Tools for Simulating and Analyzing RNA Folding Kinetics*

Xinyu Tang, Shawna Thomas, Lydia Tapia, and Nancy M. Amato

Parasol Lab, Dept. of Comp. Sci., Texas A&M University, College Station, TX 77843

Abstract. It has recently been found that some RNA functions are determined by the actual *folding kinetics* and not just the RNA's nucleotide sequence or its native structure. In this paper, we present our new computational tools that can study such functions by simulating the RNA folding kinetics. We present our prediction results of (1) the relative plasmid replication rates of ColE1 RNAII and its mutants and (2) the relative gene expression rates of MS2 phage RNA and its mutants. These results show that our approach computes the same relative functional rates as seen in experiments. Our computational tools can study RNA folding kinetics such as population kinetics, folding rates, and the folding of particular subsequences. Our method follows our previous work by first building an approximate map (or model) of the RNA's folding energy landscape. Next, we analyze the population kinetics of the roadmaps by solving the Master Equation. We can also use our new analysis technique, Map-based Monte Carlo (MMC) simulation, to stochastically extract folding pathways from the map. We first compare our methods with other computational methods working on the complete energy landscape showing that our small roadmap can capture the major landscape features. Moreover, we show that our method scales well to large RNA, e.g., 200+ nucleotides. Finally, we demonstrate that our method can correctly predict kinetics-based functional rates of ColE1 RNAII and MS2 phage RNA and their mutants.

1 Introduction

Ribonucleic acid (RNA) performs diverse and important functions such as synthesizing proteins, catalyzing reactions, splicing introns, and regulating cellular activities [27, 15, 3]. It was once believed that an RNA's functions are primarily determined by its nucleotide sequence and native state. However, it has been recently found that some RNA functions are determined by the folding process

* Supported in part by NSF Grants ACI-9872126, EIA-9975018, EIA-0103742, EIA-9805823, ACR-0081510, ACR-0113971, CCR-0113974, EIA-9810937, EIA-0079874 and by the DOE. Thomas supported in part by a Department of Education GAANN Fellowship and previously supported by a NSF Graduate Research Fellowship and a P.E.O. Scholarship. Tapia supported in part by a NIH Molecular Biophysics Training Grant (T32GM065088) and previously supported by a Department of Education GAANN Fellowship.

itself and not just the sequence and native state. For example, RNA folding kinetics may regulate the plasmid copy number, e.g., accelerating the refolding speed of RNA II can increase the E. coli ColE1 plasmids copy number [10]. In a similar way, the velocity of RNA folding can also regulate gene expression at the translational level. Some mutations of IS10 transposase mRNA slow folding kinetics of structure formation, and thus increase the ribosome-binding rate, which results in higher expression of IS10 transposase [16]. It has also been shown that the mRNA folding kinetics regulate the expression of phage MS2 maturation protein [9, 15, 11]. The mRNA acts as a regulator only when a particular subsequence is open. Since it is closed in the native state, this can only happen before folding finishes. The longer the RNA stays in a open metastable state, the higher the gene expression rate. Thus, it is imperative to have a computational method that can study both the global properties of RNA folding and more detailed features related to kinetics-based functions.

In this work, we provide computational tools to approximate the folding energy landscape and extract both global properties and detailed features of the folding process. The key advantage of our approach over other computational techniques is that it is fast and efficient while bridging the gap between high-level folding events and low-level folding details. Our method first builds a map (or model) of the RNA's folding energy landscape. We then use a new analysis technique called Map-based Monte Carlo simulation to extract folding pathways in a stochastic way. With this analysis tool, along with others previously developed, we can study population kinetics, folding rates, and the folding of particular subsequences. These techniques allow us to study kinetics-based functions that we could not study before. In addition, we extend the size of RNA our method can handle hundreds of nucleotides by using a new statistical sampling method.

We validate our methods against other computational methods (Monte Carlo Simulation) and experimental data. We first compare the kinetics of our small approximated roadmap with the complete energy landscape. We show that our roadmaps can capture the major features of much larger energy landscapes efficiently. We also demonstrate that our method scales well to large RNA with hundreds of nucleotides. Finally, we present two cases to demonstrate how our method can study kinetics-based functions. First, we compare simulated folding rates for ColE1 RNAII and its mutants against experimental rates. We show that we compute the same relative folding order as seen in experiment. Second, we predict the gene expression rates of wild-type MS2 phage RNA and three other of its mutants and match them to experiment. Again, we show that we predict the same relative functional rates as seen in experiment. In this paper, we provide results for RNAs with up to 200 nucleotides, and we anticipate that our technique can be used for even larger RNA.

2 Preliminaries

An RNA molecule is a sequence of nucleotide bases. There are four types of bases: adenine (A), cytosine (C), guanine (G), and uracil (U). The complementary

Watson-Crick bases, C-G and A-U, form stable, hydrogen bonds (*base pairs*) when they form a contact. The wobble pair, G-U, constitutes another strong base pair. These are the three most commonly considered base pairings [28, 32] and are what we consider in our model.

RNA Structure. *Tertiary structure* is a 3D spatial RNA conformation of a set of base pairs. *Secondary structure* is a planar representation of an RNA conformation. Although there are several slightly different accepted definitions [4, 12], secondary structure is usually considered to be a planar subset of the base pair contacts present. Non-planar contacts, often called *pseudo knots*, are not allowed in secondary structure. We adopt the definition in [12] that eliminates other types of contacts that are not physically favored: (1) contacts must be separated by at least 3 other bases, (2) each base cannot be involved in more than 1 contact, and (3) contacts must be planar.

Tertiary structure gives the most complete representation of RNA structure. However, secondary structure is commonly used [32, 28, 12] because in many cases it provides sufficient information to study many aspects of folding while dramatically reducing the size of the conformation space to explore. One justification for this simplification is that research has shown the RNA folding process is hierarchical, i.e., secondary structure forms before tertiary structure [27]. In this work, we focus on the first stage: secondary structure formation.

Energy Calculations. To model the RNA folding energy landscape, we must be able to calculate the energy of any conformation. We use a common energy function called the Turner or nearest neighbor rules [32]. This method involves determining the types of loops that exist in the molecule and looking up their free energy in a table of experimentally determined values. Intuitively, adjacent contacts typically form stable subunits called stacks or stems that have low energy. Much work has been done to improve the accuracy of these rules.

3 Related Work

Computational research on RNA folding falls into two main categories: structure prediction and folding kinetics. Structure prediction attempts to compute the native state given only the nucleotide sequence. Folding kinetics, on the other hand, is concerned with the folding process itself and not just the end result.

Structure Prediction. Structure prediction is commonly solved with dynamic programming. Nussinov introduced a dynamic programming solution to find the conformation with the maximum number of base pairs [20]. Zuker and Stiegler [28] formulated an algorithm to address the minimum energy problem. Today, Zuker's MFOLD algorithm is widely used for structure prediction. McCaskill's algorithm [17] uses dynamic programming to calculate the partition function, i.e., the the sum of Boltzmann factors over all possible secondary structures, while Chen [4] uses matrices to approximate the partition function over all possible conformations. Eddy and Dirks et. al. [22, 7] include pseudo-knots in their structure prediction algorithms. Partly due to the inaccuracy of the energy model, the prediction of pseudo-knot structures are typically less accurate.

Folding Kinetics. Several approaches have been used to study RNA kinetics. For example, Flamm et. al. [8, 11, 31] used Monte Carlo algorithms to find folding pathways while Gultyaev and Shapiro et. al. [10, 23] used genetic algorithms. Isambert [31] extended Monte Carlo method to consider pseudo-knots.

Some methods involve computations on the folding landscape. Dill [4] used matrices to compute the partition function over all possible structures and approximate the complete folding landscape. Ding and Lawrence [6] extended McCaskill's algorithm to generate statistical samplings of RNA structures based on the partition function. Wuchty [30] modified Zuker's algorithm to generate all secondary structures within some given energy range of the native structure. Flamm and Wolfinger [8, 29] extended this algorithm to find local minima within some energy threshold of the native state and connect them via energy barriers. The resulting energy barrier tree represents the energy landscape. To calculate the energy barrier, they used a flooding algorithm that is exponential in the size of RNA. Thus, it is impractical for large RNA.

Some statistical mechanical methods are also used to study the RNA folding kinetics. For example, the master equation is used to compute the population kinetics of the folding landscape. It uses a matrix of differential equations to represent the transition probabilities between conformations. Once solved, the dominate modes of the solution describe the general folding kinetics [21, 13, 4].

RNA Folding with PRMs. Our approach is based on the *probabilistic roadmap* (PRM) technique for motion planning [14]. Motion planning determines valid paths to move objects from one conformation to another. PRMs build graphs (roadmaps) that approximate the topology of the feasible planning space by first sampling valid conformations (nodes) and connecting them with feasible transitions (edges). Note that it is impractical (and generally not necessary) to attempt all possible connections. Thus, connections are only attempted between neighboring conformations according to some distance metric.

In previous work, we used PRMs to study protein folding [2, 1, 24, 26] and RNA folding [25]. Roadmaps produced by these methods approximate the folding energy landscape. We obtained promising results that validated against experimental data and were even able to observe subtle folding differences between structurally similar proteins [26]. We were also able to validate the population kinetics of several small RNA against experiment [25].

4 Computational Methods

Our method first constructs a roadmap to approximate the energy landscape. Then we use our map-based tools to analyze the energy landscape. In our previous work, we presented two sucssessful roadmap construction techniques: base-pair enumeration(BPE) and stack-pair enumeration (SPE). While the results were promising, they were limited to small RNA (less than 40 nucleotides). In this work, we devlop Probabilistic Boltzmann Sampling (PBS) method to build smaller (up to 10 orders of magnitude smaller than BPE) maps which enables us to study much larger RNA. We provide several map-based analysis tools includ-

ing Map-based Master Equation (MME) and Map-based Monte Carlo (MMC) simulation to extract folding kinetics. MME can extract global properties such as folding rates and transition states, while MMC can extract microscopic features of the folding process, e.g., subsequence formation order.

4.1 Using Roadmaps to Describe Energy Landscapes

The goal of roadmap construction is to approximate the energy landscape and capture its important features. The quality of this approximation highly depends on the quality of the sampling and connection methods. We will describe each sampling and connection method in more details below.

Roadmap Node Sampling. Our method is general and thus can use conformations generated by any techniques. In our preliminary work, we used three methods for generating RNA conformations: complete base-pair enumeration (for small RNA), stack-pair enumeration, and maximal-contact sampling. While stack-pair enumeration approximated the energy landscape well, it is limited to small RNA where enumeration is feasible (e.g., 40 nucleotides or less). Here we provide a new sampling method for larger RNA.

Probabilistic Boltzmann Sampling (PBS). Wuchty [30] proposes a dynamic programming algorithm to enumerate suboptimal (low energy) conformations within a given energy threshold. However, as the size of the RNA or the threshold increases, the number of generated nodes increases exponentially. Thus, it is difficult for this method to generate high energy nodes. In our method, we use these suboptimal conformations as "seeds" and augment the sampling with additional random conformations. Then, we use a probabilistic filter to retain a subset of the conformations based on their Boltzmann distribution factors. For a given conformation i with free energy E_i, the probability P_i to keep it is:

$$P_i = \begin{cases} e^{\frac{-(E_i - E_0)}{kT}} & \text{if } (E_i - E_0) > 0 \\ 1 & \text{if } (E_i - E_0) \leq 0 \end{cases} \quad (1)$$

E_0 is a reference energy threshold which we can use to control the number of samples kept. In this way, we may generate more conformations probabilistically with the Boltzmann distribution which prefers low energy conformations but will allow some high energy conformations. Our results indicate that this sampling method captures the important features of the energy landscape well.

Roadmap Node Connection. Once we have a set of samples, we must connect them to form an approximate map of the energy landscape. As mentioned earlier, it is impractical (and generally not necessary) to attempt all possible connections. Instead, we attempt to connect a node with the k closest neighboring conformations according to some distance metric. k is a constant value specified by user. Then each pair of neighboring roadmap nodes are connected using a local planner. This is a commonly used strategy [14].

To connect a given pair of conformations, we need to compute a transition path (i.e., intermediate conformations) between them and approximate the Boltzmann transition probability which is stored as an edge weight in the

roadmap. Note that these two goals are not always the same. For example, when conformations are close to each other, one single (most energetic) transition path may dominate the transition probability. However, when conformations are far apart, there might be many possible transition paths where none dominate.

In our previous work, we presented a simple greedy algorithm that generates a single transition path and computes the transition probability from that path. It works well when conformations are close to each other. However, as the size of RNA increases and thus the feasible sampling density decreases, this method fails. Here we present methods designed to compute transition probabilities and to generate transition pathways that do not have these problems.

Computing the Transition Probability. When an edge (q_i, q_j) is added to the roadmap, it is assigned a weight W_{ij} that reflects the Boltzmann transition probability between its two end points q_i and q_j. First, we find the stable subunits (stems) that are different between q_i and q_j. We calculate the nucleation cost for each stem (which is the energy barrier to form each stem) and find the maximum cost. This maximum cost is an energy barrier E_b the folding process must go over to form all the stems. We use E_b to estimate the transition probability between q_i and q_j. This strategy is widely used in Monte Carlo simulations [11, 31] and genetic algorithms for folding pathways [10, 23].

We calculate the Boltzmann transition probability K_{ij} (or transition rate) of moving from q_i to q_j using Metropolis rules [5]:

$$K_{ij} = \begin{cases} e^{\frac{-\Delta E}{kT}} & \text{if } \Delta E > 0 \\ 1 & \text{if } \Delta E \leq 0 \end{cases} \tag{2}$$

where $\Delta E = max(E_b, E_j) - E_i$, k is the Boltzmann constant, and T is the temperature. Note that the same energy barrier E_b is also used to estimate the transition probability K_{ji}, so the calculation satisfy the detailed balance:

$$\frac{K_{ij}}{K_{ji}} = e^{\frac{-(E_j - E_i)}{kT}} \tag{3}$$

Thus, the edge weight W_{ij} is:

$$W_{ij} = -log(K_{ij)}) = \frac{-\Delta E}{kT}. \tag{4}$$

(Negative logs are used since $0 \leq K_{ij} \leq 1$.) By assigning the weights in this manner, we can easily extract the most energetically feasible path in our roadmap using simple graph search algorithms.

Generating Transition Pathways. First, we find the stems between the start and goal configurations and calculate their nucleation costs. Then we generate a transition pathway connecting the start and the goal configuration by probabilistically opening/closing the stems based on their nucleation cost.

4.2 Map-based Analysis Tools

In this section we describe several different map-based analysis tools including Map-based Master Equation (MME) and Map-based Monte Carlo (MMC) simulation to study folding kinetics. As implied by their names, MME and MMC

are variants of Master Equations and Monte Carlo simulations that work on our maps. Basically, the master equation calculates global properties of the folding process while Monte Carlo simulations provide details on individual folding pathways. However, they can both produce population kinetics, one directly and the other indirectly. Given an ensemble of Monte Carlo simulation pathways, we can can compute the population kinetics of a particular conformation by summing up its population in each pathway for every time step. This approach is less accurate and will take more time and space than using the master equation directly. However, it does not have the same numerical limitations as the master equation and can handle much larger RNA. In our experimental results (Section 5.1), we empirically compare the population kinetics by the master equation, standard Monte Carlo simulation (implementation of Vienna Package), and MMC.

Map-based Monte Carlo Simulation. The folding process is stochastic rather than deterministic [13]. Transitioning from one conformation to another is probabilistically biased by the transition probabilities. Monte Carlo method [19, 13] simulate this random walk in the real (or complete) energy landscape. Kinfold is a well-known implementation of Monte Carlo simulation in the ViennaRNA Package [8]. These simulations can be computationally intensive since at each step they must calculate the local energy landscape to choose the next step.

In previous work, we simply extracted the most energetically feasible path in the roadmap to study the folding process. However, this does not mirror the stochastic folding process. Instead in this work, we apply Monte Carlo simulation directly to our roadmaps since the roadmap is just an approximation of the energy landscape where edge weights reflect Boltzmann transition probabilities.

Similar to the Monte Carlo simulation, our method starts from a random node in this roadmap and iteratively chooses a next node based on the transition probabilities. Because the edge weight W_{ij} encodes the transition probabilityK_{ij} between two endpoints i and j (see equation 4), we can calculate K_{ij} as $K_0 e^{-W_{ij}}$ where K_0 is a constant adjusted according to experimental results.

To generate the transitional conformations between two nodes, we use the method described in Section 4.1. Results presented here are generated using a fast variant of the standard Monte Carlo method [19].

Population Kinetics and Map-based Master Equation. Population kinetics is the time evolution of different conformation populations. It provides information such as folding rate, equilibrium distribution, and transition states, which can be correlated to experimental results. Here, we present our Map-based Master Equation (MME) method to analyze the population kinetics.

Master equation formalism has been developed for folding kinetics in a number of earlier studies [13, 4] The stochastic folding process is represented as a set of transitions among all n conformations (states). The time evolution of the population of each state, $P_i(t)$, can be described by:

$$dP_i(t)/dt = \sum_{i \neq j}^{n} (K_{ji} P_j(t) - K_{ij} P_i(t)) \tag{5}$$

where K_{ij} denotes the transition rate (probability) from state i to state j. The change in population $P_i(t)$ is the difference between transitions *to* and *from* state i. We compute transition rates from the roadmap's edge weights: $K_{ij} = K_0 e^{-W_{ij}}$. K_0 is a constant adjusted according to experimental results.

If we use an n-dimensional column vector $\mathbf{p}(\mathrm{t}) = (P_1(t), P_2(t), \ldots, P_n(t))'$ to denote the population of all n conformational states, then we can construct an $n \times n$ matrix M to represent the transitions, where

$$\begin{cases} M_{ij} = K_{ji} & i \neq j \\ M_{ii} = -\sum_{i \neq j} K_{ij} & i \neq j \end{cases} \tag{6}$$

The master equation can be represented in matrix form:

$$d\mathbf{p}(t)/dt = M\mathbf{p}(t). \tag{7}$$

The solution to the master equation is:

$$P_i(t) = \sum_k \sum_j N_{ik} e^{\lambda_k t} N_{kj}^{-1} P_j(0) \tag{8}$$

where N is the matrix of eigenvectors N_i for the matrix M in equation 6 and λ_i is its corresponding eigenvalue. $P_j(0)$ is the initial population of conformation j.

From equation 8, we see that the eigenvalue spectrum is composed of n modes. If sorted by magnitude in ascending order, the eigenvalues include $\lambda_0 = 0$ and several small magnitude eigenvalues. Since all the eigenvalues are negative, the population kinetics will stabilize over time. The population distribution $\mathbf{p}(t)$ will converge to the equilibrium Boltzmann distribution, and no mode other than the mode with the zero eigenvalue will contribute to the equilibrium. Thus the eigenmode with eigenvalue $\lambda_0 = 0$ corresponds to the stable distribution, and its eigenvector corresponds to the equilibrium Boltzmann distribution.

By a similar argument, large magnitude eigenvalues correspond to fast folding modes, i.e., those which fold in a burst. Their contribution to the population will die away quickly. Conversely, small magnitude eigenvalues have a large influence on the global folding process, and thus determine the global folding rates.

5 Results and Discussion

Here we present our simulation results and validate our methods against both other computational method (Monte Carlo Simulation) and also experimental data. The computational validations show that our small roadmaps can capture the major features of much larger complete energy landscapes efficiently. The roadmaps scale well to the RNA length, which enables us to study larger RNA with hundreds of nucleotides. The experimental validation shows that our methods correctly computed the kinetics-based functions of two different RNA and their mutants by studying two different properties of the folding kinetics.

In 5.1, we compare the population kinetics using our roadmaps against other computational methods working on complete energy landscapes. We first quantitatively compare the population kinetics computed from different maps and

show we can capture the major features of larger complete folding landscapes using much smaller roadmaps. Then, we empirically compare the scalability of our methods on different RNA. We present population kinetics using three different analysis methods: Map-based Master Equation (MME), Monte Carlo (MC) simulation, and Map-based Monte Carlo (MMC) simulation. The results show that the solutions of different methods are comparable to each other. They also indicate that our roadmaps scale well for large RNA. In section 5.2, we present two cases to demonstrate how we can use our method to study kinetics-based functions. Our method correctly predicts (1) the relative plasmid replication rates of ColE1 RNAII and its mutants and (2) the relative gene expression rates of MS2 phage RNA and its mutants.

5.1 Computational Validations

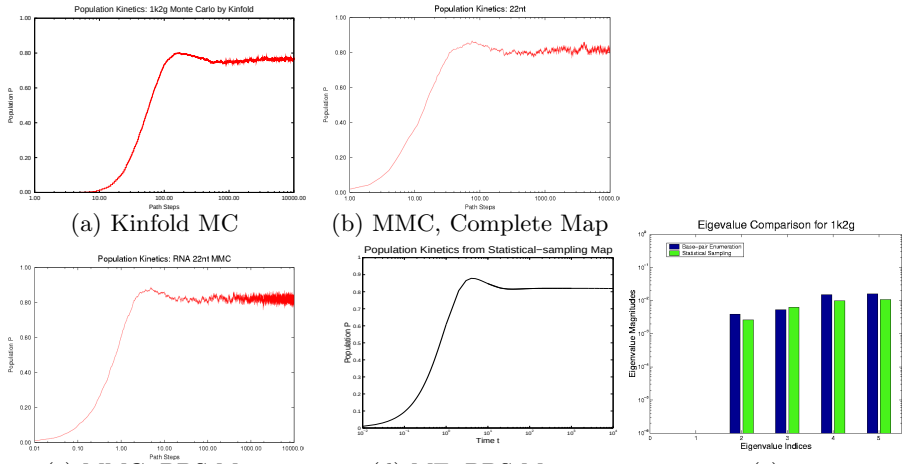
We demonstrate with two different RNA that the different analysis methods (ME, MC, MMC) produce comparable results and can be used interchangeably. This is important since some methods like the master equation do not scale as well as others like Map-based Monte Carlo simulation with RNA size.

Comparison with other Simulation methods. Here we present the results of 1k2g (CAGACUUCGGUCGCAGAGAUGG), a 22 nucleotide RNA. Figure 1 compares the population kinetics of the native state using (a) standard Monte Carlo simulation (implemented by Kinfold [8]), (b) Map-based Monte Carlo simulation on a fully enumerated roadmap (12,137 conformations), (c) Map-based Monte Carlo simulation on a roadmap with our new PBS method (42 conformations), and (d) the master equation on a PBS roadmap (42 conformations). The fully enumerated roadmap is the most accurate model. However, it is not feasible to enumerate RNA with more than 40 nucleotides. The statistical-sampling roadmaps yields much smaller subsets of the entire conformation space that effectively approximate the energy landscape. Also note that we can only use the master equation on small roadmaps (e.g., up to 10,000 conformations) due to numerical limitations in computing the eigenvalues and eigenvectors.

All population kinetics curves have similar features (see Figure 1). In each figure, the population first increases quickly, then it gradually decreases and eventually stabilizes to the equlibrium distribution. Note that the equilibrium (final) distributions are very close to each other at 80%, even though the PBS roadmap (c) and (d) contains less than 0.4% of all possible conformations. Thus, these roadmaps capture the main features of the energy landscape. This data indicates that these analysis methods are interchangeable.

Figure 1(e) compares the four smallest eigenvalues of the fully enumerated (base-pair) roadmap and the statistical-sampling roadmap computed by the master equation. All the eigenvalues, i.e., folding rates, are similar. This indicates that our extremely sparse roadmaps not only capture the major features of the equilibrium distribution, but also capture the major features of the kinetics.

Scalability of the Approximated Roadmaps. Here we compare our simulation results on a larger 56 nucleotide RNA. Leptomonas Collosoma Spliced Leader RNA is known to have many metastable structures [6]. This RNA has



(a) Kinfold MC (b) MMC, Complete Map

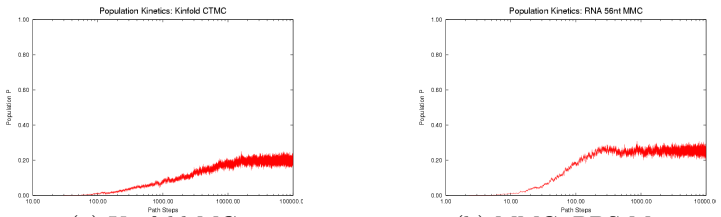
(c) MMC, PBS Map (d) ME, PBS Map (e)

Fig. 1. The population kinetics of the native state of 1k2g: (a) Kinfold Monte Carlo simulation, (b) Our MMC simulation on a fully enumerated roadmap (12,137 conformations), (c) Our MMC simulation on a PBS roadmap (42 conformations), and (d) Master equation solution on the PBS roadmap (42 conformations). All analysis techniques produce similar population kinetics curves and similar equilibrium distribution. (e) Comparison of the eigenvalues of 1k2g by the master equation on a fully enumerated roadmap (12,137 conformations) and new PBS roadmap (42 conformations). Both eigenvalues are similar between the different roadmaps.

approximately $2.0 * 10^{14}$ conformations, so it is not feasible to enumerate even the stack-pair conformations, let alone the entire conformation space. Thus, we are only able to compare kinetics from the Kinfold Monte Carlo simulation and our Map-based Monte Carlo simulation using PBS roadmaps. For each simulation technique, we compute 1000 different folding pathways. We combine these pathways to calculate the population kinetics of a particular conformation.

Figure 2 shows that although we only use 5033 conformations in the roadmap, our Map-based Monte Carlo simulation (b) results have qualitatively similar features with the Kinfold Monte Carlo simulation (a). First notice that the equilibrium distribution of both simulations are similar to each other at 20%. The one for our Map-based Monte Carlo simulation is slightly higher because it represents the entire conformation space with only a few conformations. However, considering that it uses such a tiny subset $(5.0 * 10^3)$ to represent a huge space $2.0 * 10^{14}$, this is pretty good approximation. Also we notice that both curves have three stages and share some similar features. In the first stage, the population stays at zero for quite a long time, then it gradually increases monotoically in the second stage until it reaches the third stage – equilibrium. In contrast, such features are very different to other RNA such as population kinetics of 1k2g shown in Figure 1. Again this gives strong evidence that our sparse roadmap can capture the main feature of the energy landscape. Finally, our Map-based Monte Carlo

simulation requires fewer iterations to stabilize (an order of magnitude fewer) and uses less space (1G versus 8G for Kinfold).



(a) Kinfold MC (b) MMC, PBS Map

Fig. 2. Comparison of population kinetics of a metastable state for Leptomonas Collosoma Spliced Leader RNA using (a) Kinfold Monte Carlo simulation and (b) our MMC simulation on a PBS roadmap with 5033 conformations. We are able to capture the same kinetics while only sampling a tiny fraction of the entire conformation space.

5.2 Experimental Validation: Kinetics Related Functions

Many RNA can perform a variety of functions such as regulating the gene expression rate or plasmid replication rate. It has been found that some functions are not only determined by their native states but by the metastable states formed during the folding process, where the functional units are active [9, 15, 11, 18]. Thus these functions are based on the RNA's folding *kinetics*. These functions are studied experimentally by comparing the kinetics and functional rates of different mutants that share the same thermodynamic stability and native structure. Below we give two case studies that show how we can also study these kinetics-based functions and compare to experimental data.

ColE1 RNAII: Predict Plasmid Replication Rates. ColE1 RNAII regulates the replication of E. coli ColE1 plasmids through its folding kinetics [10, 15]. The slower it folds, the higher the plasmid replication rate. A specific mutant, MM7, differs from the wild-type (WT) by a single nucleotide out of the 200 nucleotide sequence. This mutation causes it to fold slower while maintaining the same thermodynamics of the native state. Thus, the overall plasmid replication rate increases in the presence of MM7 over the WT.

We can study this difference computationally by computing the folding rates of both WT and MM7 using the master equation and comparing their eigenvalues. [10] performs a similar study. However, they solve the master equation on a much more simplified energy landscape using a specific subsequence (130 / 200 nucleotides) and 9 stems hand-picked from 30 conformations. In contrast, we simulate the kinetics of the entire sequence using around 4000 conformations.

Figure 3 shows the eigenvalues calculated using the master equation. Note that the smallest non-zero eigenvalues correspond to the folding rate. All eigenvalues of WT are larger than MM7 indicating that WT folds faster than MM7. Thus, our method correctly estimated the functional level of the new mutant.

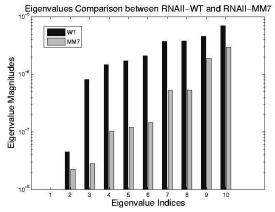


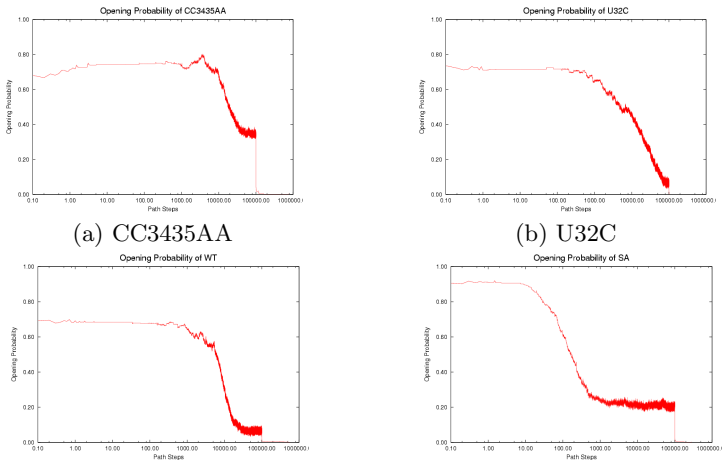
Fig. 3. Comparison of the 10 smallest non-zero eigenvalues (i.e., the folding rates) for WT and MM7 of ColE1 RNAII as computed by the master equation. The overall folding rate of WT is faster than MM7 matching experimental data.

MS2 phage RNA: Predict Protein Expression Rate. MS2 phage RNA (135 nucleotides) regulates the expression rate of phage MS2 maturation protein [9, 15] at the translational level. It works as a regulator only when a specific subsequence (the SD sequence) is open (i.e., does not form base-pair contacts). Since this SD sequence is closed in the native state, this RNA can only perform this function before the folding process finishes. Thus its function is based on its folding *kinetics* and not the final native structure. Three mutants have been studied that have similar thermodynamic properties with the wild-type (WT) but different kinetics and therefore different gene expression rates. Experimental results indicate that mutant CC3435AA has the highest gene expression rate, WT and mutant U32C are similar, and mutant SA has the lowest rate [9, 15].

Intuitively, the functional rate (e.g., gene expression rate in this case) is correlated with the opening of the SD sequence. If the SD sequence is opened longer, or has higher opening probability (i.e., having more nucleotides on the SD sequence open), then the mutant should have higher functional rate. We use our simulation method to study this opening probability during the folding process. In our study, we first simulate the folding process for each mutant by generating 1000 folding pathways for each mutant using Map-based Monte Carlo simulation. Then we analyze the pathways for each mutant and calculate the opening probability of the SD sequence. We calculate the opening probability as the percentage of open nucleotides in the SD sequence. In [11], Higgs performed a similar study using a stem-based Monte-Carlo simulation. However, in that work, they simulated the folding process only when the RNA sequence is growing. Their results may depend on the selection of growth rate. If the growth rate was too high or too low, the results may or may not be able to compare to experiment. Our simulation results, on the other hand, do not require this growth rate parameter and thus can be used to quantitatively predict the functional level of a new mutant in a more reliable way.

Figure 4 shows the time evolution of the SD opening probability for the WT and the three mutants. Note that CC3435AA has the longest duration at a relatively high level of opening probability while SA has the shortest duration. This correlates with experimental data. The opening probability of U32C decreases

earlier but finishes later than WT, so it is not clear which one has a larger total opening probability during folding, again matching experimental findings.



(a) CC3435AA (b) U32C

(c) WT (d) SA

Fig. 4. Comparison of the SD opening probability during the folding process.

The gene expression rate is determined from two factors: (1) how high the opening probability is at any given time and (2) how long the RNA stays in the high opening probability state. To compare each RNA quantitatively, we compute the integration of the opening probability (Figure 4) over the whole folding process. Note that the RNA regulates gene expression only when the SD opening probability is "high enough". We used thresholds ranging from 0.2 to 0.6 to estimate the gene expression rate. Thresholds higher than 0.6 will yield zero opening probability on WT and most mutants and thus cannot be correlated to experimental results. Similarly, we don't consider thresholds lower than 0.2. Otherwise this means that mutant SA can be active even in equilibrium condition, which does not correspond to experimental results. Table 1 shows the results for the WT and for each mutant. For most thresholds, mutant CC3435AA has the highest rate and mutant SA has the lowest rate, the same relative functional rate as seen in experiment. In addition WT and mutant U32C have similar levels (particularly between 0.4-0.6), again correlating with experimental results. Aside from simply validating our method against experiment, we can also use our method to suggest that the SD sequence may only be active for gene regulation when more than 40% of its nucleotides are open.

6 Conclusion

We have proposed new sampling techniques and a new analysis tool called Map-based Monte Carlo (MMC) simulation that can be used to study kinetics-based

Mutant	Experimental Expression Rate (order of magnitude)	Our Estimation				
		$t = 0.2$	$t = 0.3$	$t = 0.4$	$t = 0.5$	$t = 0.6$
SA	0.1	0.1	0.04	0.03	0.03	0.08
WT	1	1.0	1.0	1.0	1.0	1.0
U32C	1	2.1	1.8	1.4	0.8	1.2
CC3435AA	5	7.2	8.4	3.8	3.5	9.8

Table 1. Comparison of expression rates between WT and three mutants of MS2. It shows that we can predict similar relative functional rates as seen in experiments.

functions for RNA such as population kinetics, folding rates, and the folding of particular subsequences. These new tools enable us to study larger RNA than before – increasing from RNA with tens of nucleotides (e.g., 40) to those with hundreds of nucleotides (e.g., 200+).

We validated our method against known experimental data and analyzed two case studies in detail. For the first, we showed that our method identified the same relative folding rates as those noted in experiment for ColE1 RNAII and its mutant. In the second case study, we showed that our approach predicted the same relative gene expression rates of wild-type MS2 phage RNA and three of its mutants. We believe that our method will be a valuable tool for discovering such relationships for other RNA that have not been characterized experimentally. Although we only study secondary structure now, in the future, we expect to include pseudo knots and tertiary structures using an appropriate energy model.

7 Acknowledgments

We thank Dr. David Giedroc for many insightful discussions.

References

1. N. M. Amato, K. A. Dill, and G. Song. Using motion planning to map protein folding landscapes and analyze folding kinetics of known native structures. *J. Comput. Biol.*, 10(3-4):239–256, 2003. Special issue of Int. Conf. Comput. Molecular Biology (RECOMB) 2002.
2. N. M. Amato and G. Song. Using motion planning to study protein folding pathways. *J. Comput. Biol.*, 9(2):149–168, 2002. Special issue of Int. Conf. Comput. Molecular Biology (RECOMB) 2001.
3. D. Bartel. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116:281–297, 2004.
4. S.-J. Chen and K. A. Dill. RNA folding energy landscapes. *Proc. Natl. Acad. Sci. USA*, 97:646–651, 2000.
5. K. A. Dill and H. S. Chan. From Levinthal to pathways to funnels: The new view of protein folding kinetics. *Nat. Struct. Biol.*, 4:10–19, 1997.
6. Y. Ding and C. E. Lawrence. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Research*, 31:7280–7301, 2003.

7. R. Dirks and N. Pierce. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *Journal of Computational Chemistry*, 24:1664–1677, 2003.
8. C. Flamm. *Kinetic Folding of RNA*. PhD thesis, University of Vienna, Austria, August 1998.
9. H. Groeneveld, K. Thimon, and J. van Duin. Translational control of matruation-protein synthesis is phage MS2: a role of the kinetics of RNA folding? *RNA*, 1:79–88, 1995.
10. A. Gultyaev, F. V. Batenburg, and C. Pleij. The computer simulation of RNA folding pathways using a genetic algorithm. *J. Mol. Biol.*, 250:37–51, 1995.
11. P. G. Higgs. RNA secondary structure: physical and computational aspects. *Quarterly Reviews of Biophysics*, 33:199–253, 2000.
12. I. L. Hofacker. RNA secondary structures: A tractable model of biopolymer folding. *J. Theor. Biol.*, 212:35–46, 1998.
13. N. G. V. Kampen. *Stochastic Processes in Physics and Chemistry*. North-Holland, New York, 1992.
14. L. E. Kavraki, P. Svestka, J. C. Latombe, and M. H. Overmars. Probabilistic roadmaps for path planning in high-dimensional configuration spaces. *IEEE Trans. Robot. Automat.*, 12(4):566–580, August 1996.
15. P. Klaff, D. Riesner, and G. Steger. RNA structure and the regulation of gene expression. *Plant Mol. Biol.*, 32:89–106, 1996.
16. C. Ma, T. Kolesnikow, J. Rayner, E. Simons, H. Yim, and R. Simons. Control of translation by mRNA secondary structure: the importance of the kinetics of structure formation. *Mol. Microbiol.*, 14:1033–1047, 1994.
17. J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29:1105–1119, 1990.
18. J. H. Nagel and C. W. Pleij. Self-induced structural switches in RNA. *Biochimie*, 84:913–923, 2002.
19. M. E. J. Newman and G. T. Barkenma. *Monte Carlo Methods in Statistical Physics*. Clarendon Press, Oxford, 1999.
20. R. Nussinov, G. Piecznik, J. R. Griggs, and D. J. Kleitman. Algorithms for loop matching. *SIAM J. Appl. Math.*, 35:68–82, 1972.
21. S. B. Ozkan, K. A. Dill, and I. Bahar. Computing the transition state population in simple protein models. *Biopolymers*, 68:35–46, 2003.
22. E. Rivas and S. Eddy. A dynamic programming algorithm for rna structure prediction including pseduoknots. *JMB*, 285:2053–2068, 2000.
23. B. A. Shapiro, D. Bengali, W. Kasprzak, and J. C. Wu. RNA folding pathway functional intermediates: Their prediction and analysis. *J. Mol. Biol.*, 312:27–44, 2001.
24. G. Song, S. Thomas, K. Dill, J. Scholtz, and N. Amato. A path planning-based study of protein folding with a case study of hairpin formation in protein G and L. In *Proc. Pacific Symposium of Biocomputing (PSB)*, pages 240–251, 2003.
25. X. Tang, B. Kirkpatrick, S. Thomas, G. Song, and N. M. Amato. Using motion planning to study RNA folding kinetics. *J. Comput. Biol.*, 12(6):862–881, 2005. Special issue of Int. Conf. Comput. Molecular Biology (RECOMB) 2004.
26. S. Thomas, X. Tang, L. Tapia, and N. M. Amato. Simulating protein motions with rigidity analysis. In *Proc. Int. Conf. Comput. Molecular Biology (RECOMB)*, pages 394–409, 2006.
27. I. Tinoco and C. Bustamante. How RNA folds. *J. Mol. Biol.*, 293:271–281, 1999.

28. A. E. Walter, D. H. Turner, J. Kim, M. H. Lyttle, P. Muller, D. H. Mathews, and M. Zuker. Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proc. Natl. Acad. Sci. USA*, 91:9218–9222, 1994.
29. M. Wolfinger. The energy landscape of RNA folding. Master's thesis, University of Vienna, Austria, March 2001.
30. S. Wuchty. Suboptimal secondary structures of RNA. Master's thesis, University of Vienna, Austria, March 1998.
31. A. Xayaphoummine, T. Bucher, F. Thalmann, and H. Isambert. Prediction and statistics of pseudoknots in RNA structures using exactly clustered stochastic simulations. *Proc. Natl. Acad. Sci. USA*, 100:15310–15315, 2003.
32. M. Zuker, D. H. Mathews, and D. H. Turner. Algorithms and thermodynamics for RNA secondary structure prediction: A practical guide. In J. Barciszewski and B. F. C. Clark, editors, *RNA Biochemistry and Biotechnology*, NATO ASI Series. Kluwer Academic Publishers, 1999.