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 to request 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 at unhindered, we have out of SUs on ranger half way through the year and are running quite 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 million SUs on ranger to tide us over to the next allocation renewal cycle, when we will be eligible to ask for an advance on our second–year allocation. We present some important reasons underpinning our request.

2 Project Progress

2.1 Project 2:

CCPE

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. We are involved the development of a workflow that uses the Simple API for Grid Applications (SAGA: http://saga.cct.lsu.edu/)[1] to allow the running and coordination of these ensembles across resources both on the TeraGrid and the EU DEISA network. 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.

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].

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

cancer

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 tumour 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], the other to address activation mechanism of key proteins involved in cancer development and treatment, including EGFR and the GTPase KRAS.

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

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.

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

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

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 analysed 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[3].

In Ref [1], 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.

References

[1] "Efficient Large-Scale Replica-Exchange Simulations on Production Infrastructure", Accepted for Philosophical Transactions of the Royal Society of London A (2011)

[2] "Energy Landscape Analysis for Regulatory RNA Finding using Scalable Distributed Cyberinfrastructure" J. Kim, W. Huang, S. Maddineni, F. Aboul-ela, and S. Jha

[3] Characterizing Deep Sequencing Analytics Using BFAST: Towards a Scalable Distributed Architecture for Next-Generation Sequencing Data J Kim, S. Maddineni and S. Jha