NEED A NEW TITLE

# 1 Summary

We propose to use multiple XSEDE resources both in concurrent usage mode, as well as individual resources to study several scientific problems. This work is built on the extensive efforts over the past three years we have carried out with a wide range of computational science and computer science projects, requiring a few groups to use multiple resources on the XSEDE concurrently (for loosely-coupled simulations). Specifically, in this proposal we request 7.5M SUs for three distinct projects: (i) Understanding non-coding Functional RNAs, (ii) Atomistic Simulations of Physiological Systems and (iii) Large-scale molecular dynamics simulations of layered bio-mineral composites. The projects for which computer time is being requested are all funded projects – mostly at the National/International level, and some by local resources.

Additionally, the request for 7.5M SUs in this proposal is based upon the projected science problems as outlined below as well as a proven track record of *successfully* utilising more than 7.2M SUs in the past 18 months (Project TG-MCB090174 and its supplements). As is always the case, we did not artificially inflate the computational requirement and we continue to be on-schedule in terms of SU utilization. Last cycle we were allocated only half of the requested SUs and ran out half way through the year. We have also moved most of our work to underutilized resources (Kraken at NICS) in hope that we receive the full amount requested to continue to conduct funded research.

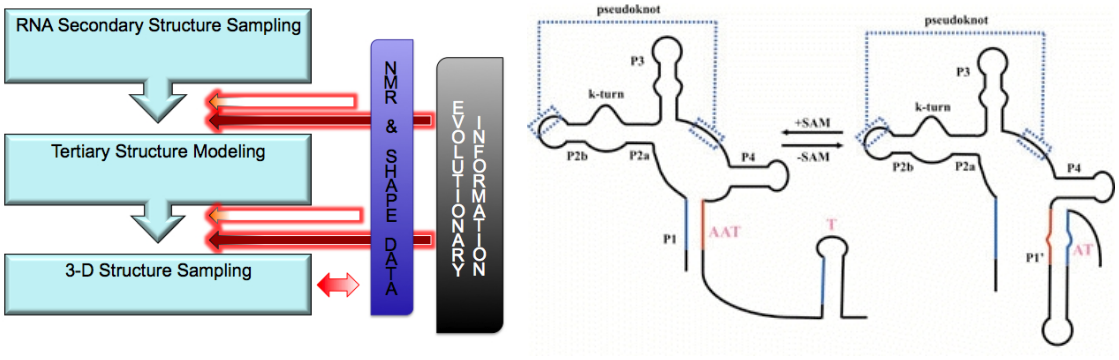
# 2 Continuing Project: Understanding non-coding Functional RNAs: Folding Dynamics and Binding mechanism of Riboswitches

## 2.1 Project Progress

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.




(a) (b)

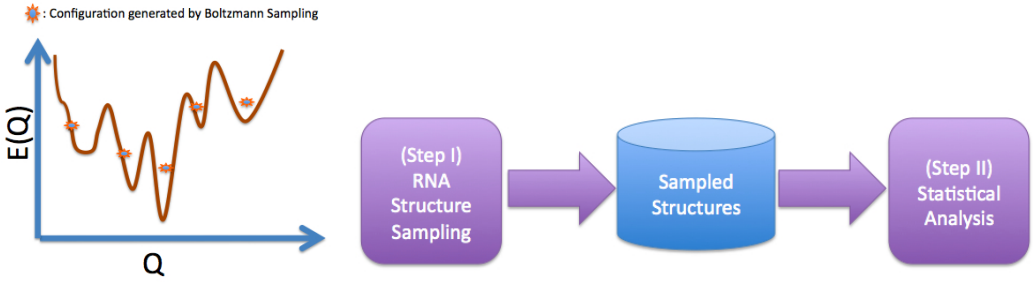
Figure 1: (a) The pipeline for riboswitch structure prediction and binding affinity estimation (b) Schematic of the secondary structure change displayed by the SAM-I riboswitch (The left figure represents the OFF state resulted by the SAM-binding and the right figure represents the ON state)

## 2.2 The Case for Continuity

Riboswitches are regulatory RNAs that control the expression of downstream genes. Small metabolite molecules, such as amino acids, nucleotides, coenzymes etc., can bind to riboswitches as effectors in vivo [2]. In our recent research efforts, the SAM-I riboswitch, one member of the riboswitch family that regulates genes related to the metabolism of sulfur and methionine, has been extensively investigated with atomistic simulations. This riboswitch choose alternative conformation depending on binding of a S-adenosyl methionine (SAM). When a SAM is bound, the aptamer domain forms anti-anti-terminator (AAT) conformation, which turns off the downstream genes by forming the terminator (T). Otherwise, the anti-terminator (AT) is formed prohibiting the T element formation for continuing transcription process (see Fig. 1(b) [3]). Although the structures of the SAM-I riboswitch in the anti-anti-terminator (AAT) conformation have been solved via X-ray crystallography, it is just a static view of how SAM binds to the SAM-I riboswitch RNA.

Our current research goals can be understood readily using the pipeline illustrated in Fig. 1 and the required computational tasks carried out with the allocation of this request are described below. With the proposed pipeline, we aim to investigate folding dynamics of ribsowtich RNAs and closely related RNA-ligand binding affinity. This is in contrast to our earlier strategy during which we have heavily relied only upon all-atom Molecular Dynamics simulations. Indeed, combining multiple computational approaches that differ in their physical principles, increase the chance to achieve a comprehensive understanding of the complex biological process carried out by riboswitch RNAs. The entire pipeline comprises three steps; the first step represents the Boltzmann Ensemble (BE) sampling of RNA secondary structures, the second step is about the 3D modeling from 2D structure information, and the third step carries out the conformational sampling. Note that sampling in the first step is executed using RNA secondary structure prediction algorithms, employing the nearest-neighbor energy model and thermodynamics parameters that are experimentally estimated. The third step is carried out with all-atom MD simulations. The second step represents 3-D molecular modeling using 2-D information. The requested allocation is primarily support the computational requirements corresponding to the first and the third stages. The BE sampling for the first step is conducted with SFold package [4]. This type of sampling is typically carried out without the abilitiy to concurrently run multiple simulations, but recently, thanks to the development of a novel runtime environment, we have demonstrated the “parallel sampling” of RNA secondary structures using scalable HPC resources (See ”Emerging Computational Methodologies in Life Science” [5]). For all-atom MD simulations, protocols similar to previous years will be used as described below. As a matter of fact, our pipeline represents the energy landscape perspective that interprets the folding dynamics of a RNA with statistical treatment of an ensemble of structures distributed in the pertinent energy landscape [6], which has been successfully applied for protein folding but not fully applied for RNA folding [7]. In our pipeline, the entire folding configuration space is efficiently explored by RNA secondary structure sampling and successively atomistic MD simulations explore the relevant basins of attraction starting from the configurations sampled.

While our pipeline can serve as a de-novo tool for RNA structure prediction from a sequence, it can also be used as a tool box that is able to carry out different physics-based calculations corresponding to each layer, i.e., RNA secondary structure prediction, 3D modeling, and all atom MD simulations for different scientific aims. This is related to the fact that the nature of RNA folding occurs in a hierarchical manner, implying that the first and the third steps can deal




(a) (b)

Figure 2: (a) Illustration of structure sampling in configuration space. (b) Schematic of a workflow for sampling and analysis of RNA secondary structures obtained with the Boltzmann-weighted sampling

with its own biophysical problems but later findings can be combined for an integrated perspective. Furthermore, our framework uses various computational approaches whose performance relies upon the availability of effective parallel implementation, thus benefiting from TeraGrid like federated, distributed and massively parallel resources.

A major proportion of our allocation for this project will be used for atomistic MD simulations. The main goal of the MD simulations is to probe the dynamic interactions between the SAM-I riboswitch and SAM at the nanoscale and to explore determinants for the specificity. In particular, we aim to examine, i) SAM-I riboswitches and their different constructs that differ from each other in potentially different secondary structures and tertiary interactions, ii) different sequences in SAM-I family, and iii) other SAM riboswiches (SAM-II and SAM-III) for which X-ray structures were recently reported and TPP ribosiwtches that we recently started to investigate. To estimate binding affinity, we employed the Molecular Mechanics - Poisson Boltzmann Surface Area (MM-PBSA) approach using configuration from MD simulations. As for efficient sampling of conformational dynamics, replica exchange molecular dynamics (REMD) protocol will be used. REMD calculations will be carried out using available scripts or with our recent development for the distributed adaptive REMD.

The protocols to be employed are similar to previous protocols. We will start with a structure derived from the X-ray crystal structures of the AAT conformation of SAM-I riboswitch (PDB: 2GIS, 3GX2, 3GX3, 3GX5, 3GX6, 3GX7) [8] or configurations generated by a process via two layers of the pipeline. For a free state riboswitch, SAM is directly removed from the x-ray crystal structure and replaced with solvent water. The amber99bsc0 correction force field is used here [9]. Parameters for SAM are from the Generalized Amber Force Field (GAFF) and missing parameters are calculated using ANTECHAMBER [10]. Positions of added hydrogens are guessed using PSFGEN within NAMD 2.6. Then the RNA molecules are solvated in a cubic solvent box of TIP3P waters with a 1.6 nm padding in all directions. Sodium and magnesium ions are distributed around the RNA molecules and neutralize charge of the system. The total number of atoms in the system is 56,000. Energy minimizations are carried out to remove bad contacts. Starting from 0 K, the temperature is raised 10 K every 10,000 steps, and is held constant after reaching the desired temperature (310 K) using temperature reassignment. MD simulations are performed in the NPT ensemble with the pressure maintained using the Langevin piston method with a period of 100 fs and decay times of 50 fs. The time step is 2fs for both equilibration and production phase. Bond lengths between hydrogens and heavy atoms are constrained using SHAKE. The long-range electrostatics is treated with the Particle Mesh Ewald (PME) method with a cutoff distance 1.2 nm. We use NAMD 2.6.

Obtained trajectories will be analyzed with various statistical techniques. Generally, beyond straightforward trajectory analysis measuring various structural variables, PCA analysis and clustering are useful to extract characteristic dynamical motions in terms of reduced dimension techniques [11, 12]. Also, we consider the calculations utilizing the Inherent Structure formalism that might reveal intriguing information with respect to the energy landscape properties such as basins, minima, and saddles [13, 14]. Conformational dynamics from obtained trajectories of a complex with a ligand and free state will be used for the MM-PBSA calculation with which ligand binding affinity of the metabolites, SAM, or TPP or other related ligands and also for cation binding of $Mg^{2+}$ is estimated. In particular, $Mg^{2+}$ binding has been evidenced as crucial for function of catalytic RNAs but remains as elusive for detailed roles in folding dynamics of riboswitches. Our preliminary results for cation binding are presented in Fig 3. According to our preliminary results, this protocol is useful for ranking ligands that bind a pertinent riboswitch, but also we observed that the accuracy depends on the way for entropy calculation and the use of configurations from MD trajectories. Also, more

**Energetic Analysis of the binding of a specific $Mg^{2+}$**

| Energy (kcal/mol) | MD_SAM | | MD_woSAM | |
|---|---|---|---|---|
| | MEAN | STD | MEAN | STD |
| ELE | -3696.76 | 49.51 | -3694.34 | 53.57 |
| VDW | 12.15 | 3.49 | 11.00 | 3.16 |
| GAS | -3684.61 | 49.11 | -3683.33 | 52.84 |
| PBSUR | -0.68 | 0.02 | -0.62 | 0.02 |
| PBCAL | 3600.15 | 48.86 | 3614.08 | 53.08 |
| PBSOL | 3599.46 | 48.86 | 3613.46 | 53.08 |
| PBELE(ELE+PBCAL) | -96.61 | 8.02 | -80.26 | 8.04 |
| **PBTOT(GAS+PBSOL)** | **-85.15** | **6.69** | **-69.87** | **6.09** |

Figure 3: MM-PBSA estimation on $Mg^{2+}$ binding with a SAM-I riboswitch RNA

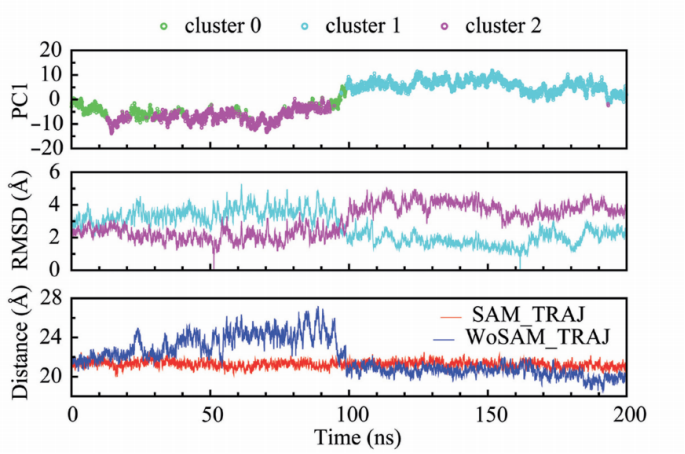



Figure 4: Figure reproduced from [12]. The root mean square distance and energetics fluctuate well into the 100-200 ns range. In future simulations, 200ns is a lower bound and we need to go to much longer: at least 300ns.

challenging questions relevant to the validity of force fields that ignores the polarizability as well as the charge transfer effect should be investigated and we hope our calculations provide interesting clues for the answer.

Sampling the Boltzmann ensemble of RNA secondary structures is the primary strategy to explore the energy landscape efficiently. We use the SFold pacakge [4] for sampling. The sampling of structures with an input sequence of riboswitches is not computationally intensive, but a considerable number of tasks are required depending on the biological questions. For example, we carry out comparative investigation on all riboswitches of the same family identified in RFam database; at the moment 2092 SAM-I riboswitches are identified, and the number grows as more genomes are added. Also, after a sampling step, the analyses involve multi-stage heterogeneous tasks (see Fig. 2(b)).

We plan to simulate SAM-I aptamer RNA. Previous time-step benchmarks on Ranger indicate a computation rate of 1ns/82 hours for this 56K atom system on 32 cores. In other words for a 56K system, 1 ns simulations require $\approx$ 300 CPU hours. Thus each 100 ns simulation requires approximately 30,000 CPU hrs. Analysis [12] results indicate that more than 300 ns trajectory is desirable for observing meaningful conformational dynamics. As figure 4 shows, the RMSD of a system starts to stabilize very late in time. Therefore, without additional post-analysis including MM-PBSA calculations, a rough estimation suggest that we can obtain about 100 simulations of a similar system with 900,000 SUs (See `http://staging.teragrid.org/userinfo/aus/namd_benchmark.php` Boltzmann Ensemble sampling. The total requested to finish this project is 1 million SUs on Ranger.

# 3 Continuing Project: Atomistic Simulations of Physiological Systems

The long term scientific objective of our project is to develop molecular dynamics simulations of medically relevant enzymes into a tool for clinicians to use in determining the cocktail of drugs to administer to an HIV-infected individual. This work is supported by grants under EU FP7 and FP6 via the VPH-NOE (EU FP7-ICT-2007-5.3 223920), Contra Cancrum (EU FP7-ICT-2007-5.3 223979), p-Medicine (EU FP7-ICT-2009-6 270089) and CHAIN (EU FP7 HEALTH-2007-2.3.2-7) projects. For such applications, reproducible accuracy at the level which can rank drug efficacies, and rapidity of acquisition of results (for clinical relevance) are all essential. This takes the application of bio-MD techniques into an entirely new domain. This project is divided into two distinct sub–projects: (i) Patient specific HIV therapy and (ii) Predicting the affinity of the EGFR kinase domain for drug inhibitors of lung cancer.

## 3.1 Subproject: Towards Patient Specific HIV Therapy

### 3.1.1 Subproject Progress

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 [15] 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 [16]. 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 [17].

### 3.1.2 The Case for Continuity

The aim of this project is to calculate binding affinities for HIV-1 enzymes with the anti-viral drugs used to target them in clinical practice. We have previously shown an excellent ability to reproduce experimental results for genetic variants of the HIV-1 protease binding the drug lopinavir using ensembles of 50 replica simulations of 4 ns trajectory duration [18]. Our work during the current grant has focussed on extending the use of the simulation and analysis protocol we have developed to the other FDA approved protease inhibitors (PIs -all of the inhibitors included in the study are listed in Table 1). The initial stage of this process is to access the protonation state of the catalytic dyad of the protease bound to each of the inhibitors, which has now been achieved for all drugs. As part of this stage of the project we developed a submission tool based on SAGA which allowed us to coordinate simulation runs on machines on both the XSEDE (Ranger) and the EU Distributed European Infrastructure for Supercomputing Applications (DEISA) [**?**]. The submission tool has been completed as part of our previous allocation: TG-MCB090174 and has been extended to include a gateway system: DARE [**?**] and infrastructure to support submission through Unicore [**?**], Genesis II [**?**], and Cloud systems [**?**]. ***YYE: Shantenu you might want to add stuff here .

Having completed our investigation of the protonation state used for each inhibitor we performed production binding affinity calculations for all of the drugs. Our results produced a promising inter-drug ranking but suggested that binding affinities are highly dependent on the initial conformation as well as protonation state of the catalytic dyad. It was also observed that results for the inhibitor ritonavir were acutely sensitive to the conformation of the aspartic acid at position 30. These observations provide the motivation for the future work we will describe in Section 3.1.3.

In addition to this progress on the originally proposed simulations, we also extended our previous work evaluating the binding affinity of HIV-1 protease mutants to the inhibitor lopinavir. As part of a previous collaboration in the EU Virolab project (EU FP7 223131) a comparative drug ranking methodology was used to compare drug resistance rankings produced by the Stanford HIVdb, ANRS and RegaDB clinical decision support systems. The methodology was used to identify a patient sequence for which the three rival online tools produced differing resistance rankings. This process identified mutations at only three positions (L10I, A71IV and L90M) which influenced the resistance level assigned to the sequence. We have simulated not only the full patient sequences but also systems containing the constituent mutations (a total of 12 sequence variants were simulated). Inserting any combination of the identified

| Inhibitor Code | Inhibitor Name |
|---|---|
| APV | amprenavir |
| IDV | indinavir |
| LPV | lopinavir |
| NFV | nelfinavir |
| RTV | ritonavir |
| SQV | saquinavir |
| AZV | atazanavir |
| TPV | tipranavir |

Table 1: Code and full names of the HIV-1 protease inhibitors (PIs) investigated.

mutations into the wildtype sequence produced no impact on the binding affinity of the protease for lopinavir. In contrast when the mutations were inserted into the background sequence present in the patient derived sequence resistance was observed. Our simulations also identified changes in the relative conformation of the two beta sheets that form the protease dimer interface which suggest an explanation of the relative frequency of different amino acids observed in patients at residue 71. This study has been submitted for publication [19].

Simulations of the HIV-1 reverse transcriptase bound to the inhibitor efavirenz (EFZ) have also been performed. The use of an ensemble approach has revealed that previous single trajectory results which allowed the discrimination of wildtype, K103N and L100I/K103N were fortuitous. Consequently we performed simulations of only the wildtype, K103N, L100I and L100I/K103N sequences rather than the more extensive range of variants proposed. Our results suggest that we need to adapt our protocol here to both include more replicas and perhaps to simulate the apo enzyme as well as the drug bound form. Simulating the apo form should allow us to evaluate the energetic cost of binding pocket formation for each sequence. Simulation conducted as part of this study have shown that experimental differences in binding affinity between EFZ and another drug (nevirapine, NVP) can consistently be reproduced.

The resource usage of this project to date is shown in Table 2. The three studies described above are listed as the PR Multiple Drug Resistance Study, PR Virtual Patient Simulations and RT Drug Resistance Study respectively.

| Sim Description | No. Sims | No. Cores | Code | TG machine | Total SUs |
|---|---|---|---|---|---|
| **PR Multiple Drug Resistance Study** | | | | | |
| PIs - 6 wildtype PR systems | 380 | 64/48 | NAMD | Ranger/Kraken | 437,760 |
| **PR Virtual Patient Simulations** | | | | | |
| LPV - 6 PR sequences | 600 | 64/48 | NAMD | Ranger/Kraken | 345,600 |
| **RT Drug Resistance Study** | | | | | |
| EFZ - 3 RT sequences | 30 | 192 | NAMD | Ranger | 600,000 |
| Grand total of SUs used | | | | | 1,383,360 |

Table 2: Simulations performed and associated computational cost.

### 3.1.3 Ongoing and Future Work

The previous work in this study has suggested that our free energy protocol is sensitive to particular conformations of the catalytic dyad in monoprotonated systems. The two key conformations that can be adopted were named ‘up’ and ‘down’ by Zhang & Zhang [20] and are illustrated in Figure 5. In order to obtain a correct binding affinity we must start simulations in both configurations. We have found that seven of the FDA approved inhibitors have monoprotonated dyads and we now wish to run ensembles intialised in the opposite configurations to those already sampled. This should allow us to more accurately rank protease sequences. Once we have refined out simulation set up we plan to study each drug bound to a series of mutant sequences with 5 experimentally determined levels of resistance. For the protonation studies we propose to use 20 replica ensembles, for the cross drug comparison we will use 50 replica ensembles. In all cases individual replicas will consist of 2 ns equilibration and 4 ns of production simulation.

Our reverse transcriptase study indicates the need for simulations of the apo protein we propose to simulate 10 replica ensembles with 4 ns of production run for the wild type and L100I/K103N double mutant in order to replicate the sampling we have achieved for the drug bound systems.

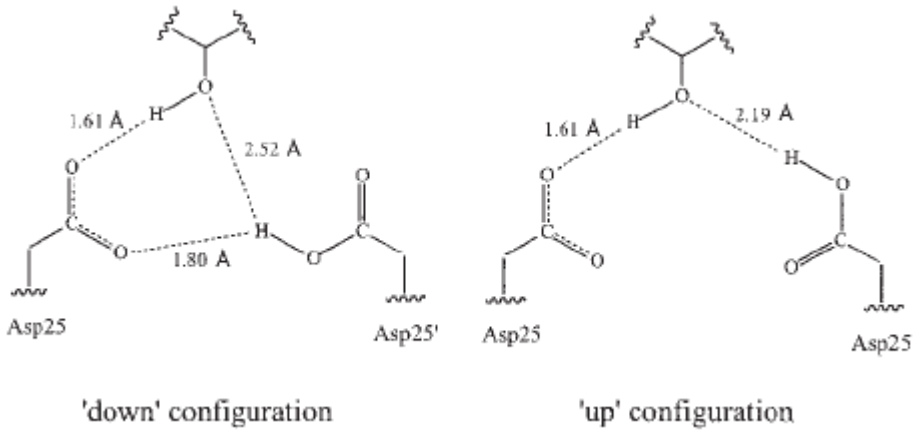



Figure 5: Two conformations of the monoprotonated catalytic aspartic acid dyad of HIV are possible, they are labelled 'down' and 'up', respectively. The difference between the two lies in the positioning, relative to the ligand of the hydrogen atom. In the 'down' position the hydrogen atom is involved in a hydrogen both with the unprotonated Asp25. In the 'up' conformation it is involved in a bond with the ligand.

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 [21]. 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.

Our proposed resource requirements are shown in Table 3.

| Sim Description | No. Sims | No. Cores | Disk | Code | TG machine | Total SUs |
|---|---|---|---|---|---|---|
| **PR Catalytic Asp Study** | | | | | | |
| 7 PIs - WT PR | 240 | 64 | 250GB | NAMD | Ranger | 276,480 |
| **PR Multiple Drug Resistance Study** | | | | | | |
| 7 PIs - 5 MDR systems | 1750 | 64/48 | 250GB | NAMD | Ranger/Kraken | 2,016,000 |
| **RT Drug Resistance Study** | | | | | | |
| Apo RT Wildtype | 10 | 192 | 300GB | NAMD | Kraken | 768,000 |
| Apo RT L100I/K103N | 10 | 192 | 300GB | NAMD | Kraken | 768,000 |
| Grand total of SUs required | | | | | | 3,828,480 |

Table 3: Planned simulations and associated computational requirements.

## 3.2 Subproject: Predicting the affinity of the EGFR kinase domain for drug inhibitors of lung cancer

### 3.2.1 Subproject Progress

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 [22–24], the other to

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

### 3.2.2 The Case for Continuity

We have performed relative binding affinity calculations using multiple (ensemble) short MD simulations. Simulations have been run for two tyrosine kinase inhibitors AEE788 and Gefitinib complexed with wild-type and 4 mutant EGFRs. 50 replicas were used for each molecular systems to ensure the calculated properties are reproducible [22]. In principle, all replicas within a single ensemble simulation can easily be run concurrently in one day, thanks to the vast number of cores on the XSEDE supercomputers (Ranger for this work). This makes it possible to accurately rank drug binding affinities on clinically relevant timescales. We show that ensemble simulations correctly rank the binding affinities for these systems: we report the successful ranking of each drug binding to a variety of EGFR sequences and of the two drugs binding to a given sequence. The study was published recently in J. R. Soc. Interface [22].

Long timescale simulations have been performed to study the mechanism of activation by cancer-causing mutations within EGFR. 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 that affect its activation and drug efficacy. The most common EGFR mutation in lung cancer patients is a leucine to arginine substitution at amino acid 858 (L858R). To investigate the changes of conformational equilibrium between the active and inactive states, 4 replicas of EGFR were used for each conformation, and 200ns molecular dynamics simulations were performed for each replica. Using XSEDE resources, as well as the EU Distributed European Infrastructure for Supercomputing Applications (DEISA) allocation, we have performed longer simulations than we initially proposed (100ns each). Structural and thermodynamic properties have been extracted from these simulations. The thermodynamic stabilities of these two conformations are characterized by free energy landscapes estimated from molecular mechanics/PoissonBoltzmann solvent area calculations. Our study reveals that the L856R mutation introduces conformational changes in both states, adjusting the relative stabilities of active and inactive conformations and hence the activation of the EGFR kinase [23].

The epidermal growth factor receptor (EGFR) is an especially important enzyme target in lung cancer therapy because it mutates and/or is overexpressed in most non-small cell lung carcinoma (NSCLC) tumours. Inhibition of kinase activation of EGFR is a frequently used method to suppress its functions [25]. The majority of tyrosine kinase inhibitors (TKIs) are ATP-competitive inhibitors which bind in the ATP-binding site. Molecular dynamics (MD) simulations will be used to study the structural and energetic properties of inhibitor-EGFR complexes. The binding affinity of inhibitors to EGFRs will be calculated by molecular mechanics Poisson-Boltzmann surface area (MM/PBSA) methods [26]. This molecular level study is one component of the EU FP7 ContraCancrum (Clinically Oriented Translational Cancer Multilevel Modeling) project which aims at developing a composite multilevel platform for simulating malignant tumor development and pharmacologic responses to a therapeutic intervention (http://www.contracancrum.eu). We have employed large scale MD techniques using both TeraGrid and DEISA resources in order to study the interactions of inhibitors with wild-type and mutant EGFRs [27].

### 3.2.3 Future Work

A better understanding of the reasons for the success or failure of a therapeutic intervention will help us in the selection of subgroups of patients who are most likely to respond to specific drugs, and paves the way for personalized treatment [28]. We have already performed a preliminary study of different inhibitors (AEE788, AFN941 and gefitinib) with EGFR which we now intend to extend to look at a wider variety of inhibitors and EGFR mutations and to probe longer time scale motions of the protein. Planned simulations include ensembles of 50, 50,000 atoms with 25 runs each. Each simulation lasts for 4ns and runs for 9 hours on 128 cores. We therefore request 1.5 million SUs on Kraken to finish this project. The ensemble size and number of runs are typical of established studies [26, 27] of NSF funded research.

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.

***YYE: Insert Table, ask SJ and Coveney for numbers and application

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

#### 4.0.4 Subproject Progress: 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 [29]. 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 [30]. 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.

## 4.1 Subproject Progress: 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

## 4.2 Ongoing and Future Work

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 [31].

***YYE: Submit Table if we need more resrouces

# 5 Continuing Project: Community Code Development

In conjunction with the scientific questions we are addressing, we are also involved in the development of a runtime 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 XSEDE 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 the community code infrastructure and assess performance using real scientific workloads, and make it available for the broader & larger community of biomolecular simulators.

To continue infrastructure development, a small allocation on a number of machines is required. We would like to include as wide an array of different machines as possible to ensure that the infrastructure is portable and easy to deploy. This is particularly important since a large number of the user community (e.g. TG-MCB100145) use different resources. For this reason, we are requesting 100K SUs on Lonestar, Blacklight and Trestles 4.

| Resource | Requested SUs |
|---|---|
| Lonestar | 100K |
| Blackligh | 100K |
| Trestles | 100K |

Table 4: Resources requested for continued community code development

# 6 Supporting Grants

The PI's research is supported by CHE-1125332 ,"CDI-Type II: Mapping Complex Biomolecular Reactions with Large Scale Replica ExchangeSimulations on National Production Cyberinfrastructure", for $1.65M for 4 years. The PI's research is also supported by current NIH and NSF awards. See PI's vitae for full grant listing. The PI leads Work Package 4 of the NSF Funded Cybertools Project (http://www.cybertools.org) (NSF Award NSF/LEQSF(2007-10)-CyberRII-01; Total Value $12M) and the $15M NIH award supporting the Louisiana Biomedical Research Network (LBRN). Project 1 is partially supported by the Biosensors work activity of WP-1 of Cybertools.

Project 1 is funded by multiple awards, including a $15M NIH Louisiana Biomedical Research Network Award (of which Jha is the co-PI and Director of the Bioinformatics and Computational Biology Core; award number 2 P20 RR016456-09), Louisiana Board of Regents award and an LSU Faculty Award (PI Jha), in conjunction with multiple awards to Fareed Aboul-ela (Experimental Collaborator). Project 2 has also benefitted from a LONI Distinguished Graduate Assistantship to Wei Huang (PhD Student co-supervised by PI Jha and experimental collaborator Aboul-ela). Project 2 forms the basis for a large NSF proposal under review, *IIS 1029810 Macromolecular Choreography: Computational methods to detect conserved dynamic properties in non-coding RNAs*.

Project 3 is led by co-PI Coveney, who is the Principal Investigator of the $12M EU FP7 Virtual Physiological Human Network of Excellence. Project 1 & 4 are also supported by RealityGrid grant GR/R67699 funded by EPSRC (the UK equivalent of NSF). Coveney and Jha have active collaboration over eight years in the field of molecular dynamics, high performance, distributed & grid computing and have co-authored 12 papers in these areas, and are currently writing several papers related to Project 1, 3 and 5c.

PI-Jha is the co-PI of LSU's HPCOPS NSF-OCI 0710874 award which also supports a large fraction of Projects 5a and 5c. Integration of SAGA with applications is part of Cybertools and the PI also holds multiple peer-reviewed awards for the development and integration of SAGA (EPSRC GR/D0766171/1) The Interoperability Project [32] is currently funded by an NSF HPCOPS award, and is being executed by the PI (Jha).

In addition, PI Jha is the LSU-lead in the ExTENCI (OCI-1007115) project that aims to further interoperability between TeraGrid and Open Science Grid. Project 5b is supported by NSF OCI award - ExTENCI: Extending Science Through Enhanced National Cyberinfrastructure (total $2M, LSU share $0.2M).

### Resource Request Summary

The total system time requested for various XSEDE resources is summarised in table 5

| Resource | Total SUs Requested |
|---|---|
| Kraken | 5 million SUs |
| Ranger | 2 million SSs |
| Lonestar | 0.3 million SUs |
| Blacklight | 0.1 million SUs |
| Trestles | 0.1 million SUs |

Table 5: Total Resource Request

## 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] M. Mandal et al. Nat Rev Mol Cell Biol. 5(6), 451-63, 2004.

[3] B. A. M. McDaniel et al. Proc. Natl. Acad. Sci. U.S.A. 100, 3083-3088, 2003.

[4] Y. Ding, "Statistical and Bayesian approaches to RNA secondary structure prediction," *RNA*, vol. 12, no. 3, pp. 323–331, 2006.

[5] J. Kim, W. Huang, S. Maddineni, F. Aboul-ela, and S. Jha, "Exploring the RNA folding energy landscape using scalable distributed cyberinfrastructure," *ACM HPDC 2010, Emerging Computational Methods for the Life Sciences*, 2010.

[6] J. N. Onuchic and P. G. Wolynes, "Theory of protein folding: The energy landscape perspective," *Annu. Rev. Phys. Chem.*, vol. 48, pp. 543–600, 1997.

[7] C. J., C. Flamm, A. Renner, and P. F. Stadler, "Density of states, metastable states, and saddle points exploring the energy landscape of an RNA molecule," *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, vol. 5, pp. 88–91, 1997.

[8] R. K. Montange et al. Nature. 255, 1172-1175, 2006.

[9] P. Alberto, et al. Biophysical Journal. 92, 3817-3829, 2007.

[10] J. Wang et al. Journal of Molecular Graphics and Modelling , 25, 247260, 2006,.

[11] J. Kim and T. A. Keiderling, "All-atom Molecular Dynamics Simulations of b-hairpins Stabilized by a Tight Turn: Pronounced Heterogeneous folding pathways," *accepted in J. Phys. Chem. B*, 2010.

[12] W. Huang, J. Kim, S. Jha, and F. Aboul-ela, "A mechanism for s-adenosyl methionine assisted formation of a riboswitch conformation: A small molecule with a strong arm," *Nucleic Acid Res.*, vol. 37, no. 19, pp. 6528–6539, 2009.

[13] T. Keyes, J. Chowdhary, and J. Kim, "Random Energy Model for dynamics in supercooled liquids: N dependence," *Phys. Rev. E*, vol. 66, p. 051110, 2002.

[14] J. Kim and T. Keyes, "On the mechanism of reorientational and structural relaxation in supercooled liquids: The role of border dynamics and cooperativity," *J. Chem. Phys.*, vol. 121, p. 4237, 2004.

[15] 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.

[16] 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.

[17] 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.

[18] S. Sadiq, D. Wright, O. Kenway, and P. Coveney, "Accurate Ensemble Molecular Dynamics Binding Free Energy Ranking of Multi-Drug-Resistant HIV-1 Proteases," *J Chem Inf Mod*, vol. 50, no. 5, pp. 890–905, 2010.

[19] H. H. A. M. G. P. J. W. P. C. Owain Kenway, David W. Wright and S. Jha, "Towards high-throughput, high-performance computational estimation of binding affinities for patient specific hiv-1 protease sequences," *TeraGrid 2011*, 2011.

[20] Y. Zhang and D. Oliver, "History matching using a hierarchical stochastic model with the ensemble Kalman filter: a field case study," *SPE Reservoir Simulation Symposium*, 2009.

[21] 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.

[22] 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.

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

[24] 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.

[25] P. J. Zhang and N. Gray, "Targeting cancer with small molecule kinase inhibitors." *Nat Rev Cancer*, vol. 9, pp. 28–39, 2009.

[26] S. Wan, P. Coveney, and D. Flower., "Peptide recognition by the T cell receptor: comparison of binding free energies from thermodynamic integration, Poisson-Boltzmann and linear interaction energy approximations." *Phil Trans R Soc A*, vol. 363, no. 1833, pp. 2037–2053, 2005.

[27] S. Wan and P. Coveney, "Patient specific prediction of drug binding affinities." in *The World Congress on Medical Physics and Biomedical Engineering, Munich, Germany*, 2009.

[28] P. Sloot, P. Coveney, G. Ertaylan, V. Müller, C. Boucher, and M. Bubak., "HIV decision support: from molecule to man." *Phil Trans R Soc A*, vol. 367, no. 1898, pp. 2691–2703, 2009.

[29] 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.

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

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

[32] TeraGrid-LONI Interoperabilty Project, *A Science-Driven Project Using Advanced CyberInfrastructure funded by NSF via a HPCOPS award to LONI*, http://www.teragridforum.org/mediawiki/index.php?title=LONI-TeraGrid-DEISA_Interoperabilty_Project.