FIGURE 2:
Illustration of a cellulosome domain acting over cellulose fibers. Cellulosomes are highly-efficient molecular machines that can degrade plant fibers. Addapted from Cover of Schöler et. al. [10].

CELLULOSOME STRUCTURE DETERMINATION EMPLOYING ATOMISTIC SIMULATIONS COMBINED TO EXPERIMENTAL ASSAYS

Allocation: Illinois/680Knh
PI: Isaac Cann[1]
Co-PI: Rafael C. Bernardi[1]
Collaborators: Klaus Schulten[1], Edward Bayer[2], Hermann Gaub[3], and Michael Nash[3]

[1]University of Illinois at Urbana-Champaign
[2]Weizmann Institute
[3]Ludwig Maximilian University

EXECUTIVE SUMMARY

Cellulosomes are multi-enzyme complexes that target the deconstruction of cellulose and hemicellulose in anaerobic cellulosome-containing bacteria. Briefly, in cellulosome assembly, a large noncatalytic polypeptide called the scaffoldin, embedded with various cohesins (Coh), anchors dockerin (Doc)-containing enzymes through Coh–Doc interactions. The precision of the Coh–Doc interaction allows the addition of different catalytic cellulases and hemicellulases onto the scaffoldin that may or may not be bound to another domain attached to the cell wall. Cellulosomes' ability to efficiently degrade plant-cell-wall biomass allows them to be used in the second-generation biofuel industry, which aims to use agricultural waste to produce ethanol. Furthermore, the recent discovery of cellulosomal bacteria in the lower gut of humans is paradigm-shifting as it has allowed demonstration of the capacity to degrade both hemicellulose and cellulose, at least in some humans. Our project employs molecular dynamics simulations, complementing single-molecule and biochemistry experiments, to characterize the structure of cellulosomes.

INTRODUCTION

Bacteria play a key role in the second-generation biofuel industry since their cellulolytic enzymes, used for plant-cell-wall degradation, are employed in the production of these advanced biofuels. Also, symbiont bacteria greatly influence human health and play a significant role in pathogenesis, disease predisposition, physical fitness, and dietary responsiveness. Here we are investigating key processes underlying bacterial activity, namely, plant fiber metabolism. Specifically, we are examining the structure and function of cellulosomes, the highly cooperative macromolecular complex that is central to this metabolic process in some bacteria.

Cellulosomes are multi-enzyme complexes that target the deconstruction of cellulose and hemicellulose in anaerobic cellulosome-containing bacteria. Integration of cellulosomal components occurs via highly ordered protein–protein interactions among three major components. Briefly, in cellulosome assembly, a large noncatalytic polypeptide called the scaffoldin, embedded with various Cohs, anchors Doc-containing enzymes through Coh–Doc interactions (Fig. 1). Specificity of the cohesin–dockerin interaction allows incorporation of different catalytic cellulases and hemicellulases onto the scaffoldin that may or may not be bound to another domain tethered to the cell wall. Cellulosome assembly promotes the exploitation of enzyme synergism because of spatial proximity and enzyme-substrate targeting [1].

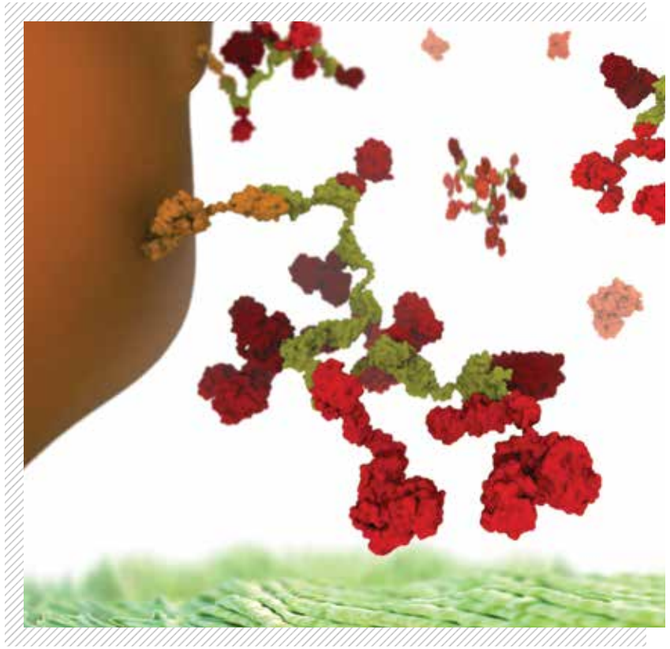
Combined with biochemical and single molecule experiments, we employed molecular dynamics (MD), steered MD (SMD) and generalized simulated annealing (GSA) [2] simulations on Blue Waters. Utilizing QwikMD [3], our new intuitive "point and click" graphical interface connecting visual MD (VMD) [4] and nanoscale molecular dynamics (NAMD) [5], we are studying the detailed mechanism of cellulase complexes, in particular, cellulosomes. Using stochastic search algorithms connected to molecular dynamics tools, we are building the **first** comprehensive structure of a cellulosome. Employing GSAFold/NAMD we were already able to obtain the structure of a cellulosome scaffoldin and, using Blue Waters, we are working to determine the structure of a whole cellulosome complex, including enzymatic domains. We expect that a complete model of cellulosome's structure will shed light on the mechanism that allows these enzymatic complexes to be highly efficient.

METHODS & RESULTS

As mediators in the interactions between comparatively large bacterial cells and cellulose particles, scaffolds, and carbohydrate-binding molecule (CBM) domains are critical cellulosomal components (Fig. 2). As many cellulosomal habitats (for example, cow rumen) exhibit strong flow gradients, shear forces will accordingly stress bridging scaffold components mechanically *in vivo*. Protein modules located at stressed positions within these networks are likely to be preselected for high mechanostability, as demonstrated by our group [6]. However, thus far little is known about cellulosome structure as they are formed by interactions of Doc and Coh that are connected by flexible linkers made of proteins with just a few to more than a thousand amino acids. The many flexible linker regions allow for very complex structural dynamics [7].

Using stochastic search algorithms coupled to NAMD we can generate thousands of different structure conformations for the cellulosome. GSA[2] analysis shows that the different linkers between Coh (and a CBM) in cellulose-integrating protein A (CipA) scaffoldin assume a different number of more stable conformations. Small angle X-ray

FIGURE 1:
Illustration of a complete cellulosomal structure. The scaffoldin (yellow) can be attached by a specific Doc-Coh interaction to a cell-anchoring domain (in orange). Another Doc-Coh interaction is responsible for attaching the enzymatic domains (red) to the scaffoldin. Addapted from Cann et. al. [1].



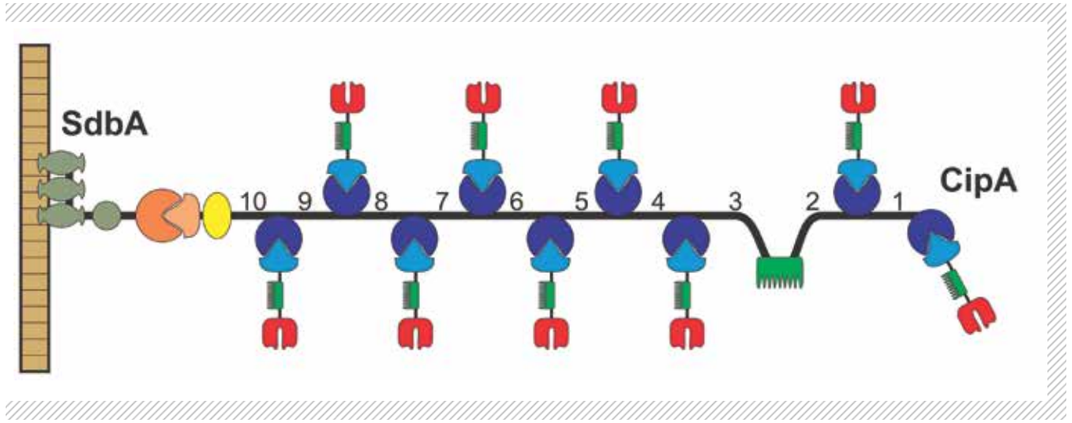


FIGURE 3: Organization of Clostridium thermocellum cellulases and hemicellulases in the SdbA/CipA cellulosome. The C. thermocellum scaffoldin (CipA) contains one CBM (Green) and nine type I cohesins (Dark Blue) and thus organizes a multiprotein complex with nine enzymes (Red). The C-terminal type II dockerin (Pink) domain of CipA binds specifically type II cohesin domains (Orange) found in cell-surface proteins. The CipA linkers already studied using GSAFold/NAMD integration are numbered.

scattering (SAXS) analysis has previously shown that three conformations are observed for linker 10 (Fig. 3). GSAFold is capable of predicting these three conformations and all the other conformations for CipA. To perform this analysis, 20,000 conformations were obtained per linker and clustered. Combined, these linker conformations would give us 1043 CipA conformations. From clustering, we reduce this number to 3888 structures that were obtained and also subjected to a cluster analysis that gave rise to the five most significant structures.

Following well-established protocols for large macromolecular systems [8,9], and using one of the CipA conformations that we obtained using GSAFold, we built a **first** model of an entire cellulosome structure. MD simulations are now employed to study the quaternary structure stability.

WHY BLUE WATERS

Investigating the structure and functional processes of large enzymatic complex machineries, such as the cellulosomes, is only possible on petascale computing resources, such as Blue Waters. Structures obtained using enhanced sampling techniques, such as GSA, are only reliable if thousands of conformations (models) are predicted. Employing GSA for the numerous linkers of the cellulosome is a well-suited task for the large-scale parallel architecture of Blue Waters.

NEXT GENERATION WORK

Our primary goal is to obtain a clear picture of the cellulosome structure at work. For that, long molecular dynamics simulations of different cellulosomes, some of them with hundreds of millions of atoms, will have to be performed. To investigate the enzymatic mechanism in the context of the cellulosome, hybrid quantum mechanics (QM)/molecular mechanics simulations will have to be performed using multiple QM regions that require massive computer power. Such complex study might only be feasible in a few years, requiring pre-exascale and exascale systems.

PUBLICATIONS AND DATA SETS

Schoeler, C., et al., Ultrastable cellulosome-adhesion complex tightens under load. *Nat. Commun.* 5 (2014), p. 5635.

Bernardi, R.C., M.C.R. Melo, and K. Schulten, Enhanced Sampling Techniques in Molecular Dynamics Simulations of Biological Systems. *Biochim. Biophys. Acta.*, 1850 (2015), p. 872.

UNDERSTANDING BIOMOLECULAR STRUCTURE AND DYNAMICS BY OVERCOMING BARRIERS TO CONFORMATIONAL SAMPLING

Allocation: NSF PRAC/2.00 Mnh
PI: Thomas Cheatham[1]
Co-PI: Adrian Roitberg[2], Carlos Simmerling[3], and David Case[4]
Collaborators: Darrin York[4], and Shantenu Jha[4]

[1]University of Utah
[2]University of Florida
[3]Stonybrook University
[4]Rutgers University

EXECUTIVE SUMMARY

Large ensembles of independent molecular dynamics, running optimized AMBER code on Blue Waters' GPUs, enable full sampling of the conformational ensemble of biomolecules, including DNA helices, RNA tetranucleotides, and RNA tetraloops. This allows detailed validation and assessment of enhanced sampling approaches and biomolecular force fields and provides detailed insight into biomolecular structure, dynamics, interactions, and function. The ensemble simulations currently being performed are possible only on computational hardware with large numbers of GPUs. While today our simulations are pushing the state of the art, such large simulations will become routine within a few years. The even larger and more powerful parallel resources available in the near future will enable molecular dynamics simulations to probe more relevant biological time scales (milliseconds to seconds) and to study larger biomolecular assemblies more completely.

INTRODUCTION

Biomolecular simulation—although known as a powerful tool for probing the structure, dynamics, interactions, and functions of proteins and nucleic acids for over 40 years—is really coming of age thanks to access to large-scale computational resources such as Blue Waters. Not only can simulations be applied to larger biomolecular assemblies, but for modest sized biomolecules the community has demonstrated the ability to fold proteins *de novo* and to fully sample the conformational distributions of various nucleic acid motifs. A challenge is parallel scaling since, for a fixed system size, adding additional cores does not increase performance. To overcome this, the community has moved toward application of ensemble methods and application of various enhanced sampling methodologies that couple together independent molecular dynamics (MD) simulation engines. AMBER, a suite of programs for biomolecular simulation whose latest version, AMBER 16, was released in April 2016, has been highly optimized for use on GPUs. The optimized GPU code, and the ensembles that are

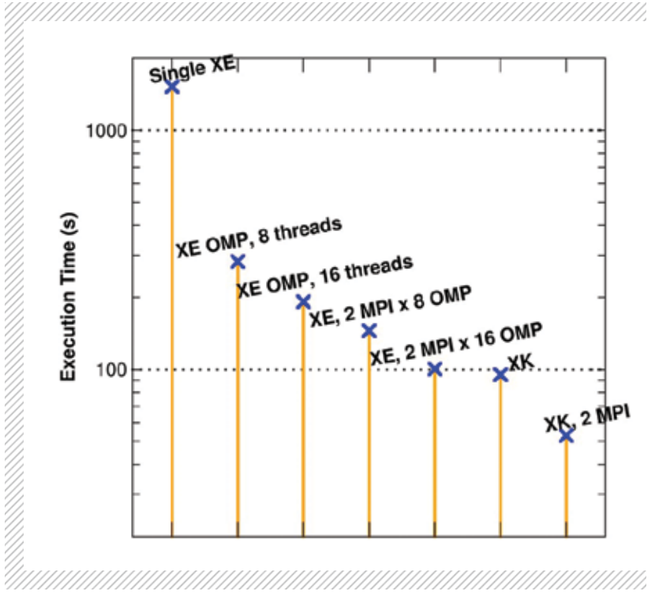


FIGURE 1: Relative performance of CPPTRAJ on different nodes when determining the 965 closest solvent molecules out of 15,022 to 4,143 solute atoms from a 2,000 frame MF trajectory (no imaging) using various parallelization modalities, including CUDA on the XK nodes.