DI

GA

TN

BI

USING LARGE ENSEMBLES OF AMBER MOLECULAR DYNAMICS SIMULATIONS TO REPRODUCIBLY AND RELIABLY CONVERGE THE CONFORMATIONAL DISTRIBUTIONS OF NUCLEIC ACIDS

Allocation: NSF PRAC/12,000 Knh
PI: Thomas Cheatham[1]
Co-PIs: Adrian Roitberg[2], Carlos Simmerling[3], David Case[4]
Collaborators: Darrin York[4], Rodrigo Galindo Murillo[1], Daniel Roe[5], Christina Bergonzo[6], Niel Henriksen[7], Hamed Hayatshahi[8]

[1]University of Utah
[2]University of Florida
[3]Stony Brook University
[4]Rutgers University
[5]National Institutes of Health
[6]National Institute of Standards and Technology
[7]Atomwise, Inc.
[8]University of North Texas Health Science Center

EXECUTIVE SUMMARY

The AMBER simulation codes and force fields allow reliable and accurate modeling of the atomistic structure and dynamics of biomolecular systems. The molecular dynamics codes within AMBER have been highly optimized on GPUs. Taking advantage of the GPUs and using large ensembles of independent but coupled molecular dynamics simulations, we have demonstrated the ability to reliably and reproducibly converge and sample the conformational ensembles of biomolecules including DNA helices and RNA dinucleotides, tetranucleotides, and tetraloops. More efficient or enhanced sampling is enabled by applying multidimensional replica exchange methodologies. The ability to fully sample or converge the conformational ensemble allows detailed validation and assessment of enhanced sampling approaches and of the biomolecular force fields, providing detailed insight into biomolecular structure, dynamics, interactions, and function. However, even more computational power is required to continue this work and push toward larger and more complex biomolecular assemblies on longer timescales.

RESEARCH CHALLENGE

Molecular dynamics simulations are a widely applied methodology for elucidating the structure, dynamics, and interactions among biomolecules. This can help us understand biomolecular function, biomolecular assemblies, and also to probe ligand–receptor interactions and biomolecular folding. Key issues include being able to sample conformational space efficiently and effectively and also to accurately model the molecular interactions with appropriate biomolecular force fields. Using the Blue Waters petascale supercomputer, we have demonstrated the ability to overcome sampling limitations to fully elucidate the conformational ensemble of DNA duplexes and RNA dinucleotides, tetranucleotides, and tetraloops.

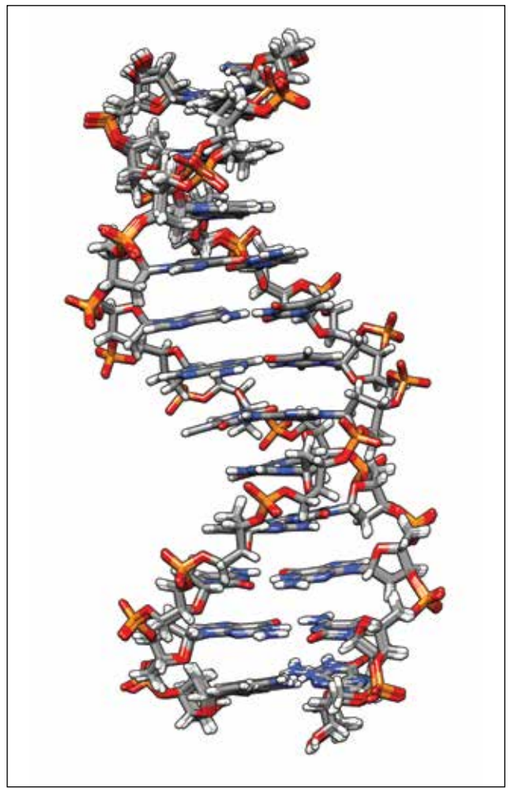

Figure 1: The image shows an overlay of six average structures of the Drew–Dickerson DNA dodecamer (CGCGAATTCGCG)2 from a 15-microsecond simulation (using the 5 to 10 µs region, all-atom RMS fit) using six different salt combinations: LiCl, LiBr, NaCl, NaBr, KCl, and KBr, each of 200mM concentration.

METHODS & CODES

As the team continued to develop and optimize the AMBER suite of molecular dynamics (MD) and free energy set-up, simulation, and analysis codes and force fields, these codes were applied on Blue Waters to understand the structure, dynamics, and interactions of biomolecules in their native environment. The AMBER molecular dynamics engine PMEMD is well-optimized in CUDA for high performance on GPUs. Application of enhanced sampling methodologies, including multidimensional replica exchange molecular dynamics (M–REMD) with ensembles of independent MD simulations that exchange information periodically—including temperature and Hamiltonian changes—provides a very efficient means of reproducibly converging the conformational ensembles of various biomolecular systems. As the ensembles of independent MD simulations generate big data in the form of time series of atomic positions or trajectories that must be sorted and analyzed, we parallelized the CPPTRAJ trajectory analysis code to enable efficient processing of the large data on Blue Waters' parallel file system.

RESULTS & IMPACT

Using M–REMD methods, we have demonstrated the ability to reliably and reproducibly converge the conformational ensembles of DNA helices and RNA dinucleotides, tetranucleotides, and tetraloops. This means that we can elucidate the thermally accessible set of conformations for a given biomolecule at a given temperature. The AMBER biomolecular force fields, particularly the parmbsc1 and X_{OL3} (contained within the ff14SB AMBER force field designation), perform incredibly well for nucleic acid helices. Fig. 1 shows the remarkable agreement observed in the central base pairs with RMS deviations of less than 0.1 Angstroms between all six conformations of 5 μs average structures of the DNA duplex $d(CGCGAATTCGCG)_2$ from MD simulations in different salts. The agreement to the average NMR structures from the highest resolution structure of this system (1NAJ) is less than 0.5 Angstroms with both of the high-performing AMBER force fields (parmbsc1, ff14SB) in OPC water. Note that some disruption of structure is observed with the first and second base pairs (and even into the third base pair) at both ends of the DNA, which shows a deviation of ~1 Angstrom due to fraying and base-flipping events that occur on the microsecond timescale. This observation emphasizes the importance of convergence for these particular systems and the lack of structural impact on the DNA regardless of the salt used to neutralize the system. The research enabled by Blue Waters also demonstrated convergence of the dynamic conformational ensemble of various RNA molecules; multiple conformations are populated over the course of the M–REMD simulations. However, some of these are incompatible with experiment, suggesting that the AMBER force fields are over-populating anomalous conformations. These are likely due to very subtle misbalances in the RNA force field in terms of nonbonded interactions (van der Waals and hydrogen bonding) that still require optimization to improve the force fields. This work is underway currently.

WHY BLUE WATERS

The Blue Waters petascale resource allowed our team to converge the conformational ensembles of various nucleic acid systems for the first time, and this was enabled by the developments in both the AMBER codes and force fields, and by the exceptional performance of the AMBER PMEMD engine on GPUs. The thousands of GPUs and well-performing Lustre parallel file system facilitated our simulation and analysis workflow. The Blue Waters team helped us overcome problems and facilitated some CUDA, OpenMP, and MPI optimizations of the CPPTRAJ analysis code.

PUBLICATIONS & DATA SETS

The AMBER WWW pages: http://ambermd.org.

Data sets from some runs on Blue Waters: http://amber.utah.edu.

Galindo-Murillo, R., D.R. Roe, and T.E. Cheatham III, On the absence of intrahelical DNA dynamics on the µs to ms timescale. *Nature Commun.*, 5:5152 (2014), DOI:10.1038/ncomms6152.

Galindo-Murillo, R., D.R. Roe, and T.E. Cheatham III, Convergence and reproducibility in molecular dynamics simulations of the DNA duplex d(GCACGAACGAACGAACGC). *Biochimica Biophys. Acta*, 1850 (2015), pp. 1041–1058.

Cheatham, T.E., III, and D.R. Roe, The impact of heterogeneous computing on workflows for biomolecular simulation and analysis. *Computing in Science and Engineering*, 17:2 (2015), pp. 30–39.

Galindo-Murillo, R., et al., Intercalation processes of copper complexes in DNA. *Nuc. Acids Res.*, 43 (2015), pp. 5364–5376.

Bergonzo, C., et al., 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.

Waters, J.T., et al., Transitions of double-stranded DNA between the A- and B- forms. *J. Phys. Chem. B*, 120 (2016), pp. 8449–8456.

Bergonzo, C., K.B. Hall, and T.E. Cheatham III, Divalent Ion Dependent Conformational Changes in an RNA Stem-Loop

Continued on page 215

IMPROVING NWCHEM SCALABILITY USING THE DATASPACES FRAMEWORK

TN

Allocation: Innovation and Exploration/525 Knh
PI: Gregory Bauer[1]
Collaborators: Victor Anisimov[1], Manish Parashar[2], Melissa Romanus[2]

[1]National Center for Supercomputing Applications
[2]Rutgers University

EXECUTIVE SUMMARY

Molecular Dynamics (MD) is the major computational technique to bridge the gap between experiment and theory in materials science, engineering, and biomedical research. However, the predictive ability of MD simulations strongly depends on the quality of underlying parameters. The purpose of the present phase of this project is to develop a parameter optimization tool for the AMBER classical force field for DNA, with the potential for extending the underlining methodology to optimization of other popular force fields. The project employs the previously generated data set of experimental quality base–base interaction energies prepared by conducting high-level quantum-mechanics CCSD(T) computations in NWChem on molecular clusters extracted from experimental crystallographic data for DNA bases. The parameter optimization procedure performs a grid-based scan on a set of parameters by trying all their possible combinations, computing the interaction energy for each grid point using the Amber force field, and comparing the result to the reference interaction energy. The computation runs in parallel and returns a set of parameters that best reproduces the target data.

RESEARCH CHALLENGE

The present computational challenge is handling a combinatorial explosion in problem size that accompanies the task of identifying the global minimum in parameter space. The natural solution to this challenge is to employ parallelization. However, the use of a large number of parallel processes in the grid search exacerbates the I/O load on the filesystem since each process frequently performs read and write operations. That places a limit on the number of parallel tasks that can be practically used in the computation without overloading the filesystem. When the grid search is done, each process carries a multidimensional array of results that associates the point in the parameter space with the optimization function. The aggregate distributed array holds billions of records for an average parameter optimization problem. Sorting an array of such size to determine the promising parameter sets for further analysis represents a practically unsolvable problem that requires a creative solution.

BI

METHODS & CODES

Generation of target data for parameter optimization uses a set of in-house scripts to extract molecular clusters from the experimental crystallographic data of DNA bases. Quantum mechanical computations, which follow, determine the base–base interaction energy in the molecular clusters. The computation employs the previously optimized CCSD(T) method [1] in the NWChem package [2]. A scalable tool to optimize Lennard–Jones parameters in the AMBER (Assisted Model Building with Energy Refinement) force field to fit the parameters to intermolecular interaction energies for experimental geometry of monomers has been developed. It has been tested to run on 16,384 XE nodes using 32 cores per node resulting in the use of 524,288 processing units on Blue Waters. The optimization tool generates an adjustable number of alternative parameter sets of comparable quality for further testing in molecular dynamics simulations.

RESULTS & IMPACT

This project introduces a procedure for systematic improvement of classical force field by determining the global minimum in the parameter space for an expandable set of the training data. The beneficiary of the optimized parameter set is the entire molecular dynamics community. As the number and quality of the training data increase with time, rerunning the parameter optimization tool will deliver the improved parameter set. The developed fractional parallel sorting procedure drastically reduces time spent in sorting as well as the required RAM per node. The use of RAM disk for read / write operations on compute nodes eliminates the filesystem overhead and makes the code applicable to compute systems beyond Blue Waters' size.

WHY BLUE WATERS

Blue Waters, with its fast interconnect and large memory per core, is unique in its ability to conduct CCSD(T) computations of molecular systems encountering a thousand basis functions, which is vital for the success of the developed parameter optimization procedure. Since the parameter optimization procedure is extremely resource demanding, the availability of large numbers of nodes is essential for the exhaustive exploration of parameter space.

Continued from page 213

Observed by Molecular Dynamics. *J. Chem. Theory Comp.*, 12 (2016), pp. 3382–3389.

Galindo-Murillo, R., et al., Assessing the current state of Amber force field modifications for DNA. *J. Chem. Theory Comp.*, 12 (2016), pp. 4114–4127.

Heidari, Z., et al., Using Wavelet Analysis to Assist in Identification of Significant Events in Molecular Dynamics Simulations. *J. Chem. Inf. Model.*, 56 (2016), pp. 1282–1291.

Hao, Y., et al., Molecular basis of broad-substrate selectivity of a peptide prenyltransferase. *PNAS*, 113 (2016), pp. 14037–14042.

Hayatshahi, H.S., et al., Computational Assessment of Potassium and Magnesium Ion Binding to a Buried Pocket in the GTPase-Associating Center RNA. *J. Phys. Chem. B*, 121 (2017), pp. 451–462.

Zgarbova, M., et al., Influence of BII Backbone Substates on DNA Twist: A Unified View and Comparison of Simulation and Experiment for all 136 Distinct Tetranucleotide Sequences. *J. Chem. Info. Model.*, 57 (2017), pp. 275–287.

Wang, Y., et al., Application of thiol-yne/thiol-ene reactions for peptide and protein macrocyclizations. *Chemistry*, 23 (2017), pp. 7087–7092.

Hayatshahi, H.S., C. Bergonzo, and T.E. Cheatham III, Investigating the ion dependence of the first unfolding step of GTPase-associating center ribosomal RNA. *J. Biomol. Struct. Dyn.*, 1:11 (2017), pp. 243–253.

Bergonzo, C., and T.E. Cheatham III, Mg^{2+} binding promotes SLV as a scaffold in Varkud Satellite Ribozyme SLI-SLV kissing loop junction. *Biophys. J.*, 113 (2017), pp. 313–320.

Galindo-Murillo, R., and T.E. Cheatham III, Computational DNA binding studies of (-)-epigallocatechin-3-gallate. *J. Biomol. Struct. Dyn.*, 3 (2017), pp. 1–13 (2017).

Hayatshahi, H.S., N.M. Henriksen, and T.E. Cheatham III, Consensus conformations of dinucleotide monophosphates described with well-converged molecular dynamics simulations. *J. Chem. Theory Comp.*, 14 (2018), pp. 1456–1470.

Cornillie, S.P., et al., Computational modeling of stapled peptides toward a treatment strategy for CML and broader implications in the design of lengthy peptide therapeutics. *J. Phys. Chem. B*, 122 (2018), pp. 3864–3875.

Galindo-Murillo, R., T.E. Cheatham III, and R.C. Hopkins, Exploring potentially alternative non-canonical DNA duplex structures through simulation. *J. Biomol. Struct. Dyn.*, (2018) DOI:10.1080/07391102.2018.1483839.