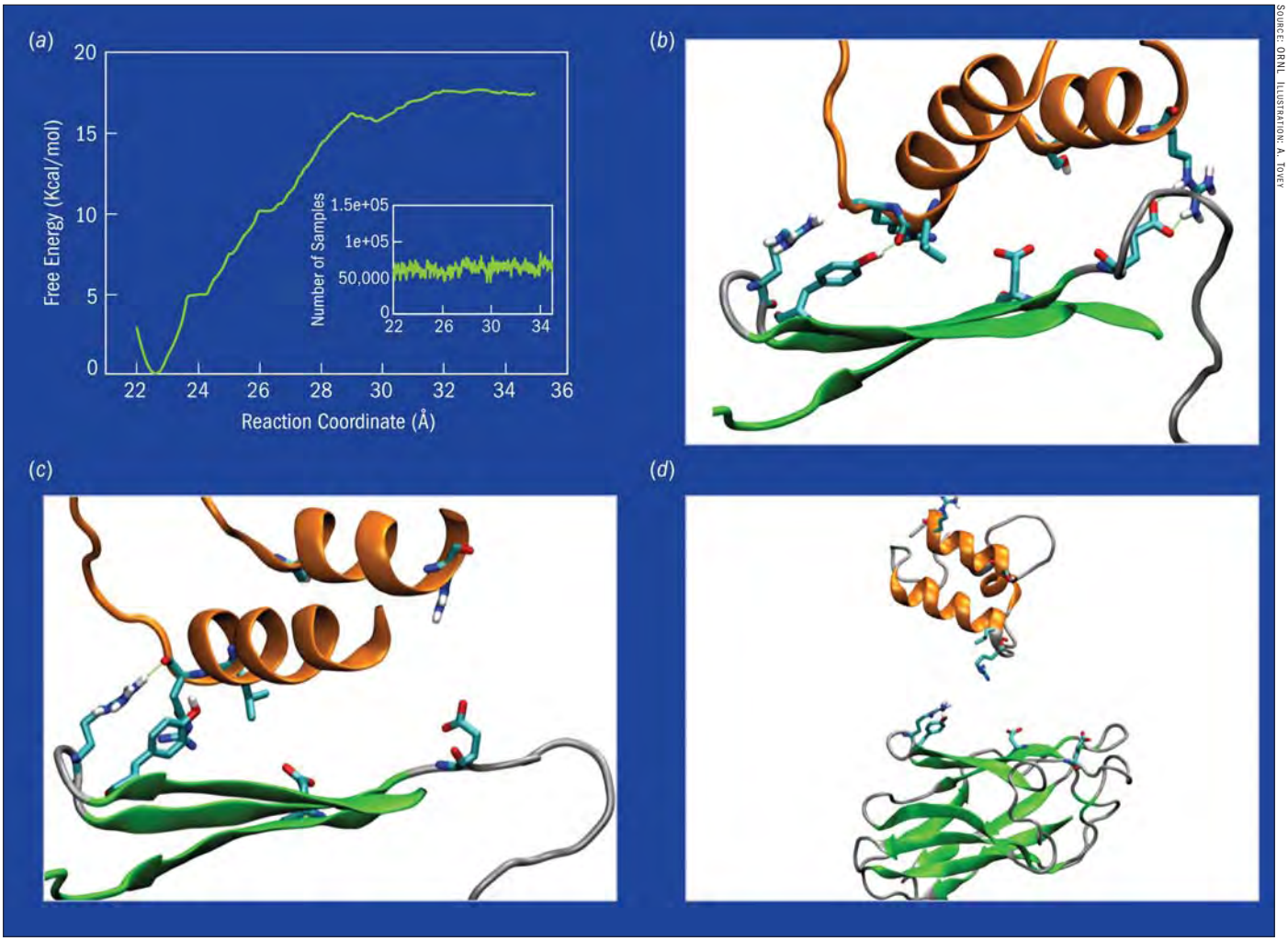
The work under way described above is directed at providing a foundation for the accurate multiscale simulation of biological systems, with a special focus on application to lignocellulosic biomass. On scales ranging from those of chemical reactions and interactions between pairs of atoms up to microbial–biomass systems-level modeling, high-performance computer simulation will provide detailed information on factors influencing biomass recalcitrance in cellulosic ethanol production. The research process will be iterative; as much relevant experimental information as possible will be fed into the simulation models, and the models will be refined in the light of experimental results. The hope is simulation will also be able to unify experiment by rationalizing disparate experimental snippets of information into a coherent multiscale picture. Indeed, simulation can provide a federating role: combining known experimental data with underlying physical principles to form a consistent picture and make specific suggestions regarding future experiments.

We are now further developing the methodology with the specific aims of applying REACH force fields to the simulation of lignocellulosic biomass and parallelizing the simulation code for petaflop/s supercomputers.

Atomistic simulation results provide insights essential to design processes in cellulosic ethanol. For example, the data presented above on protein–protein recognition in cellulosomes can now, in turn, be used to guide protein modifications in order to engineer binding. Efforts are under way at NREL to design engineered cellulosomal modules that can conduct more efficient biomass degradation than corresponding wild-type protein complexes. In conjunction with appropriate biochemical and biophysical experiments, both atomic-detail and coarse-grained computer simulations are expected to provide a foundation for understanding the principles of domain synergy and cellulosomal activity. This foundation will allow the rational, structure-based design of improved cellulosomal assemblies for cellulosic ethanol production. Likewise, the understanding of the principles of cellulose hydrolysis by cellulases may allow rational design of more efficient enzymes or even synthetic mimics.

Powerful DOE supercomputers that came on line in 2009, such as the Jaguar Cray XT5 at ORNL's National Center for Computational Sciences (NCCS), can in principle enable atomistic simulations of million-atom biomass systems on timescales up to the microsecond. This technology will have specific ramifications for the analysis of a wide variety of experimental data.

Of particular interest is the use of state-of-the-art materials science experimental facilities in the DOE complex. For example, biomass data emerging from synchrotron sources, and from neutron sources such as the Spallation Neutron Source (SNS) and High Flux Isotope Reactor (HIFR) at ORNL, will require simulation analysis. Neutron and X-ray scattering from complex systems (such as many biological, chemical, or materials science samples) cannot be unequivocally analyzed using simplified analytical models. Rather, a simulation model is required from which scattering quantities — such as structure factors — are calculated and compared with experiment, thus allowing interpretation of the experimental results and design of future experiments. CMB researchers are working with Barbara Evans, Dean Myles, and their ORNL colleagues to interpret biomass neutron data. Comparison with experimental data also permits the refinement of the simulation force fields and protocols.



SOURCE: ORNL ILLUSTRATION: A. TOVEY

Figure 13. *(a) A free energy profile for the dissociation of cohesin and dockerin domains. (b) A snapshot of the cohesin–dockerin complex at reaction coordinate value X = 24 Å. (c) A snapshot at X = 27 Å. (d) A snapshot of the cohesin–dockerin complex in the dissociated state, that is, X > 30 Å. The two alpha-helices; beta-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.*

The atomistic simulation work will be extended to longer time scales and larger length scales by developing and applying first-principles, self-consistent multiscale simulation techniques. In this way a seamless framework can be created, allowing the researcher to create a systems-level simulation of the interactions of lignocellulosic plant cell wall variants with microbial interfaces, cellulases, and cellulosomes, and to zoom in on crucial atomistic processes at will.

Research into the origins of biomass recalcitrance is certainly timely. At the time of this writing, roughly 300 million gallons of planned commercial-scale cellulosic ethanol plants are in various stages of planning and development across the country. Furthermore, a study released earlier this year by DOE's Sandia National Laboratories found large volumes of cellulosic biofuels could be produced from already identified biomass sources and resources without displacing crop production. The study indicated that even without incentives, cellulosic biofuels could potentially compete with gasoline with oil prices of between $70 and $90 per barrel by 2030, given the expected accelerated development of technology and feedstocks. A combination of technological development and economic opportunity may lead to those recalcitrant plants giving up their sugars after all.

Contributor Jeremy C. Smith, UT/ORNL Governor's Chair and Director, ORNL Center for Molecular Biophysics

Acknowledgements The author acknowledges funding from the BioEnergy Science Center (BESC). BESC 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 author is the recipient of supercomputing awards from the 2008 and 2009 DOE INCITE program and the National Science Foundation Teragrid.