Exploring the RNA Folding Energy Landscape Using Scalable Distributed Cyberinfrastructure

Joohyun Kim
Center for Computation and Technology
Louisiana State University
Baton Rouge, LA

Wei Huang
Biological Sciences
Louisiana State University
401 Choppin
Baton Rouge, LA

Sharath Maddineni
Center for Computation and Technology
Louisiana State University
Baton Rouge, LA

Fareed Aboul-ela
Biological Sciences
Louisiana State University
Baton Rouge, LA

Shantenu Jha
CCT & Comp. Science
Louisiana State Unviersiy
Baton Rouge, LA

ABSTRACT

The increasing significance of RNAs in transcriptional or post-transcriptional gene regulation processes has generated considerable interest towards the prediction of RNA folding and its sensitivity to environmental factors. We use Boltzmann-weighted sampling to generate RNA secondary structures, which are used to characterize the energy landscape, via the distributions of energies and base-pair distances. Depending upon the length of an RNA, the number of sequences investigated, and the sample size of generated structures — generating and analyzing sufficient samples can be computationally challenging. We introduce and develop a lightweight and extensible runtime environment that is effective across a range of RNA sizes and other parameters, as well as over a range of infrastructure – from traditional HPC grids to clouds, without requiring any changes at the application or user level. The Adaptive Distributed Application Management System (ADAMS) is built upon an extensbile and interoperable pilot-job and supports the concurrent execution of a broad range of task sizes across a range of infrastructure. We use ADAMS to investigate the folding energy landscape for two RNA systems of different sizes: a set of S-adenosyl methionine (SAM) binding RNA sequences known as SAM-I riboswitches and the S gene of the Bovine Corona Virus (BCoV) RNA genome that comprises 4092 nucleotides. Results of the energy and base-pair distance distributions suggest different energy landscapes, implying different folding dynamics. With obtained results, we demonstrated the possibility of utilizing this protocol to explore microscopic origins for reported sequence-dependent variation of binding affinity and gene expression in the two RNA systems.

Categories and Subject Descriptors

D.1.3 [**Software**]: Concurrent Programmingdistributed programming/parallel programming; J.3 [**Computer Applications**]: Biology and genetics

General Terms

Theory, Cyberinfrastructure, Simulations

Keywords

RNA folding energy landscape, SAM-I riboswitch, S gene of Bovine Corona Virus, Runtime Environment, Distributed Computing, Simple API for Grid Applications, SAGA, Pilot-Job abstraction

1. INTRODUCTION

RNAs are critically involved in many biological processes in living cells[54, 32, 18, 4]. Understanding how an RNA functions is directly connected to the understanding of its folding as well as interactions with other molecules[55, 8, 27, 17]. In some cases, the well defined 3-D structures determined by X-ray diffraction provide great insight regarding the biological function and mechanisms[47, 16, 50, 53]. However, there is mounting evidence that the complex mechanisms of RNAs are better understood when structural information on various alternative conformers or states as well as on dynamical transitions among them are considered[47, 53, 30, 56].

In recent years, there has been an explosion of reports on newly identified structured RNAs including many non-coding (nc) RNAs. These RNAs are found to play significant roles in various gene regulation mechanisms[32, 54, 18, 9, 4]. The functions of these RNAs are often associated with complicated structure formation, for example through the folding process occurring in response to binding a metabolite or complex formation with other proteins or nucleotides. These findings underscore the importance of multiple folding pathways exhibited by an ensemble of structures [5, 10, 19, 30, 57, 39]. In such cases, the energy landscape perspective, in which the statistical description of an ensemble of structures is the main idea, may provide physical insights for RNA folding or complex formation[48, 63, 31, 13, 57, 11].

A good example of a class of the nc-RNAs that exhibit com-

plex structural organization is the RNA riboswitch[28, 46, 51, 24, 10]. This RNA system is found in the untranslated region up-stream of mRNA of a target gene, mostly in bacterial systems. It adopts alternative folding structures in response to pertinent metabolite binding. The consequence of this structural organization eventually triggers the on or off state in a downstream region, called the expression domain, resulting in regulation of target gene transcription or translation. A number of studies have been devoted to understanding riboswitch mechanisms and to various applications including drug discovery[36, 7]. In previous work[29], we investigated the SAM-I riboswitch RNA with atomistic molecular dynamics simulations for its SAM-bound folded state. From these simulations we proposed hypotheses for how SAM binding affects the stability of the folded state as well as for the folding mechanism. Interestingly, Henkin and coworkers observed variation of gene regulation efficiency as well as binding affinity among a series of SAM-I riboswitch sequences[58]. An intriguing question raised by their experimental data is whether the variability in binding affinity is correlated in differences of the folding energy landscapes of different sequences. The simple rationale behind this question is that the forward binding rate could be affected by the distribution of an ensemble of structures. The population of certain secondary structures that better accommodate metabolite-binding would affect the forward binding rate. In a slightly different context, a recent study demonstrated the importance of multiple pathways in RNA hairpin formation, showing changes in relative contributions with such pathways as different experimental probes are used[30].

On the other hand, folding of an entire RNA genome or a part of the entire genome is an important subject for certain biological studies[62, 61]. One example is the recent experimental observation from gene expression studies of Bovine Corona Virus (BCoV) genome[15]. Kousoulas and coworkers found that the codon-optimized sequence of the S gene that comprises 4092 nucleotides (nts) successfully expressed the pertinent Spike glycoprotein but the original RNA sequence did not allow gene expression under the same experimental condition[15, 14]. One immediate question, therefore, is whether the structure formation with the original sequence differs from that predicted for the codon-optimized sequence. This question arises due to the possibility that locally formed secondary structures could affect translation processes, and we pursue the answer by comparing the energy landscapes.

Motivated by these biological questions in the context of two RNA systems, the SAM-I riboswitch and the BCoV S gene, we carried out comparative studies for related sequences with the aim of characterizing differences in ruggedness of energy landscapes. The energy landscape point of view has been successfully applied to explain protein folding or misfolding, and has used the Random Energy Model (REM) [48, 63], which interprets the energy distribution or parameters representing the correlated nature of the energy landscape as indicators of the energy landscape ruggedness[33]. With RNAs, we propose to use base-pair distance as a measure of the correlated energy landscape that has been used for a measure of similarity between two secondary structures predicted from a sequence. Our strategy to explore the energy

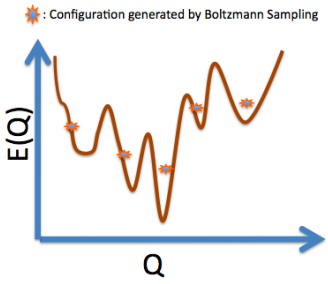

Figure 1: Illustration of the Energy Landscape perspective for folding/misfolding of proteins or RNAs. Sampled configurations are shown with the symbol. Q represents the representative coordinates of configurations and E(Q) represents the potential energy or the free energy.

landscape consists of sampling secondary structures and the subsequent calculation of the distribution of energies as well as base-pair distances. The first step is illustrated schematically in Fig. 1.

In this study, we consider structure information in terms of secondary structures[43, 23, 40, 66, 44, 26]. Along with others, Zuker and Stiegler[65] suggested an efficient way to predict a minimum free energy (MFE) structure utilizing dynamic programming. Subsequently, Zuker, Turner and their coworkers[42, 41] reported experimentally determined thermodynamic parameters based on the nearest-neighbor energy model and used them for the MFE calculation (see the ref. [43] and references therein). On the other hand, McCaskill[44] introduced a dynamic programming algorithm for the calculation of the partition function. Note that, unlike the MFE approach, the partition function approach does not predict a single or a set of secondary structure, but the partition function is a primary quantity for thermodynamic properties that can be measured from experiments and theoretical calculations. At this moment, RNA folding predictions using the two approaches are available through various packages[66, 21, 26, 43]. Notably, Ding and coworkers introduced Sfold recently and the package proposed a means of Boltzmann-weighted sampling utilizing the partition function calculation[21].

In this work, we employed Sfold for sampling of a Boltzmann-weighted ensemble of secondary structures. Furthermore, we proceed to utilize the sampled structures to characterize the pertinent energy landscape. The energy landscape perspective is a theoretical departure from the approaches relying upon a MFE structure or a small set of representative structures around the MFE structure, providing a more rigorous description of a dynamic, polymorphic system than MFE[31].

Many challenges remain in the computational investigation of RNA folding dynamics. The challenges can be divided in two categories: the first is associated with the computational cost resulting from the complexity of required computation, and the other is related to the effective and efficient execution of the resulting scientific computation *for a range of input sizes*. To address these challenges, we implement the ability to execute concurrent tasks for sampling and anal-

ysis phases. We employ an efficient runtime environment ADAMS — a lightweight application management system that has been developed for supporting dynamic execution of a set of tasks comprising a scientific workflow. ADAMS can support a range of execution modes, task sizes and types — including high-throughput of highly-parallel tasks, many task computing (MTC) and simple parameter sweeps. It is important to establish that our approach enables *all* of these task types to seamlessly utilize resources irrespective of whether they are logically or physically distributed[37].

We have previously reported on the development of a runtime environment for DARE tasks for sampling of 3-D structures using atomistic Molecular Dynamics simulations[37]. ADAMS is built upon Simple API for Grid Applications (SAGA) and its pilot-job abstraction SAGA-BigJob[52]. Although distributed federated HPC grids have been traditionally the primary computing resources employed, the underlying technology has been employed on a range of cloud systems [38] Therefore, from a cyber-infrastructure perspective this work represents an important extension by demonstrating the capability of the SAGA-BigJob to support heterogenous & ordered phases of a scientific problem. However, the critical contribution of this work is the application of advances in cyberinfrastructure to make progress towards the general purpose solution of well-defined biological problem for a wide-range of input sizes. We note that two RNA systems, the SAM-I riboswitch and S gene of BCoV, difer in size, the number of sequences of interest and show different biological complexity and computational costs and thus widely-different time-to-solutions.

This paper is organized as follows. In Section 2, we describe methods including information about RNA sequences that we investigate here and the two measures for characterizing the energy landscape ruggedness. Section 3 describes computational challenges with respect to the computational complexity associated with exploring RNA folding energy landscape, strategies for concurrent execution of many tasks, and the overall structure of ADAMS. In section 4, we present results establishing the effectiveness of our approach and analyzing scientific results obtained as a consequence. Discussions and concluding remarks are presented in section 5 and section 6, respectively.

2. METHODS

2.1 RNA Sequences

The two RNA systems investigated in this work are summarized in Table 1. Eight SAM-I riboswitch sequences are chosen from the work recently reported by Henkin and coworkers[58]. To match the experimental set-up for binding affinities, all 8 sequences of SAM-I riboswitches are constructed to contain the region from the Anti-Anti-Terminator (AAT) element (which binds a metabolite, S-adenosyl methionine (SAM)) but its 3'-end stops just before the Termination (T) element. Consequently, the entire expression domain is not included, and thus the termination efficiency is not investigated here. Note that in spite of variations with each sequence, all sequences function intrinsically as a SAM-I riboswitch but exhibit the natural variability in binding constant.

The other RNA system is the S gene region of Bovine Coronavirus Genome. Two different sequences are investigated in this work motivated by gene expression experiments conducted by Kousoulas group[15, 14]. This S gene has been investigated due to it significance for understanding viral infection mechanisms as well as a vaccine target. Interestingly, their experiments found that while the original sequenced RNA did not express the target S glycoprotein, the codon-optimized sequence using DNA work program expresses the protein.

Table 1: RNA systems. For SAM-I riboswitches, the downstream gene is also indicated[58].

RNA	Seq. ID	Length (nts)	Description
SAM-I	I-A	161	metE
	I-B	137	yitJ
	I-C	132	yjcI
	I-D	179	ykrT
	I-E	117	ykrW
	I-F	144	yoaD
	I-G	124	yusC
	I-H	122	yxjG
BCoV S gene	II-A	4092	codon-optimized
	II-B	4092	original sequence

2.2 Statistical Sampling of the RNA Secondary Structures: Using SFold for Boltzmann Weighted Sampling

The probability of a secondary structure, I_i, of a sequence, $S = r_1 r_2 ... r_N$ (r_α = A, U, G, C, $\alpha = 1, ..., N$), in the Boltzmann Ensemble is given as:

$$P(I_i) = \frac{e^{-\beta E(I_i)}}{Z(N,T)}. \quad (1)$$

$\beta = \frac{1}{k_B T}$, k_B is Boltzmann's constant, and $E(I_i)$ is the energy of a secondary structure I_i. $Z(N,T)$ is the partition function defined as

$$Z(N,T) = \sum_i e^{-\beta E(I_i)}. \quad (2)$$

The general idea for the dynamic programming used for MFE RNA secondary structure and a simple example for its connection to an energy calculation can be found in the reference [23]. Generally, the partition function can be calculated if all secondary structures are known, but obtaining all possible secondary structures is a formidable task even with a moderate size of a sequence[31]. However, thanks to McCaskill's algorithm[44], the partition function, in case of the simple model of Turner and Zuker that is employed in this work[43], can be calculated efficiently through a dynamic programming algorithm similar to one used for MFE structure prediction.

Exploiting the idea of McCaskill's partition function algorithm, Ding and Lawrence proposed the Bolzmann Ensemble sampling scheme[20]. The algorithm is available in the Srna module of the Sfold package[21]. In the module, the first stage calculates the partition function, and then the second stage samples secondary structures by calculating the

probability of a new sampled structure with the value of the partition function (see Eqn. 1). The detailed scheme on generating trial structures can be found in Ref. [20]. All sampling results presented in this work are obtained at 37 °C. Taken together, we construct the energy landscape by utilizing a Boltzmann-weighted sampling at a given temperature.

Figure 2: Schematic for a workflow comprising steps for sampling and analysis. During the sampling step, Step I, a set of secondary structures calculations generates a sampling of the energy landscape (sampled structures). Calculation of statistical measures with base-pair distance and energy take place subsequently during Step II.

2.3 Energy Landscape Ruggedness: Energy Distribution and Base-pair Distance Distribution

Once sampling of structures is carried out, various statistical analyses are applied on the sampled structures (see Fig. 2). Results of these statistical analyses, are used for verifying quality of sampling, but more importantly are used for characterizing the energy landscape. In this work, we focus on two statistical measures – the distribution of energies and base-pair distances, which are computed using in-house python-based scripts. The definition of the base-pair distance from I_A to I_B, used in this work, is the number of base pairs that exists in I_A, but not in I_B. The computational cost of the former (calculation of the distribution of energies) is low given the output of the sampling process. A calculation of the distribution of base-pair distances is relatively costlier, as discussed below.

3. DESIGNING AND DEVELOPING CYBERINFRASTRUCTURE TO EXPLORE RNA ENERGY LANDSCAPE

3.1 Analysing the Computational Complexity of Exploring the RNA Energy Landscape

We examine the computational cost dependence on the various parameters in order to understand the computational challenges. The parameters that characterize the computational complexity are: (i) the number of RNA sequences (M), (ii) the number of nucleotides in each RNA sequence (N_a where $a = 1, ..., M$), and (iii) the number of structures to be sampled Ω (for each sequence).

The first step of sampling of secondary structures, termed Step I, is carried out with an external standalone package (Sfold). Sfold has two internal stages for sampling that differ in parameter dependency and can not separately executed, limiting the achievable scaling with the parallel strategy that

Table 2: Statistics of the number of base pairs among formed in each structure. The mean (m) and the standard deviation (σ) are summarized.

Seq. ID	Num. of sampled structures	m	σ
I-A	1000	45.8	2.8
I-B	1000	36.2	1.9
I-C	1000	40.7	1.5
I-D	1000	56.0	2.5
I-E	1000	31.7	2.6
I-F	1000	42.1	2.1
I-G	1000	35.6	1.3
I-H	1000	35.5	2.3
II-A	1000	1168.2	12.8
	10000	1167.8	12.5
II-B	1000	1306.3	14.9
	10000	1306.0	15.1

focuses on the parallel execution of the second stage. We will analyze these aspects later in the results section. The time for executing Step I (T_E^I) is dependent upon these parameters as follows:

$$T_E^I = \sum_{a=1}^{M} \{f_a {N_a}^3 + g_a \Omega \times N_a\} \quad (3)$$

where f_a and g_a are constants that represent times for relevant unit tasks, and the summation is over the different RNA sequences. Here, we assume that i) McCaskill's algorithm for the partition function requires $O({N_a}^3)$, and ii) the sampling step after calculating the partition function requires $O(\Omega) \times O(N_a)$. Note that additional parameters,$\Theta_{I_i} (i = 1, 2, ...\Omega)$ that represent the numbers of base pairs formed in each secondary structure, should be considered, but we ignore them for simplicity as these are complicated variables dependent on N as well as a structure I_i $(i = 1, 2, ..., \Omega)$ and unknown until the secondary structures are sampled from Step I. Insight into the contribution regarding Θ can be obtained from data shown in Table 2: statistics for the number of base pairs formed in each sampled structure are presented. First of all, while among SAM-I riboswitches, no significant pattern is indicated, two S gene sequences (II-A, II-B) differ in the mean values. That is because the codon-optimization changes the base pairing due to different codons are used. In fact, such change is reflected in the energy distributions (Fig. 9) showing the overall shift of the distribution.

The computational cost of Step II (T_E^{II}) — the analysis phase, is both more complicated and greater. The base-pair distance calculation requires $O(\Omega^2) \times O(\Theta^2)$ — where Θ represents the number of base-pair in the structure, $I_{a,i}$. The calculation of base-pair distances for Ω sampled structures is composed of four loops; the outer two loops iterate over all structures, i.e., taking $\Omega \times (\Omega - 1)$ iterations, and two inner iterations are applied for base-pairs found in two different structures.

As show in Fig. 3 and Fig. 4, a larger sample size decreases the statistical noise, thus establishing a dependence on Ω. Lower energy states are critically important in determin-

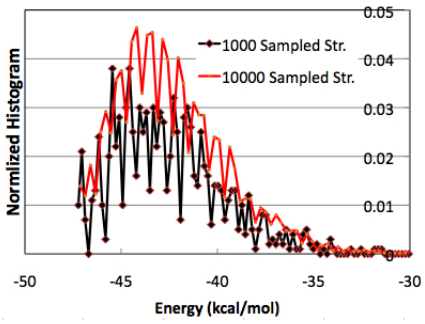

Figure 3: Histograms for energy distributions that depend on the number of sampled structure. Two histograms using different number of sample sizes; Ω with Seq. I-A are compared.

ing the final stages of folding dynamics[48, 63]. These low-energy states (left-side on the x-axis) are difficult to sample sufficiently, i.e., are sensitivie to noise and thus to lower sample counts. Thus greater sampling of structures is required for accurate statistics.

It is important to note that increasing Ω poses a challenge, as in some cases there is a unpredictable yet very dramatic increase in the computational cost of the subsequent steps of analysis. For example, in our case, the base-pair distance calculation takes a significantly longer time as Ω increases; this is shown in Table 5. The unpredictable yet large variation in execution time needs to be addressed by making the execution environment adaptieve.

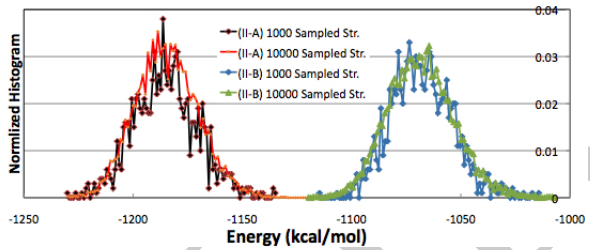

Figure 4: Quality of obtained histograms for energy distributions that depends on the number of sampled structure. Two histograms using different number of sampling size, Ω for Seq. II-A and II-B are respectively compared.

As for M, a simple reason to account for large M is because of ever-growing genomic data. For example, approximately a 1000 SAM-I riboswitches were identified with the Rfam database [1], and it is not surprising to see the number continuing to grow quickly. Another reason is related to the nature or RNA. RNAs are involved in many gene regulation mechanisms and particularly with transcription or translation stages, meaning the significance of length dependent folding behavior as transcribed or translated[49] and thus requiring investigation of many sequences that varies in length but differ only with additionally added residues. N is the parameter that is decided by the biological context. We will investigate the N-dependency on the computational complexity later while presenting our results.

3.2 Effective Execution of Many-Stage Concurrent Multiple Tasks

Our strategy for exploring the RNA folding energy landscape combines the sampling and the analysis stages (Fig. 2). Each stage represents a scientific calculation that is carried out by a stand alone program such as Srna or analyses-scripts. Better performance as measured by lower time-to-solution (TTS) is achieved via the orchestration of multiple tasks and by the exploitation of concurrent task execution.

The concurrent execution of multiple tasks is accomplished by dividing a scientific calculation of each step into many parallel runs. For example, in Step I, multiple SFold instances are executed by assigning each instance a number of sampled structures — N/P, where P is the number of concurrent tasks/runs to be used. Similarly in Step II, the base-pair distance calculation is executed concurrently with P tasks. The concurrent execution of multiple tasks strategy utilizes the fact that the calculation of base-pair distance among Ω sampled structures is composed of four loops; the outer two loops iterate over all structures, i.e., taking $\Omega \times (\Omega - 1)$ iterations and two inner iterations are applied for base-pairs found in two different structures. Therefore, each parallel run for Step II (base-pair distance calculation) computes Ω/P iteration of the outermost loop while the three inner loops are intact. While these simple parallel strategies help to decrease time-to-solution at each step, the orchestration or the concurrent execution of many tasks are managed by the runtime environment (ADAMS).

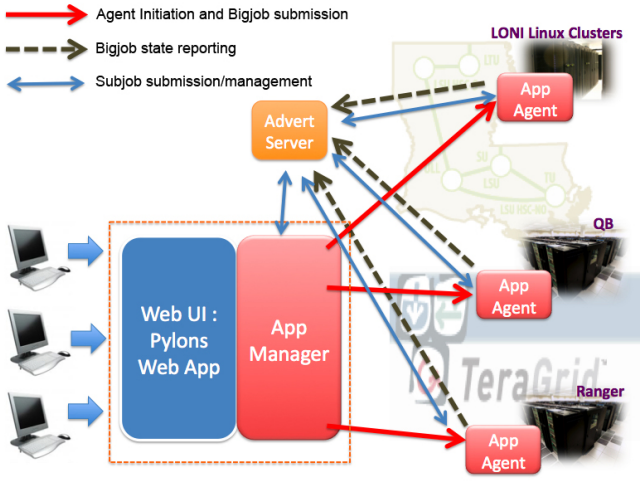

Figure 5: Schematic of Adpative Distributed Application Manger System (ADAMS). HPC machines affiliated with two federated grids, Teragrid and LONI, are seamlessly utilized by the lightweight, modular, and adaptive runtime environment. The arrows illustrate how BigJobs and Subjobs are submitted, monitored, and managed. More details on the components such as App Manager, App Agent, and Advert Server are described previously[35, 37].

The primary design goal of ADAMS is to support the two-stage pipeline. A schematic for how ADAMS manages the jobs is illustrated in Fig. 5. ADAMS aims to provide a run-

Table 3: Distributed High Performance Computing (HPC) resources utilized for this work. All CPUs are Intel Xeon processors. Ranger has a 2.66 GHz CPU and other LONI machines has a 2.33 GHz CPU as a core. For HPC grids affiliated with, TG represents Teragrid and L represents LONI. LONI small clusters, Eric, Oliver, Louie, and Poseidon are configured equivalently, so we call them LS collectively or E, O, L, and P, respectively if needed

HPC Name	Num. of Cores	HPC Grid	Node Spec. (Type, Memory)
Ranger (R)	5840	TG	2 dual-core, 8 GB
QB (Q)	2720	L & TG	2 quad-core, 8 GB
Eric (E/LS)	512	L	2 dual-core, 4 GB
Oliver (O/LS)	512	L	2 dual-core, 4 GB
Louie (L/LS)	512	L	2 dual-core, 4 GB
Poseidon (P/LS)	512	L	2 dual-core, 4 GB

time environment for executing scientific applications for a broad range of physical model sizes without actually having to change or tune the enviroment. Further, architectural goals are to implement a lightweight and modular runtime environment which can exploit a range of distributed resources. ADAMS supports the effective execution and management of applications that are executed as loosely-coupled/pleasingly-parallel tasks.

Web-based tool comprising an open source web app using Pylons (www.pylonshg.com) and core middleware scripts implementing the ADAMS framework have been developed as the DARE-Rfold and are freely accessible at `http://cyd01.cct.lsu.edu/dare-rfold`.

The ability to utilize multiple, distinct and heteregenous distributed computing resources represents a critical objective of achieving efficient execution of a target application. ADAMS supports the concurrent execution of multiple tasks by executing individual tasks in an adaptive manner by monitoring the dynamic resource conditions.

3.3 Computing resources

The High Performance Computing (HPC) machines utilized in this work are summarized in Table 3. These machines are part of two different federated grids, the US national Teragrid[2] and a statewide Louisiana federated grid systems, Louisiana Optical Network Initiative (LONI)[3]. Ranger is a Teragrid machine and Queenbee (QB) is affiliated with LONI as well as Teragrid.

4. RESULTS

4.1 Performance Analysis

We present benchmark results that underscore the efficacy of the ADAMS runtime environment. To this end, we compare the results using ADAMS with the "conventional mode" for executing an application. The conventional mode represents the situation without a runtime environment, and thus relies on the straightforward execution of a target application as a single task submitted to one specified machine.

For submission to a typical multi-user HPC system, the total time-to-solution, T_{tts} for completing a scientific application is composed of two components, the queuing waiting time (T_Q) and the actual execution time (T_E):

$$T_{tts} = T_Q + T_E. \quad (4)$$

For the following discussions, we ignore time for communication including file transfer since its contribution to time-to-solution is insignificant. T_E is further decomposed into two components, T_E^I and T_E^{II}, as shown in Fig. 2.

In previous works, we demonstrated the benefits of an interoperable and extensible pilot-job implementation (SAGA-BigJob) for the multi-physics (CFD/MD) applications[35] and Replica Exchange Molecular Dynamics (REMD)[37]. Both of these applications are examples of loosely coupled applications due to the fact that each sub-task are essentially independent, in that they do not require the use of MPI. We demonstrated that the use of multiple pilot-jobs running concurrently on multiple distinct resources, but coordinated and working towards a single problem instance has significant perfomance (as measured by lowered time-to-solution) advantages. The pilot-job approach reduces both T_Q and T_E.

4.1.1 The Case for Scaling-Out

The queue wait time is unavoidable on multi-user HPC systems due to scheduling systems, such as PBS. Conventionally, each application task encounters a wait time whenever submitted to a computer system. In contrast, an ADAMS-based runtime environment reduces multiple task waiting on a queue, to a single instance of queuing – since all tasks comprising are now executed as sub-tasks of the larger pilot-job abstraction (SAGA-BigJob) – and thus amenable to efficient execution using adaptive resource management and monitoring mechanisms.

Measured queue wait-times during two different time windows (separated by 24 hour interval) are presented in Table 4. We conducted the measurements twice to examine the impact of different loads on the targeted HPC systems that fluctuates with time. Also, we assume that the conventional execution mode uses one machine and the ADAMS-based execution mode utilize multiple machines. Three small LONI clusters are used for this test. The ADAMS-based execution requests three SAGA-BigJobs for three machines, and T_Q is measured as the time when the first BigJob identified by ADAMS executes (goes through the queueing system). As shown in Table 4, T_Q is lower when using three machines; this result is not surprising, since the utilization of multiple resources is an effective way if there is a runtime environment that enables dynamical resource assignment using resource monitoring. As shown in previous work[35, 37], subsequent BigJobs that go through the queing-system after the first BigJob, can also be utilized by dynamically re-assigning tasks to them, thus loering the time-to-solution further.

To facilitate a comparision of the conventional mode to the ADAMS-based approach we discuss our observations while gathering the second data set of Table 4. One of machines used for this data-set – Poseidon was heavily loaded and it took 236 minutes in the queue before running a BigJob. In other words, without resource monitoring and flexible exe-

Table 4: Comparison of the queue wait time, T_Q in the conventional mode using a computing resource vs. the ADAMS-based mode using three different resources. E/P/O represents three small LONI Linux clusters. The measured time is in seconds. One bigjob comprises 28 cores. The mean (m) and the standard deviation (σ) are compared

HPC Resources	Num. of Submission	m	σ
E/P/O	20	38.0	16.9
P	20	103.5	84.3
E/P/O	20	55.3	27.0
L	20	154.1	188.9
Ranger	20	263.4	220.1

cution, T_Q would be higher. In Table 4, results for the large Teragrid machine (Ranger) is compared to results using the LONI clusters. While the LONI clusters have the same architecture as well as system-wide configurations and are connected with the state-wide fast optical network, Ranger differs in many ways from the LONI systems we used in this study. Therefore it is likely to respond slow with a larger fluctuation in time as indicated. In fact, the utilization of Ranger represents an example of the cases with machines that is hard to be utilized due to different system-wide configurations. The result demonstrates the practical use of a large cluster with which the overall time-to-solution can be decreased through opportunistic decision and supporting of dynamic execution of tasks comprising the entire workflow. On the other hand, efficient execution to reduce T_E is enabled by supporting the concurrent execution of multiple tasks as described in the previous section. According to Table 5, the potential scaling-up varies for each step against the conventional mode – whereby the conventional mode is assumed to be a serial run since the target applications for two steps are serial applications. The results present the cases for SAM-I and S gene, respectively which is useful to estimate the computing cost depending on the size of system. Noticeably, the base-pair distance calculation are costly when non-concurrent implementations are executed. For example, sampling 10000 times for S gene using the conventional mode is expected to be considerably long, given that it takes 15 hours for sampling 1000 times.

Overall, the scaling for Step I is modest, mostly due to limitations of using the stand-alone (and monolithic) Sfold program. Currently, Sfold is available as an executable in which the partition function calculation and the subsequent sampling process are not separable; this limits the potential solutions for better performance. Possible future solutions and ideas include implementing parallelization of the partition function calculation, or incorporating a strategy in which the partition function calculation is carried out once and used for the subsequent sampling phase. On the contrary, the scaling for Step II is greater due to the ability to execute multiple, concurrent instances of a target application.

4.1.2 *Performance Dependency on Sequence Length*

We investigate the N-dependency for each step of our two-step applications, the results of which are shown in Fig. 6. These results validate our model for the computational complexity as presented in Eqn. 3. Furthermore they provide motivation for a general purpose, extensible infrastructure such as ADAMS.

Table 5: Execution time comparison between a conventional calculation (C) vs. calculations using ADAMS (A) that support concurrent execution of tasks. Two kinds of tasks, sampling (SA) in Step I and the base-pair distance (BPD) calculations carried out in Step II are compared. All calculations are carried out with LONI small clusters. (*scaling results are normalized by the result of a conventional calculation equivalently configured)

Seq. ID	Type Task	Num. Struct.	Time to Solution	Num. Tasks	Scaling*
I-A	SA (C)	10000	1 m	1	1.0
I-A	SA (A)	10000	31 s	10	1.94
II-A	SA (C)	10000	5 h 20 m 45 s	1	1.0
II-A	SA (A)	10000	4 h 01 m 11 s	10	1.32
II-A	SA (A)	10000	3 h 53m 19 s	25	1.37
II-B	SA (C)	10000	4 h 17 m 6 s	1	1.0
II-B	SA (A)	10000	3 h 49 m 26 s	10	1.1
II-B	SA (A)	10000	3 h 51 m 8 s	25	1.1
I-A	BPD (C)	10000	21 m 6 s	1	1.0
I-A	BPD (A)	10000	2 m 20 s	10	9.0
II-A	BPD (C)	1000	15 h 32 m 5 s	1	1.0
II-A	BPD (C)	1000	1h 46 m 30 s	10	8.8
II-A	BPD (A)	1000	12 m 6 s	100	77.8
II-A	BPD (A)	1000	3 m 35 s	500	310.7

Different Ω (1000 and 10000) counts for Step I are shown in Fig. 6 (a) and (b); as can be seen a modest scaling and thus advantage is seen in (b) when using ADAMS-based calculations, but not so for (a). During Step I, the first stage takes $O(N^3)$ – which is most of computing time, compared to the second stage with $O(N) \times O(\Omega)$. On the other hand, the second stage which depends on Ω, becomes more dominant as Ω increases, which then scales as concurrent tasks begin (as indicated in (b)).

The computational cost is expected to grow significantly as N increases. As suggested by Eqn. 3, the N-dependency of Step I and Step II is expected to be complex and different. Results using the conventional execution in Fig. 6(a), (b), and (c) show slightly different behavior as N increases. For example, as implied by the result in Fig.6(c), using a simple conventional implementations for base-pair calculations, for many sequences and large Ω, will be infeasible. But overall, the results shown in Fig.6 suggest better performance for ADAMS-based execution. Specifically the computationally demanding Step II (base-pair distance) calculations scales well due to the abilty to execute concurrent tasks using ADAMS over multiple resources, whilst the less computationally demanding Step I is not influenced as much by the ability to execute concurrently. The underlying reason for better scaling with Step II becomes evident with the results in Fig. 6(c). As seen in Eqn. 3, this calculation is suitable for task parallelization.

Therefore, results in Fig. 6 show not only the efficacy of ADAMS for scaling, particularly with Step II, but also provide the insight into the computational complexity that is

reflected in benchmark results. Taken together, algorithmic advances with the ability to execute many concurrent tasks over multiple resources and adapt to variable workloads has important performance advantages.

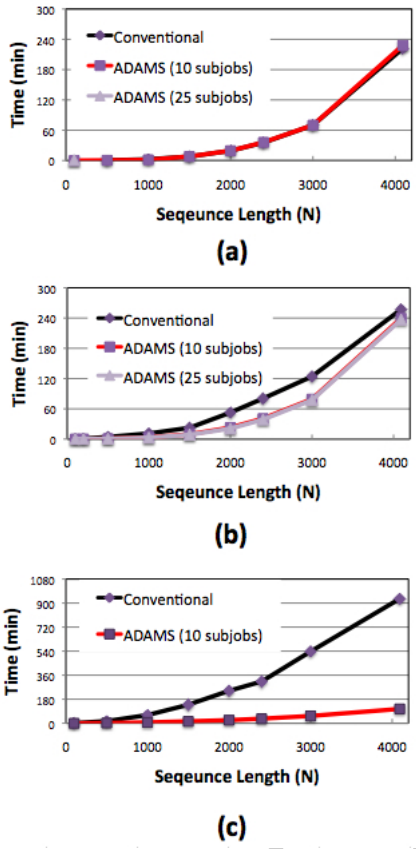

Figure 6: N dependent execution times of Step I and Step II. RNA sequences with a varying size of the number of nucleotides, N, are generated from Seq. II-B by taking a part of the sequence that starts from the first residue. Conventional execution mode with a single serial job is compared to ADAMS-based execution modes for which 10 or 25 parallel runs are executed at the same time. (a) Step I for 1000 structure sampling, whilst (b) Step I for 10000, and (c) Step II for base-pair distance calculation with 1000 sampled structures are presented.

4.2 Energy Landscape characteristics via the two measures

4.2.1 Variability of SAM-I sequences

According to calculated results, the energy distributions (Fig. 7) and the base-pair distributions(Fig. 9) of the eight SAM-I riboswitches are observed with different shapes, locations of peaks, and the range of values. The energy distributions show the distinctive asymmetric shape with more population of states close to the lowest energy state. In fact, according to the REM theory, a zeroth order approximation predicts a Gaussian distribution of the distribution of energies, and thus the results appear to suggest that these RNA sequences are too small to satisfy the simple assumption of the theory developed with the case of long random heteropolymers.

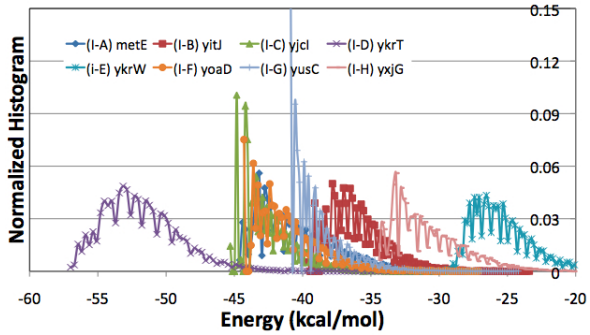

Figure 7: Energy distributions of sampled secondary structures of the eight SAM-I sequences. 10000 structures are sampled and distribution of energies is presented with the normalized histogram

To some extent, the energy distribution itself explains the variability among sequences and characteristic features. For example, Seq. I-G and two other sequence, I-C and I-F, result in distributions that have considerably populated lower energy states in the vicinity of the lowest energy state. On the other hand, other sequences are found with somewhat modest increment of number of states from the lowest energy state. The low energy states are considered to be important for the folding dynamics into the native state that is generally located around the lowest energy state or itself[48, 31]. More clear understanding on the roles of those states should be attempted with their connectivity. In that sense, the base-pair distribution provides additional information on the energy landscape. For instance, while energy states of Seq. I-A, I-C, and I-F are similarly distributed resulting much overlaps, corresponding base-pair distributions clearly reveal that they are distinctive different in terms of similarity or connectivity. Overall, the eight SAM-I sequences are expected to explore different folding dynamics along with different energy landscapes.

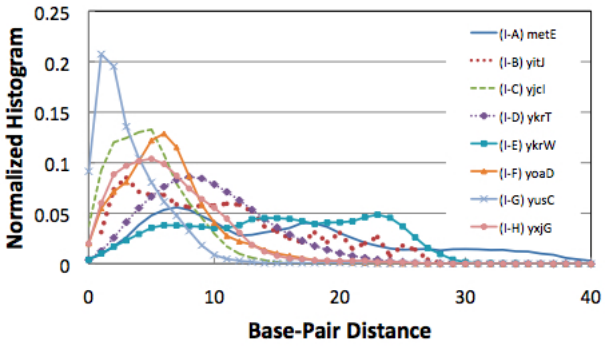

Figure 8: Base-pair distance distributions of the eight SAM-I sequences. The number of sampled structures used is 1000.

4.2.2 S gene region: Original vs. Codon-optimized

The two energy distributions corresponding to II-A and II-B are energetically separated since the overall G-C pairs and A-T pairs are changed due to the codon-optimization process. The energy distributions for the longer S gene sequences (Fig. 9) show a Gaussian-like shape, in contrast to shorter SAM-I riboswitch sequences (Fig. 7), indicating that the zeroth approximated theory of REM can be applied for this system[48]. Therefore, the observed slight difference in the shape of the distribution, in particular resulting in different widths can be used for arguing different ruggedness of the energy landscapes.

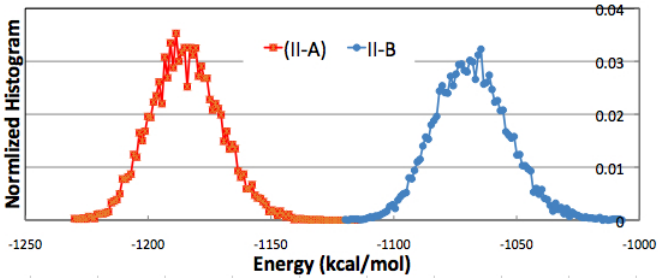

Figure 9: Energy distributions of sampled secondary structures of the two S gene sequences. 10000 structures generated by Boltzmann sampling are used.

However, it is the base-pair distributions that reveal somewhat striking difference between two S gene sequences. As shown in Fig. 10, a bimodal distribution is observed for the original sequence, but not for the optimized one, suggesting that the energy landscapes can be characteristically different in connectivity among sampled structures.

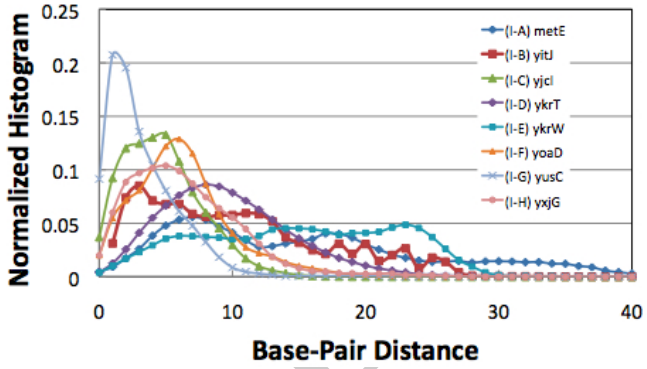

Figure 10: Comparison of the base-pair distance between two S gene sequences, Seq. II-A (red) and Seq. II-B (green). 1000 structures generated by Boltzmann sampling are used

5. DISCUSSION

Random Energy Model: Measuring the Ruggedness of Energy Landscape

The energy landscape perspective is a theoretical paradigm to understand complicated dynamics and interactions of macromolecular systems occurring on a multi-dimensional potential energy landscape[31, 63, 59, 6, 22]. Among many theoretical approaches embracing the perspective, the energy landscape theory of protein folding introduced by Wolynes and coworkers adopts the Random Energy Model (REM), the theory developed by Derrida for the spin-glass system[48, 12, 63]. The theory attempts a statistical mechanical treatment of the folding process, resulting in the theoretical conclusion that protein folding is the organization of an ensemble of structures on the rugged funnel-like energy landscape[48, 63]. As shown in the seminal work of Bryngelson and Wolynes for protein folding[12], in the simplest zeroth order approximation, the overall density of states is a Gaussian distribution. Consequently, the distribution of thermally weighted probabilities at a given temperature, i.e. Boltzmann ensemble of structures, also becomes a Gaussian distribution. In this simplest case, the ruggedness of the energy landscape is thus reflected in the fluctuation in the energy in the density of states, the width of a Gaussian distribution. A variant of REM, called the generalized REM (GREM) has been suggested for a better understanding of complicated folding energy landscapes. The key idea of this new approach is to incorporate the correlation among states[48, 33]. For example, according to GREM, the distribution of connected neighbors is considered as the critical component instead of the distribution of all states, i.e. the density of states. Therefore, the ruggedness is now interpreted in terms of the connectivity among states the system can visit in configuration space.

We propose that the base-pair distance, widely used as a measure of similarity between structures predicted from the same sequence[26], could be used as a measure of correlation or connectivity between two sampled structures. In other words, a base-pair distance from a structure I_a to another structure I_b is somehow linked to the enthalpic contribution of the activated barrier, since it corresponds to the number of base pairs to be broken for a potential transition pathway. Of course, the transition itself would be a multi-step process involving multiple pathways, but the base-pair distance would be a good quantity to estimate the proximity of two structures in terms of activation barrier for the transition[64, 30]. Interestingly, the importance of the Boltzmann weighted neighbor distribution was suggested by Clote and coworkers for RNA secondary structure prediction[25].

Energy Landscapes of SAM-I riboswitches and two S gene RNAs

According to our results, SAM-I riboswitches of the size of approximately 100 nts do not produce a Gaussian-like distribution with secondary structures obtained with Boltzmann-weighted sampling (Fig. 7), suggesting a caution for applying the REM theory for this size of systems. The breakdown of the REM with this small system, in fact, is not surprising. REM assumes the overall interactions is a sum of random interactions arising from local regions. In SAM-I riboswitches, an interaction arising from a local region is likely to be coupled with interactions of other regions. For example, local hairpin formation in one region could affect secondary structure formation occurring in other region. Indeed, in the AAT element, a complex 3-D structure is stabilized by interactions from four helical segments linked by a four way junction and long range tertiary interactions such as the formation of Pseudoknot as revealed by X-ray experiment[45] and our atomistic simulation study[29].

Interestingly, the long S gene sequences produce Gaussian-like energy distributions (Fig. 9). The REM therefore seems appropriate for this size of RNA. While the two different energy distributions of S gene RNAs are indicative of differing ruggedness for their folding energy landscape as shown, for example, via different widths, a more striking observation is the bimodal character of the base-pair distribution for the original S gene sequence(Fig. 10). One possible argument is that such a bimodal distribution, in particular with the additional peak at about 600, suggests the existence of large free energy barriers between groups of clusters, resulting in kinetic trapping. In contrast, the codon-optimized sequence would avoid kinetic trapping by finding the pathways with lower energy barrier requiring fewer base pair rearrangement. We observe distinct energy landscapes for each of eight SAM-I sequences for which experimental functional variability has been reported[58]. Our results can be further examined for more precise connection to the experimental findings. While each of eight sequences function intrinsically as a SAM-I riboswitch[28, 58], it is intriguing to understand how they manage to vary binding affinity as well as transcriptional termination efficiency. The implication of this variability for gene regulation mechanisms in living cells is of great importance, but poorly understood.

In this study, we found a part of clues for that quest for the variation of binding affinity. As a matter of fact, to estimate the binding affinity, or binding constant, the forward binding rate as well as the backward dissociation rate need to be considered together. The backward rate could be informed by using computational approaches such as Docking scores or Molecular Mechanics Possison Boltzmann Surface Area (MM-PBSA) method[60]. Note that these approaches aim to estimate the binding free energy, but due to large conformational changes, the estimate cannot reflect the initial unbound state properly. We have estimated binding free energy using MM-PBSA with atomistic MD simulations starting with the published X-ray structure (unpublished data). With this study, we intended to scrutinize the forward binding process applying the energy landscape perspective. The underlying assumption is that if there are structures that accommodate a SAM binding more readily, the population of those structures could affect the overall forward rate. Therefore, it is intriguing that the energy distributions as well as base-pair distributions indicate differences in population distribution, but it is difficult to draw definitive conclusions without the backward rate.

In summary, variation in the energy landscape ruggedness was observed among the eight SAM-I sequences, which along with differing landscapes for two BCoV S gene sequence, suggests the usefulness of the energy landscape perspective.

Toward a Pipeline for RNA 3-D Folding Prediction

Our investigation carried out in this study found that the ADAMS-based runtime environment is useful for scientific discovery with its easy deployment, scale-out strategy, and adaptive execution of tasks composed of the entire workflow. Also, we demonstrated that ADAMS manages efficiently parallel execution of target applications employed for specific calculations in a non-intrusive way. However, limitations of the current implementation are also found, for example, with suboptimal scalability of Step I. Perhaps, a simple solution for better parallel performance with Step I is to separate the first stage of the partition function calculation from the second stage of sampling; the first stage is only carried out and its output is used for the parallelized sampling stage.

Another limitation of our current approach is that we consider only secondary structure information. Currently, we are developing a pipeline approach aimed at predicting RNA 3-D folding. The main idea is to combine the use of RNA folding energy landscape, as demonstrated in this work, with the 3-D conformational structure sampling. Atomistic simulations start with configurations that are generated by 3-D modeling with the secondary structures obtained with this work. The scheme is illustrated schematically (Fig 1). After Boltzmann sampling of secondary structures, extensive atomic simulations are carried out to explore the starting basin of attractions or neighboring basins[34, 6]. This multi-resolution approach overcomes many difficulties raised when only the secondary structure prediction or only the atomistic simulations are applied. To this end, the development of a runtime environment efficiently managing Distributed Adaptive Replica Exchange (DARE) MD reported in the previous work[37], is incorporated into the current ADAMS runtime environment, resulting an extended workflow comprising many components but not expecting any major challenges due to the lightweight, modular, and extensible ADAMS.

Related Work

There are several workflow tools that could in principle be used. However, due to the unique requirements of this problem most workflow tools/systems do not provide the flexibility or generality that is needed. For example, an effective solution to the RNA landscape problem requires the ability to utilize multiple heterogeneous resources âĂŞ as part of the same workflow, i.e., there is a great variation in the computational requirements between stages. Most workflow tools and systems are not designed for such variation. For example, Pegasus/DAGMan confines the execution to resources of small to intermediate tasks.

Additionally, different input models lead to different execution time requirements. This could lead to either longer time-to-solutions (i.e. variation in time), or could lead to different computational resources (i.e. variation in size of resource – from large-scale MPI to mid-level parallelism that fits onto a many-core processor). Once again, typical workflow systems/tools are not designed to support such variations with input. It should be noted that it may not always be easy to predict temporal variations in advance.

Our approach is more amenable to being programmatically specified and modified than traditional workflows; in this sense it is more like Swift, but provides explicit control and capability for distributed coordination and execution. A consequence of this is that our approach is dependent on a reliable and robust âĂIJdistributed programming environmentâĂİ, the current lack of which makes the execution more onerous on the application scientist.

ADAMS, or more precisely the SAGA-based Pilot-Job is being extended for MapReduce tasks, including iterative MapReduce and MapReduce operating on dynamic data.

This work will be published in the near future.

6. CONCLUDING REMARKS

The energy landscape perspective is a theoretical departure from approaches relying upon a minimum free energy (MFE) structure or a small set of representative structures around the MFE structure, and provides a more rigorous description of a dynamic, polymorphic system than MFE[31]. It can be argued that methods that use the energy landscape perspective, in which the statistical description of an ensemble of structures is the main idea, may provide physical insights for RNA folding or complex formation that are not possible otherwise. Our scientific aim is to understand and to validate the energy landscape as a model for RNA folding and metabolite-recognition. We show how to construct the RNA folding energy landscape with secondary structure sampling using a Boltzamann-weighted scheme. We establish that due to significant increase in computing cost as the size of sampling and sequence length, or the number of sequences of interest become larger, an efficient runtime environment is critically required. In response to this computational challeng and complex multi-stage requirements, we have implemented ADAMS — an inter-operable, adaptive, extensible and scalable runtime environment. We have demonstrated its ease-of-deployment, modular and lightweight architecture, seamless utilization of heterogeneous HPC computing resources, and its ability to support dynamic execution (using general-purpose pilot-jobs). As a consequnce of being able to explore the energy landscape of RNA folding — both effectively and efficiently, our understanding of structured RNAs is expected to significantly advance.

7. ACKNOWLEDGMENTS

The authors would like to thank Drs. Gus Kousoulas and Vladimir Kouljenko for introducing the BCoV S gene sequences and sharing experimental results for gene expression. The authors also are grateful to Andre Luckow for his early work on SAGA Bigjob abstraction and related scripts, as well as Abhinav Thota for assistance with testing, and the SAGA development team at CCT. JK is grateful to Dr. Ye Ding and his Sfold team who share information on Sfold and Miss Sohyun Park for gathering results with SAM-I riboswitches with Sfold.

8. REFERENCES

[1] http://rfam.sanger.ac.uk.
[2] http://www.teragrid.org.
[3] http://www.loni.org.
[4] P. P. Amaral, M. E. Dinger, T. R. Mercer, and J. S. Mattick. The Eukaryotic Genome as an RNA Machine. *Science*, 319:1787–1789, 2008.
[5] D. Baek, J. Villen, C. Shin, F. D. Camargo, S. P. Gygi, and D. P. Bartel. The impact of microRNAs on protein output. *Nature*, 455:64–71, 2008.
[6] D. A. Beck, G. W. N. White, and V. Daggett. exploring the energy landscape of protein folding using replica-exchange and conventional molecular dynamics simulations. *J. Struct. Biol.*, 157(3):514–523, 2007.
[7] C. L. Beisel and C. D. Smolke. Design Principles for Riboswitch Function. *PLoS Computational Biology*, 5(4):e1000363, 2009.
[8] H. M. Berman, W. K. Olson, D. L. Beveridge, J. Westbrook, A. Gelbin, T. Demeny, A. R. Hsieh, S. H. Srinivasan, and B. Schneider. The nucleic acid database: a comprehensive relational database of three-dimensional structures of nucleic acids. *Biophysical J.*, 63:751–759, 1992.
[9] E. Birney and et al. Identification and analysis of functional elements in 1human genome by the ENCODE pilot project. *Nature*, 447(7):799–816, 2007.
[10] S. Blouin, J. Mulhbacher, J. C. Penedo, and D. A. Lafontaine. Riboswitches: Ancient and Promising Genetic Regulators. *ChemBioChem*, 10:400–416, 2009.
[11] S. Bonhoeffer, J. S. McCaskill, P. F. Stadler, and P. Schuster. RNA multi-structure landscapes: A study based on temperature partition functions. *Eur. Biophys. J.*, 22:13–24, 1993.
[12] J. D. Bryngelson and P. G. Wolynes. Intermediates and barrier crossing in a Random Energy Model (with application to protein folding). *J. Phys. Chem.*, 93:6902–6915, 1989.
[13] S. Chen and K. A. Dill. RNA folding energy landscapes. *Proc. Natl. Acad. Sci., USA*, 97(2):646–651, 2000.
[14] V. N. Chouljenko and K. G. Kousoulas. private communication, 2010.
[15] V. N. Chouljenko, X. Q. Lin, J. Storz, K. G. Kousoulas, and A. E. Gorbalenya. Comparison of genomic and predicted amino acid sequences of respiratory and enteric bovine coronaviruses isolated from the same animal with fatal shipping pnemonia. *Journal of General Virology*, 82:2927–2933, 2001.
[16] B. F. C. Clark. The crystal structure of tRNA. *Journal of Biosciences*, 31(4):453–457, 2006.
[17] A. J. Collier, J. Gallego, R. Klinck, P. T. Cole, S. J. Harris, G. P. Harrison, F. Aboul-ela, G. Varani, and S. Walker. A conserved RNA structure within the HCV IRES elF3-binding site. *Nature Struct. Biology*, 9(5):375–380, 2002.
[18] J. A. Cruz and E. Westhof. The dynamic landscapes of RNA architecture. *Cell*, 136:604–609, 2009.
[19] M. D. Dambach and W. C. Winkler. Expanding roles for metabolite-sensing regulatory RNAs. *Curr. Op. Microbiol.*, 12:161–169, 2009.
[20] Y. Ding. Statistical and Bayesian approaches to RNA secondary structure prediction. *RNA*, 12(3):323–331, 2006.
[21] Y. Ding, C. Y. Chan, and C. E. Lawrence. Sfold web server for statistical folding and rational design of nucleic acids. *Nuc. Acids Res.*, 32:W135–W141, 2004.
[22] C. M. Dobson. Protein folding and misfolding. *Nature*, 426:884–890, 2003.
[23] S. R. Eddy. How do RNA folding algorithms work? *Nature Biotechnology*, 22(11):1457–1458, 2004.
[24] V. Epshtein, A. S. Mironov, and E. Nudler. The riboswitch-mediated control of sulfur metabolism in bacteria. *Proc. Natl. Acad. Sci., USA*, 100:5052–5056, 2003.
[25] E. Freyhult, V. Moulton, and P. Clote. Boltzmann probability of RNA structural neighbors and riboswitch detection. *Bioinformatics*, 23(16):2054–2062, 2007.
[26] A. R. Gruber, R. Lorenz, S. H. Bernhart, R. Neubock, and I. L. Hofacker. The Vienna RNA websuite. *Nucleic Acids Res.*, 36:W70–W74, 2008.
[27] J. M. Hart, S. D. Kennedy, D. H. Mathews, and D. H. Turner. NMR-Assisted Prediction of RNA Secondary Structure: Identification of a Probable Pseudoknot in the Coding Region of an R2 Retrotransposon. *J. Am. Chem. Soc.*, 130:10233–10239, 2008.
[28] T. M. Henkin. Riboswitch RNAs : using RNA to sense cellular metabolism. *Genes and Development*, 22:3383–3390, 2009.
[29] 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 Acids Res.*, 37(19):6528–6539, 2009.

[30] C. Hyeon and D. Thirumalai. Multiple Probes are required to explore and control the rugged energy landscape of RNA hairpins. *J. Am. Chem. Soc.*, 130:1538–1539, 2008.

[31] 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.*, 5:88–91, 1997.

[32] G. F. Joyce and L. E. Orgel. pages 49–77. Cold Spring Harbor Laboratory Press, New York, 1999.

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

[34] 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.*, 121:4237, 2004.

[35] S. Ko, N. Kim, J. Kim, A. Thota, and S. Jha. Efficient Runtime Environment for Coupled Multi-Physics Simulations: Dynamic Resource Allocation and Load-Balancing. accepted in IEEE CCGrid 2010.

[36] K. H. Link and R. R. Breaker. Engineering ligand-responsive gene-control elements: lessons learned from natural riboswitches. *Gene Ther.*, 16:1189–1201, 2009.

[37] A. Luckow, S. Jha, J. Kim, A. Merzky, and B. Schnor. Adaptive distributed replica-exchange simulations. *Philosophical Transactions of the Royal Society A: Mathematical,Physical and Engineering Sciences*, 367(1897):2595–2606, 2009.

[38] A. Luckow, L. Lacinski, and S. Jha. Saga bigjob: An extensible and interoperable pilot-job abstraction for distributed applications and systems. 2010.

[39] H. Ma, D. J. Proctor, E. Kierzek, R. Kerzek, P. C. Bevilacqua, and M. Gruebele. Exploring the energy landscpae of a small RNA hairpin. *J. Am. Chem. Soc.*, 128:1523–1530, 2006.

[40] D. H. Mathews. Revolutions in RNA secondary structure prediction. *J. Mol. Biol.*, 359:526–532, 2006.

[41] D. H. Mathews, M. D. Disney, J. L. Childs, S. J. Schroeder, M. Zuker, and D. H. Turner. Incorporating Chemical Modification Constraints into a Dynamic Programming Algorithm for Prediction of RNA Secondary Structure. *Proc. Natl. Acad. Sci. USA*, 101:7287–7292, 2004.

[42] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, 288(5):911–940, 1999.

[43] D. H. Mathews and D. H. Turner. Prediction of RNA secondary structure by free energy minimization. *Curr. Opin. Struct. Biol.*, 16(3):270–278, 2006.

[44] J. S. McCaskill. The Equilibrium Partition Function and Base Pair Binding Probabilities for RNA Secondary Structure. *Bioploymers*, 29:1105–1119, 1990.

[45] R. K. Montange and R. T. Batey. Structure of the S-adenosylmethionine riboswitch regulatory mRNA elemen. *Nature*, 441:1172–1175, 2006.

[46] R. K. Montange and R. T. Batey. Riboswitches: emerging themes in RNA structure and function. *Annual Review of Biophysics*, 37:117–133, 2008.

[47] J. M. Ogle, A. P. Carter, and V. Ramakrishnan. Insights into the decoding mechanism from recent ribosome structures. *Trends in Biochemical Sciences*, 28:259–266, 2005.

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

[49] G. Quarta, N. Kim, J. A. Izzo, and T. Schlick. Analysis of Riboswitch Structure and Function by an Energy Landscape Framework. *Nucleic Acids Res.*, 35:5370–5378, 2007.

[50] J. D. Robertus, J. E. Ladner, D. Rhodes, R. S. Brown, B. F. Clark, and A. Klug. Correlation between three-dimensional structure and chemical reactivity of transfer RNA. *Nuc. Acids Res.*, 1(7):927–932, 1974.

[51] A. Roth and R. R. Breaker. The Structural and Functional Diversity of Metabolite-Binding Riboswitches. *Annual Review of Biochemistry*, 78:305–334, 2009.

[52] SAGA - A Simple API for Grid Applications. http://saga.cct.lsu.edu/.

[53] T. M. Schmeing and V. Ramakrishnan. What recent ribosome structures have revealed about the mechanism of translation. *Nature*, 461:1234–1242, 2009.

[54] A. Serganov and D. J. Patel. Ribozymes, riboswitches and beyond: regulation of gene expression without proteins. *Nat. Rev. Genet.*, 8:776–790, 2007.

[55] B. A. Shapiro, Y. G. Yingling, W. Kasprzak, and E. Bindewald. Bridging the gap in RNA structure prediction. *Curr. Op. Struct. Biol.*, 17:157–165, 2007.

[56] I. Shcherbakova, S. Mitra, A. Laederach, and M. Brenowitz. Energy barriers, pathways and dynamics during folding of large, multi-domain RNAs. *Curr. Opin. Chem. Biol.*, 12(6):655–666, 2008.

[57] S. Solomatin, M. Greenfeld, S. Chu, and D. Herschlag. Multiple native states reveal persistent ruggedness of an RNA folding landscape. *Nature*, 463:681–684, 2010.

[58] J. Tomsic, B. A. McDaniel, F. J. Grundy, and T. M. Henkin. Natural variability in S-Adenosylmethionine (SAM)-dependent riboswitches: S-Box elements in Bacillus subtilis exhibit differential sensitivity to SAM in vivo and in vitro. *Journal of Bacteriology*, 190:823–833, 2008.

[59] D. J. Wales and T. V. Bogdan. Potential Energy and Free Energy Landscape. *J. Phys. Chem. B*, 110(42):20765–20776, 2006.

[60] J. Wang, P. Morin, W. Wang, and P. A. Kollman. Use of MM-PBSA in reproducing the binding free energies to HIV-1 RT of TIBO derivatives and predicting the binding mode to HIV-1 RT of efavirenz by docking and MM-PBSA. *J. Am. Chem. Soc.*, 123(22):5221–5230, 2001.

[61] J. M. Watts, K. K. Dang, R. J. Gorelick, C. W. Leonard, J. W. Bess Jr., R. Swanstrom, C. L. Burch, and K. M. Weeks. Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature*, 460:711–716, 2009.

[62] K. A. Wilkinson, R. J. Gorelick, S. M. Vasa, N. Guex, A. Rein, D. H. Mathews, M. C. Giddings, and K. M. Weeks. High-Throughput SHAPE Analysis Reveals Structures in HIV-1 genome RNA strongly conserved across distinct biological states. *PLoS Biology*, 6(4):e96, 2008.

[63] P. G. Wolynes. Landscape, Funnels, Glasses, and Folding : From Metaphor to Software. *Proceedings of American Philosophical Society*, 145(4):555–563, 2001.

[64] Y. Yao, J. Sun, X. Huang, G. R. Bowman, G. Singh, M. Lesnick, L. J. Guibas, V. J. Pande, and G. Garlsson. Topological Methods for Exploring Low-density States in Biomolecular Folding Pathways. *J. Chem. Phys.*, 130(14):144115, 2009.

[65] M. Zucker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, 9(1):133–148, 1981.

[66] M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, 31(13):3406–3415, 2003.