Request for Supplemental Allocation

Principal Investigator: Shantenu Jha[1,2]
Co-Principal Investigator: Joohyun Kim[1]
Co-Principal Investigator: Yaakoub El Khamra[3]
[1]*Center for Computation & Technology, Louisiana State University, Baton Rouge, USA*
[2]*Rutgers, State Univeristy of New Jersey, USA*
[3]*Texas Advanced Computing Center TACC, University of Texas, Austin, USA*

01 April 2011

1 Summary

We would like to request an urgent supplemental allocation on TeraGrid resources. In the first year of our multi–year allocation (TG-MCB090174: *Scale-Up and Scale-Out of Ensemble-based Simulations*) we requested 5 million SUs on Ranger and 3 million SUs on Kraken. Our allocation was awarded half the requested SUs on Ranger (2.5 million) and 2 out of 3 million SUs on Kraken.

As we have pursued our science unhindered, we have run-out of SUs on Ranger half-way through the year, and are running dangerously low on SUs on Kraken. We document progress made along multiple fronts in the short period of time. As it stands, our research will stall in the first week of April and remain that way for several months. We kindly request 1.5 million SUs on Kraken and 1.5 million SUs on Ranger to tide us over to the next allocation renewal cycle, when we will be eligible to request an advance on our second–year allocation. If one of those machines is over-subscribed, we kindly request equivalent SUs on the other. We present some important reasons underpinning our request.

2 Project Progress

2.1 Project 2: Understanding non-coding Functional RNAs: Folding Dynamics and Binding mechanism of Riboswitches

In order to fully understand the coupling between the ligand binding and the folding of riboswitch RNAs, we adopt all atomic simulations to explore this linkage. We have been engaged in seeking physical insights into novel ideas that elucidate the interplay between the Aptamer domain and the Expression domain; this has not been explored in previous studies. In particular, we are using very long time-scales ($>$ 100ns) all-atom multiple MD simulations/trajectories, to seek an understanding of the role of the branch migration during a dynamical transition toward the OFF state of S-adenosyl Methione (SAM) binding riboswitch (SAM-I riboswitch). Based upon interesting phenomenon that has been observed in trajectories generated on Ranger, we have developed new analysis to monitor the trajectory on the fly. Additionally, these data also provide us the basis of choosing a system to be submitted for tens of us (microsecond) time scale on Anton, a machine specially designed for running MD simulations. This work is being carried out by Wei Huang a final year PhD (expected graduation Dec 2011) student co-supervised by the PI. The computing resources we ask here are critical for our ongoing study on RNA riboswitch mechanisms.

We need to extend this to cover multiple distinct starting configuration; the need for multiple trajectories arises because we have to simulate various initial structures that are sampled by 3D modeling and thus cover the conformational space appropriately. In addition, these multiple all-atom MD simulations should be conducted with SAM-bound and SAM-free states, since the role of SAM could be understood by comparing trajectories of two different states.

Initial work in this project has just been accepted as a Special Issue of Concurrency and Computing: Practice and Experience (CCPE) for Emerging Methods for the Life Sciences [1]. The next round of simulations, should provide conclusive evidence of a potential role of the SAM in facilitating P1 helix formation over the AT helix formation, which will eventually clarify whether the proposed branch migration mechanism as a major switching pathway between two alternative secondary structures of SAM-I riboswitch.

2.2 Project 3: Atomistic Simulations of Physiological Systems

Both sub-projects described here employ ensembles of short simulations to achieve superior statistical sampling compared to a single simulation of equivalent duration.

2.2.1 Project 3a: Towards patient specific HIV therapy

Progress in ongoing work: The long term scientific objective of our project is to develop molecular dynamics simulations of HIV-1 Pol enzymes into a tool for clinicians to use in determining the cocktail of drugs to administer to an HIV infected individual. We have recently completed a study which applies our free energy calculation protocol [2] to a patient derived HIV-1 protease (PR) sequence and the drug lopinavir, identified as producing ambiguous resistance rankings from currently used clinical decision support tools [3]. This study has suggested a potential new mechanism for drug resistance. We have completed studies of the protonation state of the protease catalytic dyad when bound to all FDA approved HIV-1 PR inhibitors. This is a prerequisite for the application of our protocol to these drugs. With a further allocation we would look to extend this work to compare wildtype and known resistant mutants for each drug.

We have also extended our protocol to investigate the binding of drugs to HIV-1 RT. In this system we have identified the impact of large scale protein motions, distant from the binding site, on the binding free energy [4].

Next Steps: A further extension of our program of work has involved preliminary studies of the viability of the prototype foamy virus integrase as a model system for investigating resistance in the HIV-1 integrase. This is necessary as no complete structure of the HIV-1 intrasome is available. Our studies indicate that calculated free energies are largely independent of motions of the N terminal domain, which exhibits only low sequence similarity with the HIV homolog, and the DNA substrate [5]. With further computational resources we would seek to compare the binding thermodynamics of the PFV wildtype with a sequence containing mutations inserted at locations experimentally identified as causing resistance in both PFV and HIV.

2.2.2 Project 3b: Predicting the affinity of the EGFR kinase domain for drug inhibitors of lung

Progress in ongoing work: Our research aims at creating molecular level simulators which have an impact in personalized drug treatment of targeted therapy. The epidermal growth factor receptor (EGFR) is a major target for drugs in treating lung carcinoma since it promotes cell growth and tumor progression. Structural studies have demonstrated that EGFR exists in an equilibrium between catalytically active and inactive forms, and dramatic conformational transitions occur during its activation. It is known that EGFR mutations promote such conformational changes which affect its activation and drug efficacy. Using TeraGrid resources, we have been doing two simulations: one is to study changes in drug binding affinities due to cancer mutations of EGFR using ensemble molecular dynamics simulations [6, 7, 8], the other to address activation mechanism of key proteins involved in cancer development and treatment, including EGFR and the GTPase KRAS.

Next Steps: We are performing extended timescale molecular dynamics simulations to study the structural and energetic properties of KRAS at both active and inactive conformations. Although studies have provided insights into the structural basis for KRAS activation, the energetic aspects of the conformational changes are not fully understood. We have completed about 250ns simulations so far for wild-type and mutant KRAS within both active and inactive states. To investigate the changes of conformational equilibrium between the two conformational states, substantially longer timescale molecular dynamics simulations are needed. We plan to extend the simulations to 500ns each. The study will be able to reveal the relative stabilities of active and inactive conformations and hence the KRAS activation.

2.3 Project 4: Large-scale molecular dynamics simulations of layered bio-mineral composites

2.3.1 Effects of Varying Mineral Surface Charge on RNA Adsorption, and Rate of Folding

We are currently utilizing TeraGrid resources to progress simulations, which build on our previous publications in the biomaterials domain. Previous simulations performed on TeraGrid resources elucidated the mechanism for adsorption of RNA on montmorillonite surfaces and showed the increased rate of RNA folding into biologically important secondary structures [9]. Our current simulations have been developed to observe the effects of the clay mineral surface charge on the adsorption and folding of the RNA oligomers. We are simulating large layered double hydroxide (LDH) surfaces with single-stranded RNA molecules of differing sequences [10]. These large ensembles of models consist of hundreds of thousands of atoms that we hope to progress beyond the standard simulation times of current bioinorganic MD simulations.

2.3.2 Enhanced folding of large RNA ribozymes using Replica Exchange Molecular Dynamics

We have shown that RNA folds much more rapidly on clay than in a bulk aqueous environment, but in complex oligomers such as RNA molecules, with many degrees of conformational freedom, folding of the molecule into functional tertiary structures happens over relatively long timescales, which are greater than a typical molecular dynamics simulation (10-100ns). Replica exchange molecular dynamics (REMD) overcomes this limitation through the use of multiple molecular dynamics simulations (replicas) in parallel at multiple temperatures. High-temperature replicas enable rapid barrier crossing and sample additional configurations, which are unlikely to be observed in conventional room temperature molecular dynamics simulations. Periodically, replicas attempt to exchange temperatures according to a Metropolis-like criterion, thereby allowing low-energy configurations to be sampled. Results of our earlier simulations have motivated experimental research in the folding of RNA/DNA on clay minerals at the University of Edinburgh

Next Steps: We propose to use the enhanced sampling available through REMD to explore the conformations of the hammerhead ribozyme absorbed on montmorillonite clay surface. Despite the computationally intensive nature of REMD, it will allow us to determine the role the clay surface plays in mediating hammerhead RNA folding into the most catalytically active tertiary structure. Such a scenario would not be possible with conventional molecular dynamics.

Layered nanomaterials consist of mineral layers, such as montmorillonite clays, separated by polymeric or organic material, the thickness of which is of the order of nanometres. The mechanism by which water and other small molecules penetrates (intercalates) between the clay sheets is unknown. To estimate the free energy of intercalation and elucidate the mechanism, we used thermodynamic integration methods. To speed up these expensive simulations, we used multiple replicas of the system at different values of the umbrella constraint and run concurrently using the MD code LAMMPS. We have used between 16-32 replicas of the structure per simulation, with approximately 32 cores per replica. A publication detailing these results is currently in preparation [11]

2.4 Project 5: Expeditions in Distributed Computing using SAGA

2.4.1 Analysis of Data-Intensive Applications on the TeraGrid

Next-Generation (gene) Sequencing (NGS) machines produce unprecedented amounts of data. In addition to the challenge of data-management that arise from unprecedented volumes of data, there exists the important requirement of effectively analyzing large volumes of data. It is worth mentioning that the computational complexity of the analysis (e.g. mapping) depends, upon other things, the size and complexity of the reference genome & the data-size of short reads. Given that these can vary significantly, the computational requirements of NGS-analytics also varies (even between data-sets of similar size). Thus an efficient, scalable and extensible analytical approaches must be supported by any framework supporting NGS-analytics.

We have created the DARE-NGS Gateway (`http://cyder.cct.lsu.edu/dare-ngs`) which supports Genome-wide analysis on the TeraGrid and other distributed cyber-infrastructure. DARE-NGS builds upon the Distributed Adaptive Runtime-Environment (DARE) Framework, which support a range of tasks with varying computing and

data requirements over a wide range of high-performance and distributed infrastructure. Using DARE-NGS we have analyzed the full Human-Genome (requiring data-sets of upwards of 250GB) on TeraGrid machines such as Ranger. This is work in progress, done entirely in the period of this award period; we are still extending the DARE-NGS capabilities to support other advanced algorithms[12].

2.4.2 Frameworks for Supporting the Scale-up and Scale-out of Ensembles based Simulations

In Ref. [13], we have continued to develop the capabilities to perform arguably the world's largest number of replica-exchange simulations using the SAGA Repex-Framework. Most experiments, validation and refinements have been performed in the award period. Important refinements in the communication and coordination approaches remain, as well as in the ability to scale-out to even more resources/cores. This is the proposed area of activity over the next phase so that several application research groups can use this framework (eg Bishop (Tulane), Ron Levy (Rutgers) and Darrin York (Rutgers)), as well as for Project 2 and 4.

We have used the underlying framework that is being developed and tested using this allocation, to support the project "Running Many MD Simulations on Many Supercomputers" – a collaboration with Tom Bishop as part of Bishop's TRAC award. This work is scheduled for submission to TeraGrid 2011 Conference (and to Journal of Chemical Informatics and Modeling).

3 The Case for Continuity

As outline above, we have delivered impressive scientific and technological advances in the short period since this grant was awarded. Importantly, we are on the trajectory that we were aiming for and are set to deliver on the goals that we hoped the allocation would facilitate. Specifically, the uninterrupted continuation is important as, (i) Project 2 forms the basis for *specialized* runs on the DE Shaw Anton machine, to which we will have access starting in Q2 of 2011, (ii) Projects 3 is an important component of the International Interoperability Project (between TeraGrid and DEISA), (iii) Project 5 will lead to the timely delivery of an infrastructure that in turn will be used by multiple biomolecular simulation groups on the TeraGrid for efficient and effective execution of ensemble-based simulations.

It is our hope that our scientific progress and trajectory will not be perturbed; additionally, it is important to mention that several graduate dissertations and papers critically depend upon non-disruption. In the next 3-6 months, we anticipate 3 Graduate student led publications and 2 theses (1 PhD (Wei Huang) and 1 Masters (Thota)) based upon a continued allocation.

Type of Calculation	Method or Package	HPC Resources To Be Used	SUs required
Atomistic MD Simulation MM-PBSA	NAMD AMBER	Ranger/Kraken	1000K
AEE/EGFRs (50K atoms)	NAMD	Ranger/Kraken	800K
Erlotinitb/EGFRs	NAMD	Ranger/Kraken	700K
Clay Edge Simulations, including NEMD and RNA montmorillonite	LAMMPS	Ranger/Kraken	500K
Total SUs		Ranger/Kraken	3000K

Table 3 shows the current status of the main stages of the projects in the allocation and the estimated computational requirements. Details on the estimate process can be found in the initial allocation request. Since we can use Ranger and Kraken interchangeably, a total sum of 3 million SUs equally distributed between both machines would be ideal. However, if either machine is oversubscribed we can make do with a higher allocation on the other.

In conjunction with the scientific questions we are addressing, we are also involved in the development of a run-time execution system (DARE) that uses the Simple API for Grid Applications (SAGA: http://saga.cct.lsu.edu/) and SAGA-based Pilot-Job (BigJob) to allow the running and coordination of hundred if not thousands of large-scale ensembles across resources, both on the TeraGrid and the EU DEISA network, as part of the NSF-HPCOPS funded Interoperability Project. Significant progress has recently been achieved in allowing the use of SAGA to interoperate between these two different grids. A key goal of an extended allocation would be to further develop this infrastructure and assess performance using real scientific workloads, and make it available for the broader & larger community of biomolecular simulators.

References

[1] J. Kim, W. Huang, S. Maddineni, F. Aboul-ela, and S. Jha, "Energy landscape analysis for regulatory rna finding using scalable distributed cyberinfrastructure," *accepted for publication, Special Issue of Concurrency and Computing: Practise and Experience (CCPE) for Emerging Methods for the Life Sciences*, 2011.

[2] S. K. Sadiq, D. W. Wright, O. A. Kenway, and P. V. Coveney, "Accurate ensemble molecular dynamics binding free energy ranking of multidrug-resistant hiv-1 proteases," *Journal of Chemical Information*, vol. 50, no. 5, pp. 890–905, 2010.

[3] D. W. Wright and P. V. Coveney, "An ensemble molecular dynamics investigation of ambiguous resistance rankings for hiv-1 protease bound to lopinavir," *In preparation*, 2011.

[4] D. W. Wright, S. K. Sadiq, P. Kellam, and P. V. Coveney, "The impact of ligand binding on the dynamics of hiv-1 reverse transcriptase," *In preparation*, 2011.

[5] M. Bem, D. W. Wright, , and P. V. Coveney, "Molecular dynamics investigation of the prototype foamy virus integrase as a model for hiv-1 resistance," *In preparation*, 2011.

[6] S. Wan and P. V. Coveney, "Rapid and accurate ranking of binding affinities of epidermal growth factor receptor sequences with selected lung cancer drugs," *Journal of the Royal Society Interface*, 2011.

[7] ——, "Molecular dynamics simulation reveals structural and thermodynamic features of kinase activation by cancer mutations within the epidermal growth factor receptor," *Submitted*, 2011.

[8] K. Marias, D. Dionysiou, V. Sakkalis, N. Graf, R. Bohle, P. V. Coveney, S. Wan, A. Folarin, P. Buchler, M. Reyes, G. Clapworthy, E. Liu, J. Sabcznski, T. Bily, A. Roniotis, M. Tsiknakis, E. Kolokotroni, S. Giatili, C. Veith, E. Messe, H. Stenzhorn, Y.-J. Kim, S. Zasada, A. Haidar, C.May, S. Bauer, T. Wang, Y. Zhao, M. Marasek, R. Grewer, A. Franz, and G. Stamatakos, "Clinically driven design of multi-scale cancer models: the contracancrum project paradigm," *Journal of the Royal Society Interface Focus*, 2011.

[9] J. B. Swadling, P. V. Coveney, and H. C. Greenwell, "Clay minerals mediate folding and regioselective interactions of rna: A large-scale atomistic simulation study," *Journal of American Chemical Society*, 2010.

[10] J. Swadling, P. Coveney, and H. Greenwell, "Stability of free and mineral-protected nucleic acids: Implications for the rna world," *Submitted*, 2011.

[11] J. L. Suter and P. Coveney, "Unraveling the mechanism of mechanism of water intercalation in clays using molecular dynamics," *In preparation*, 2011.

[12] J. Kim, S. Maddineni, and S. Jha, "Characterizing deep sequencing analytics using bfast: Towards a scalable distributed architecture for next-generation sequencing data," *Submitted*, 2011.

[13] A. Thota, A. Luckow, and S. Jha, "Efficient large-scale replica-exchange simulations on production infrastructure," *Accepted for Philosophical Transactions of the Royal Society of London A*, 2011.