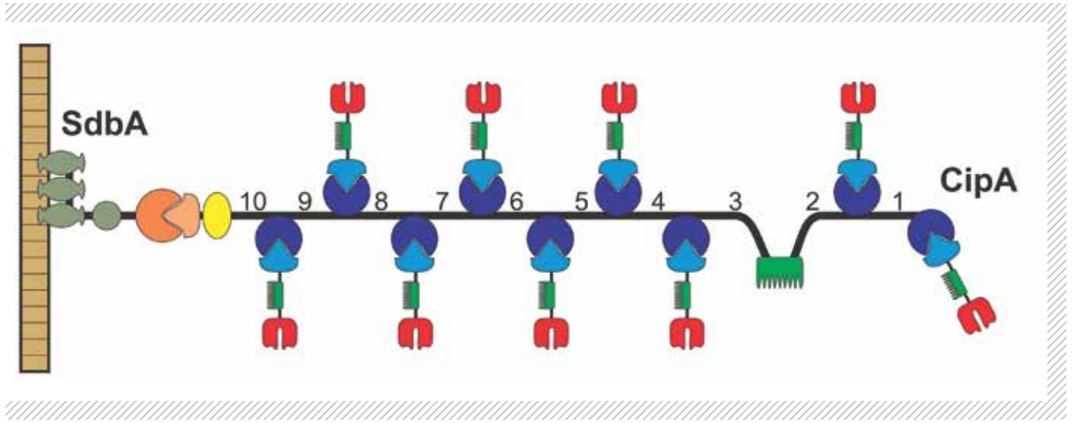


FIGURE 3: Organization of Clostridium thermocellum cellulases and hemicellulases in the SdbA/CipA cellulosome. The C. thermocellum scaffoldin (CipA) contains one CBM (Green) and nine type I cohesins (Dark Blue) and thus organizes a multiprotein complex with nine enzymes (Red). The C-terminal type II 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

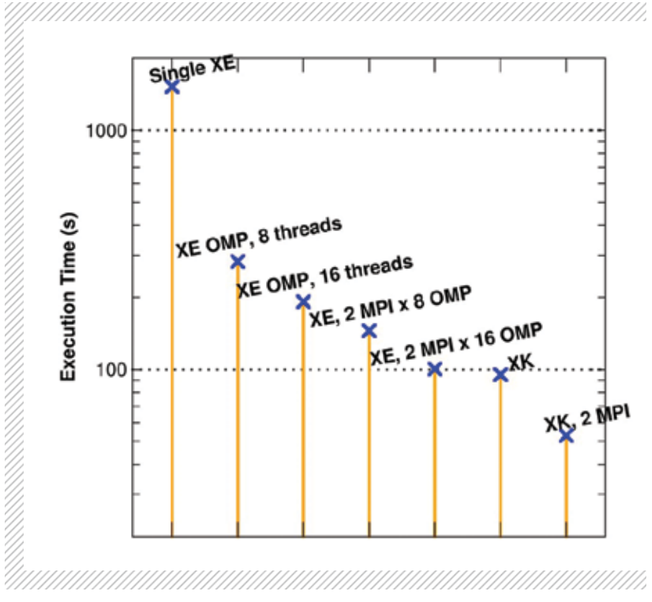
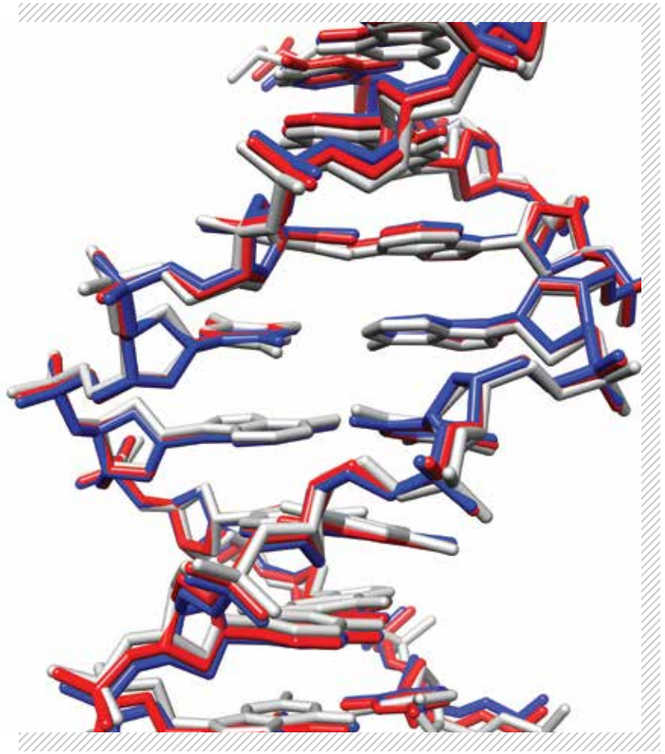


FIGURE 1: Relative performance of CPPTRAJ on different nodes when determining the 965 closest solvent molecules out of 15,022 to 4,143 solute atoms from a 2,000 frame MF trajectory (no imaging) using various parallelization modalities, including CUDA on the XK nodes.



produced, provide a powerful means to assess and validate the available force fields and to apply these methodologies to give novel biological insight into protein and nucleic acid structure and dynamics. Improving the codes, methods, and force fields while also pushing ensemble methods to their limits are critical research activities since these tools are being used by an ever-increasing pool of researchers throughout the world.

METHODS & RESULTS

The coupling together of independent molecular dynamics simulations into loosely coupled ensembles to enhance conformational sampling is an increasingly used modality for applications in biomolecular simulation. If you peruse any recent journal in the field, including *The Journal of Chemical Theory and Computation, The Journal of Computational Chemistry,* and *The Journal of Physical Chemistry*, among others, you will see multiple publications developing and applying ensemble methods. With a variety of methodological variations and names ranging from replica-exchange MD, metadynamics, swarms, and Markov state modeling to constant pH MD and lambda dynamics, all of the approaches promise more efficient means to explore structural ensembles, free energy pathways, and kinetics. Although these techniques are promising, there is no free lunch, since as the size of the biomolecule increases, sampling a complete ensemble takes longer and longer. Various methods attempt to speed the process through applications of enhanced sampling in different degrees of freedom. However it is often difficult to verify claims of efficiency, especially with method variants implemented into vastly different code bases. Therefore, we have explored making our ensemble data available to the community online (http://amber.utah.edu) to allow other researchers to directly compare our results to results obtained using different ensemble approaches and to assess relative convergence and efficiency.

The combination of large ensemble methods with the availability of many fast GPUs on Blue Waters leads to an explosion in data. In order to process vast amounts of data efficiently, the AMBER MD trajectory analysis code CPPTRAJ has been modified to include multiple levels of parallelism, aided by a Petascale Application Improvement Discovery (PAID) collaboration with Blue Waters. Specifically, CPPTRAJ now implements four levels of parallelism. MPI parallelism over ensemble instances (with sorting of data across all ensembles), MPI parallelism over reading/writing of trajectory files in a given ensemble, OpenMP parallelism for time-intensive analyses, and most recently GPU parallelization of time-intensive analyses involving calculations of a large number of distances. Relative performance is shown in Fig. 1.

Key results of our ensemble approaches are described in greater detail in the listed publications and range from accurate modeling of magnesium-dependent conformational changes in RNA, correct modeling of proteins binding and modulating DNA structure, and optimization, assessment, and validation of nucleic acid force fields. This includes fully converging the conformational and dynamic distribution of the Dickerson-Drew dodecamer with average structures over milliseconds of aggregated MD data less than 0.8 Å from the published experimental structures as shown in Fig. 2.

WHY BLUE WATERS

Within the NSF ecosystem of computational resources, Blue Waters is the GPU-optimized resource with a sufficiently large set of GPUs to allow ensembles on the 300-3,000 scale (assuming a single ensemble instance per GPU). Our team has shown the ability to converge the conformational ensemble of an RNA tetraloop with multidimensional replica exchange; using 360 GPUs, this requires about 2-3 microseconds of MD simulation per ensemble instance, or approximately five to 10 days of continuous MD simulation on those resources.

NEXT GENERATION WORK

Ensemble-based biomolecular simulation methods will continue to evolve in terms of their generality and power. With next-generation computational resources, the community will be able to not only study larger biomolecular systems but also to more fully sample and converge the accessible conformational space of these systems. While at present we can aggregate to milliseconds of effective sampling, to reach biological time scales we still need orders of magnitude greater sampling (to seconds and beyond) in simulations that likely will require multiple GPUs or accelerators to reach system sizes of hundreds of thousands to millions of atoms and beyond. The kind of ensemble analyses we are currently undertaking, made possible by Blue Waters, will be effectively routine by 2019-20 and will enable the larger community while we push the edge of what is possible on future petascale resources.

PUBLICATIONS AND DATA SETS

http://amber.utah.edu

Galindo-Murillo, R., J.C. Garcia-Ramos, L. Ruiz-Azuara, T.E. Cheatham, III, and F. Cortes-Guzman. Intercalation processes of copper complexes in DNA. *Nuc. Acids Res.* 43 (2015) pp. 5364-5376.

Bergonzo, C., N. Henriksen, D.R. Roe, and T.E. Cheatham, III. Highly sampled tetranucleotide and tetraloop motifs enable evaluation of common RNA force fields. *RNA* 29 (2015) pp. 1578-1590.

Bergonzo, C. and T.E. Cheatham, III. Improved force field parameters lead to a better description of RNA structure. *J. Chem. Theory Comp.* 11 (2015) pp. 3969-3972.

Bergonzo, C., K.B. Hall, and T.E. Cheatham, III. Stem-loop V of Varkud satellite RNA exhibits characteristics of the Mg^{2+} bound structure in the presence of monovalent ions. *J. Phys. Chem. B* 119, (2015) pp. 12355-12364.

Robertson, J.C. and T.E. Cheatham, III. DNA backbone BI/BII distribution and dynamics in E2 protein-bound environment determined by molecular dynamics simulation. *J. Phys. Chem. B* 119 (2015) pp. 14111-14119.

Galindo-Murillo, R., D.R. Davis, and T.E. Cheatham, III. Probing the influence of hypermodified residues within the $tRNA_3^{Lys}$ anticodon stem loop interacting with the A-loop primer sequence from HIV-1. *Biochemica Biophys. Acta* 1860 (2016) pp. 607-617.

Dissanayake, T., J. Swails, M. Harris, A.E. Roitberg, and D. York. Interpretation of pH-activity Profiles for Acid-Base Catalysis from Molecular Simulations. *Biochemistry* 54 (2015) pp 1307-1313.

Hopkins, C., S. LeGrand, R.C. Walker, A.E. Roitberg. Long Time Step Molecular Dynamics through Hydrogen Mass Repartitioning. *J. Chem. Theory Comp.* 11 (2015) pp. 1864-1874.

Alvarez, L., A. Louis Ballester, A.E. Roitberg, D. Estrin, S.-R. Yeh, M. Marti, L. Capece. Structural study of a flexible active site loop in human indoleamine 2,3-dioxygenase and its functional implications. *Biochemistry* 55 (2016) pp. 2785-2793.

Nguyen, H., A. Perez, S. Bermeo, and C. Simmerling. Refinement of generalized Born implicit solvation parameters for nucleic acids and their complexes with proteins. *J. Chem. Theory Comp.* 11 (2016) pp. 3714-3728.

Maier, J.A., C. Martinez, K. Kasavajhala, L. Wickstrom, K.E. Hauser, and C. Simmerling. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comp.* 11 (2015) pp. 3696-3713.

Lai, C.-T., H.-J. Li, W. Yu, S. Shah, G.R. Bommineni, V. Perrone, M. Garcia-Diaz, P.J. Tonge, and C. Simmerling. Rational Modulation of the Induced-Fit Conformational Change for Slow-Onset Inhibition in Mycobacterium tuberculosis InhA. *Biochemistry* 54 (2015) pp. 4683-4691.

FIGURE 2 (LEFT): Overlap of the average structures, omitting hydrogens and the two terminal base pairs on each end from nearly milliseconds of aggregated MD simulation data from 100 independent 11 microsecond length MD simulations (omitting the first 2 microseconds) with the parmbsc1 (blue) and AMBER ff15 (or ol15, red) force fields compared to the PDB average structure from 1NAJ (gray).