Allocation Application Details

PI Nickname:	cheatham
Department:	Center for High Performance Computing
PI Full Name:	Thomas Cheatham
Campus Phone:	801-587-9652
Project Title:	Insight into biomolecular structure, dynamics, and interactions from simulation
Requester:	Thomas Cheatham
Quick Allocation:	No
Status:	Submitted
Start Date:	04/01/2016
End Date:	03/31/2017
All Users Have Access to Full Award:	Yes
Custom Allotment:	
General Allocation Pool	
Spring 2016 requested:	250,000
Summer 2016 requested:	250,000
Fall 2016 requested:	250,000
Winter 2017 requested:	250,000

Proposal

General Description of Project:

Our studies aim to push the boundaries of atomistic simulation methods to give insight into the structure, dynamics, and interactions of biomolecules in their native environment and also drug-receptor interaction mainly on proteins and nucleic acids. As developers of the AMBER biomolecular simulation suite of software, we have access to the latest code and methods developments. HPC resources at CHPC are used mainly to set-up, equilibrate, and analyze simulation results. [Additionally, a couple of my users are Iranian nationals and therefore cannot get access to some of the national resources we have available.] Our main production resources are on NSF national resources through large allocations of computer time on XSEDE (however the XSEDE allocation has currently lapsed) and also Blue Waters. Our calculations, in collaboration with various different experimental groups at Utah and beyond, provide reliable insight into various biomolecular systems and our focus is on assessment and validation of the simulation results.

Significance and Expected Impact:

Generally, we aim to expose limitations in the commonly applied potentials (or force fields) and overcome these. These methods are in wide use by the community in a large series of research projects funded by all sorts of agencies throughout the world. As an AMBER developer (one of the most widely used MD simulation codes for biomolecules), it is important that we prove the validity and utility of the methods. Our research projects span from basic science (assessing and validating the results) to more applied projects including the application of computer-aided drug-design methods to proteins and nucleic acids and the design of novel drug delivery agents and are supported by multiple grants. In the past couple of years we have been able to reproducibly demonstrate convergence of the conformational ensembles of the internal portion of a DNA helix, of RNA tetra nucleotides, and RNA tetra loops with multiple different force fields. The results suggest that none of the currently available RNA force fields adequately populate the experimental geometries, however this year we have improved the force fields in a couple of ways that are showing increasing promise, for example increasing the population of the "experimental" loop structure of UUCG more than any other currently available force field. The ability to now routinely and reliably reproducibly (from different initial conditions) converge the conformational distributions of biomolecules is a significant advance.

Numerical Techniques / 3rd Party Software:

We generally apply the “AMBER” (http://amber.scripps.edu) biomolecular simulation package. I am one of the AMBER developers and have access to early pre-release (optimized) developer’s versions and am also the primary author of the main analysis package, “ptraj”. We also utilize NAMD, Dock, Autodock, and CHARMM in our research and also Gaussian. The only CHPC licensed software that we make use of (beyond the software to job schedule, compilers and other system software) is Gaussian. We have a valid license for CHARMM37 and make use of the AMBER developers Git tree for access to AMBER.

Computational Results and Analysis/Presentation Methodology:

Since we have access to considerable HPC resources outside of CHPC, our use of CHPC resources is focused on the initial and post-production phases of simulation (and also for prototyping and/or experimenting with the methods or for production simulations by the Iranians). The initial phase involves setting up the simulations and performing a series of simulations to equilibrate the results. Production is then normally performed on external resources and then the data is brought back to CHPC for detailed analysis. As one of the major disk hogs at CHPC--our group has purchased more than 400 TB is disk space at CHPC-- these resources and HPC cycles to perform analysis are critical. Key foci of the lab are to properly assess and validate the simulation results in comparison with experiment, to seek out problems with molecular mechanical force fields and to improve them, and also to provide means to facilitate searching, exploring and sharing of the raw and derived MD simulation data (http://ibiomes.chpc.utah.edu and also at http://amber.utah.edu).

New and special features of AMBER include an incredibly fast and accurate GPU implementation of Ewald MD simulation--the Cheatham lab owns half the GPU nodes on ember and also has a Kepler K20 (kepler.chpc.utah.edu)--and also the development of efficient multi-dimensional replica exchange methods that allow 10's to 1000's of independent MD simulations to run and periodically

exchange information to increase the efficiency of conformational sampling. Additionally, we have made considerable updates to the CPPTRAJ MD trajectory analysis programs (available in Amber Tools and on GitHub), including its parallelization with MPI and OpenMPI, with scripts used to generate analysis in our papers provided in supporting information.

Parallelism of Code:

The main MD simulation engines are parallelized with MPI; there are two versions including a fully open source less efficient and general purpose simulation engine (sander) and a more parallel and efficient more limited engine (PMEMD) [that is not open source; this has the GPU code]. Both scale very efficiently to a node, with sander scaling dropping off beyond ~16-32 cores and PMEMD beyond 64 cores. However, both codes can be run in ensemble mode effectively allowing multiple MD instances to be run simultaneously, thereby effectively scaling up to an arbitrary number of processors. For details about the efficiencies, see http://ambermd.org and in particular the GPU page. Note that the MD implementation on GPUs uses a mixed single and fixed point precision model and that the current code is completely deterministic (sequential on GPU) and therefore is an excellent code for testing GPUs for problems!

The analysis code CPPTRAJ is now 3-way parallelized for time consuming tasks using OpenMP and MPI. OpenMP is used for computationally demanding analysis tasks, and we now MPI parallelize not only across ensembles (i.e. independent sets of simulations) but in reading in the MD trajectory files. One task we make use of is sorting of ensembles of trajectories (for example from temperature replica exchange to sort either into replica trajectories or temperature trajectories). The new code, to be formally released with AMBER16 in April 2016 and also available now in the CPPTRAJ GitHub page, provides super-linear speed-ups on some parallel file systems. We look forward to testing on the new DDN kingspeak Lustre scratch space.

Existing / Outside Computer Resources:

NSF XSEDE MCA01S027 (1/01/15-12/31/15)
- 3.4M core hours on Stampede @ TACC
- 70K core hours on Blacklight @ PSC
- 6.1M core hours on Comet @ SDSC
- 10TB data supercell
- 500K core hours on Maverick
- 50TB on Data Oasis

[This has lapsed now; we will submit a renewal for a smaller amount given our massive allocation on Blue Waters - likely for specialized resources (big memory on Bridges; multi-GPUs where available; Knights Landing).]
NSF PRAC OCI-1036208 (through 2018)
- 12M node hours on the Blue Waters Petascale Resource + 2M supplement for Ebola project. At the time of award in September 2015, this was the largest allocation on Blue Waters.

At CHPC:
- dedicated use of a subset of the lonepeak.chpc.utah.edu nodes (formerly telluride); this is also used for teaching and occasionally by the sequencing/informatics core)

- over 400 TB of spinning disk
- 12 M2090 GPUs on 6 nodes of ember
- ~2 front-end nodes

Local resources are used primarily for setup and analysis of results run on the national production resources. Also for smaller production runs and use by the Iranian students.

Resources Required:

We requested an arbitrary allocation of 250K node hours per quarter. This is at the level we were allocated last year. Given that the machines are still over-subscribed and we and CHPC have not recently bought new nodes, I realize that 250K is a large request. Any time that is awarded will be used efficiently and productively to further the research of our lab and collaborators. It would be nice to be among the larger allocations, however I understand that given over-subscription and due to our access to considerable national and local resources that this may be unfeasible.

Sources of Funding:

Note: The list below looks like a lot of funding! In reality, it is not as large as it seems since many of the grants are in no-cost extension or are rather small. We have pending grants and are frantically submitting new ones to keep the lab alive for more than 1 more year.

- National Science Foundation, ACI-1443054 (10/01/14-9/30/19) CIF21 DIBBS: Middleware and high performance analytics libraries for scalable data science. PI: Fox, Co-PIs: Wang, Qiu, Jha, Marathe; Cheatham is significant personnel. [~$125K to Cheatham lab, total cost]

- National Science Foundation, ACI-1341034 (10/01/13-9/30/16) CC-NIE Integration: Science slices converting network research innovation into enhanced capability for computational science and engineering at the University of Utah. PI: Corbato, Co-PIs: Bolton, van der Merwe, Ricci, Cheatham [small salary support to Cheatham]

- National Science Foundation, CHE-1266307 (10/01/13-9/30/16) CDS&E: Tools to facilitate deeper data analysis, exploration, management, and sharing of ensembles of molecular dynamics trajectory data. PI: Cheatham [$300K total cost, in NCE, funding gone by 6/30/16]

- National Institutes of Health, R01 GM098102 (9/30/11-8/31/15) RNA-ligand interactions: simulation and experiment. M-PIs: Cheatham, Kathleen Hall (Wash U, contact), Carlos Simmerling (Stony Brook).
[~$125K/yr direct cost, in NCE, funding gone by 6/30/16]

- National Science Foundation, OCI-1440031 (9/01/14-8/31/18) PRAC – Hierarchical molecular dynamics sampling for assessing pathways and free energies of RNA catalysis, ligand binding, and conformational change. PI: Cheatham, Co-PIs: Simmerling (Stony Brook U), Roitberg (UFl), Case (Rutgers), and York (Rutgers) [$40K total cost, travel only]

- NSF Cyberinfrastructure Partnership / TeraGrid / XSEDE (1/01/15-12/31/15) XRAC/LRAC MCA01S027: Insight into biomolecular structure, dynamics, interactions and energetics from simulation. PI: Cheatham, Computer time award: ~10M core hours awarded in 2015, award since 2002. Currently lapsed; may submit new one in March 2016.

- NSF ACI (4/1/15-3/30/17) RAPID: Optimizing experimental approaches to Ebola membrane fusion inhibitor design through high-throughout biomolecular simulation workflows on Blue Waters . PI: Cheatham [$200K total cost; in NCE]

Publications based on CHPC Resources Work Results (include full citation):

(102) R Galindo-Murillo, DR Roe, and TE Cheatham, III. "Convergence and reproducibility in molecular dynamics simulations of the DNA duplex d(GCACGAACGAACGAACGC)." Biochimica Biophys. Acta 1850, 1041-1058 (2015) doi: 10.1016/j.bbagen.2014.09.007.

(103) TE Cheatham, III and DR Roe. "The impact of heterogeneous computing on workflows for biomolecular simulation and analysis." Computing in Science and Engineering 17:2, 30-39 (2015).

(104) R Galindo-Murillo, JC Garcia-Ramos, L Ruiz-Azuara, TE Cheatham, III, and F Cortes-Guzman. "Intercalation processes of copper complexes in DNA." Nuc. Acids Res. 43, 5364-5376 (2015).

(105) C Bergonzo, N Henriksen, DR Roe, and TE Cheatham, III. "Highly sampled tetranucleotide and tetraloop motifs enable evaluation of common RNA force fields." RNA 29, 1578-1590 (2015).

(106) AC Simmonett, FC Pickard IV, Y Shao, TE Cheatham, III and BR Brooks. "Efficient treatment of induced diploles." J. Chem. Phys. 143, 074115 (2015).

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

(108) C Bergonzo, KB Hall, and TE Cheatham, III. "Stem-loop V of Varkud satellite RNA exhibits characteristics of the Mg2+ bound structure in the presence of monovalent ions." J. Phys. Chem. B 119, 12355-12364 (2015). PMC4634716

(109) R Galindo-Murillo and TE Cheatham, III. "Using information about DNA structure and dynamics from experiment and simulation to give insight into genome-wide association studies." Chapter 4 in Translation Cardiometabolic Genomic Medicine, edited by A Rodriguez-Oquendo, p 81-98 (2015).

(110) JC Thibault, DR Roe, K Eilbeck, TE Cheatham, III, and JC Facelli. "Development of an informatics infrastructure for data exchange of biomolecular simulations: Architecture, data models,

and ontology." SAR and QSAR in Environmental Research DOI: 10.1080/1062936X.2015.1076515 (2015).

(111) JC Robertson and TE Cheatham, III. "DNA backbone BI/BII distribution and dynamics in E2 protein-bound environment determined by molecular dynamics simulation." J. Phys. Chem. B 119, 14111-14119 (2015).

(112) M Zgarbova, J Sponer, M Otyepka, TE Cheatham, III, R Galindo-Murillo, and P Jurecka. "Refinement of the sugar-phosphate backbone torsion beta for the AMBER force fields improves the description of Z-DNA and B-DNA." J. Chem. Theory Comp. 11, 5723-5736 (2015).

(113) R Galindo-Murillo, DR Davis, and TE Cheatham, III. "Probing the influence of hypermodified residues within the tRNA3Lys anticodon stem loop interacting with the A-loop primer sequence from HIV-1." Biochemica Biophys. Acta 1860, 607-617 (2016).