Using Motion Planning to Study RNA Folding Kinetics

Xinyu Tang[†] Bonnie Kirkpatrick[§] Shawna Thomas[†] Guang Song[‡] Nancy M. Amato[†]

ABSTRACT

We propose a novel, motion planning based approach to approximately map the energy landscape of an RNA molecule. Our method is based on the successful probabilistic roadmap motion planners that we have previously successfully applied to protein folding. The key advantage of our method is that it provides a sparse map that captures the main features of the landscape and which can be analyzed to compute folding kinetics. In this paper, we provide evidence that this approach is also well suited to RNA. We compute population kinetics and transition rates on our roadmaps using the master equation for a few moderately sized RNA and show that our results compare favorably with results of other existing methods.

Categories and Subject Descriptors: J.3 [**Life and Medical Sciences**]: Biology and genetics

General Terms: Algorithms, Experimentation

Keywords: RNA, Folding Kinetics, Motion Planning

1. INTRODUCTION

Ribonucleic acid (RNA) molecules perform diverse and important functions such as synthesizing proteins, catalyzing reactions, splicing introns, and regulating cellular activities [21]. An RNA's nucleotide sequence, and the three-dimensional structure of its energetically stable conformations determines how the RNA functions and interacts with its environment. The process by which an RNA molecule (re)configures itself into an energetically stable conformation is called *folding*.

There are two general, but related, issues for RNA folding: structure prediction and folding kinetics. The structure prediction problem is to predict the structure of the native conformation given the RNA's nucleotide sequence. Unlike the related protein folding problem, efficient algorithms do exist for some forms of RNA structure prediction [22, 26]. However, they do not provide insight into the folding process or the "energy landscape" which determines the folding kinetics.

Each RNA conformation is associated with an energy; the lower its energy, the more stable it is. The energy landscape can be thought of as adding this energy as another dimension to the other parameters specifying the conformation. As will described in detail later, the energy landscape encodes information about folding pathways, transition rates, intermediate states, and population kinetics. The size of the landscape grows exponentially with the sequence length, so it is infeasible to compute the complete landscape. To reduce this complexity, researchers focus on RNA planar secondary structures instead of three-dimensional conformations. Although this dramatically reduces the size of the landscape, it remains impractical to compute the complete landscape for sequences longer than about 40 nucleotides [10].

There are at least three important reasons to study RNA energy landscapes and folding kinetics. First, a better understanding of the folding process will aid the development of more efficient structure prediction algorithms. Second, it has recently been discovered that catalytic RNA often fluctuate away from their native conformation to interact with other RNA, proteins, and ligands [21], and we cannot model or predict these fluctuations without studying energy landscapes. Third, we must study energy landscapes and folding kinetics to understand how and why RNA molecules misfold.

In this paper, we propose a novel, motion planning based approach to approximately map the RNA's energy landscape. In particular, we develop a *probabilistic roadmap* (PRM) [12] based approach that first samples RNA configurations and then connects them together to form a graph, or *roadmap*. The key advantage of our method is that it provides a sparse representation of the landscape that captures its main features and which can be analyzed to compute folding kinetics. We have previously applied this strategy to protein folding with considerable success [1, 2, 19, 20], e.g., our method predicted the subtle folding differences between the structurally similar proteins G and L [20]. In this paper, we provide evidence that this approach is also well suited to RNA. In particular, we present results such as population kinetics and transition rates obtained using the master equation (Section 4.1) for a few moderately sized RNA and show that our results match results obtained with other existing methods quite well [25].

[†] Parasol Lab, Dept. of Computer Science, Texas A&M Univ., College Station, TX 77843-3112. Email: {xinyut,sthomas,amato}@cs.tamu.edu. Research supported in part by NSF Grants ACI-9872126, EIA-9975018, EIA-0103742, EIA-9805823, ACR-0081510, ACR-0113971, CCR-0113974, EIA-9810937, EIA-0079874. Thomas supported in part by an NSF Graduate Research Fellowship.

[‡] Baker Center for Bioinformatics and Biological Statistics, Iowa State Univ., Ames, IA 50011. Email: gsong@iastate.edu. Work performed at the Parasol Lab and supported in part by an IBM TJ Watson PhD Fellowship.

[§] Dept. Computer Science, Montana State Univ., Bozeman, MT 59717. Email: bkirk@cns.montana.edu. Work performed at the Parasol Lab during research internships in summer 2002 and 2003 supported by the CRA Distributed Mentor Project.

RECOMB'04, March 27–31, 2004, San Diego, California, USA.

2. PRELIMINARIES & RELATED WORK

2.1 RNA Primer

An RNA molecule is a sequence of nucleotides which differs from other RNA molecules in its 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 [22, 27, 9], and are also what we consider in our model.

RNA Structure. *Tertiary structure* is a 3D spatial RNA conformation with a set of base pairs. *Secondary structure* is a planar representation of an RNA conformation. Although there are slightly differing definitions [4, 9], secondary structure is usually considered to be a planar subset of the base pair contacts present (see Table 1, case 3). Non-planar contacts, often called *pseudo knots*, are usually considered tertiary interactions and not allowed in secondary structure. Many definitions of secondary structure, including the one we adopt, eliminate other types of contacts that are not physically favored. Contacts considered invalid in our secondary structure are defined in Table 1; this definition is also used in [9]. Three common representations for RNA secondary structure are shown in Figure 1 [27].

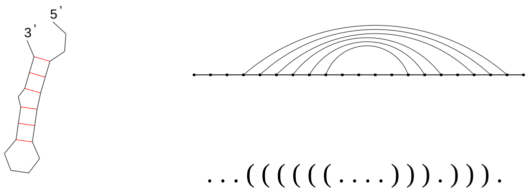


Figure 1: Three representations of the same secondary structure for the sequence GGCGUAAGGAUUACCUAUGCC which denote contact pairs with bonds, arcs, and pairs of brackets, respectively.

The tertiary structure gives the most complete representation of RNA structure. However, the secondary structure is commonly used [26, 27, 9] because in many cases it provides sufficient information to study many aspects of folding while dramatically reducing the size of the RNA conformation space that must be explored. One justification for this simplification is that research has shown that the RNA folding process is hierarchical, i.e., secondary structure forms before tertiary structure [21, 27]. In this work, we focus on the first stage, the formation of secondary structure.

Energy Calculations. To represent the RNA folding energy landscape, we must be able to calculate the energy of any conformation. One commonly used energy function is the Turner or nearest neighbor rules [26]. 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, more contacts, especially adjacent contacts, typically yield more stable structures with lower energy. Much work has been done to make these rules more detailed and accurate.

RNA (Secondary Structure) Conformation Space. For a given RNA nucleotide sequence, an RNA (secondary structure) conformation is a planar set of valid base pairs.[1] The secondary structure conformation space, $\mathcal{C}$, of an RNA sequence contains all sets of base pairs that meet the criteria in Table 1. The size of $\mathcal{C}$ grows exponentially as sequence length increases [27, 5]. Knowledge of the size of $\mathcal{C}$ is used to determine the feasibility of enumerating all conformations, or if some sampling will be needed. The size of $\mathcal{C}$ depends not only on the RNA sequence length but also on the sequence itself. Since exact computation of $|\mathcal{C}|$ requires enumerating $\mathcal{C}$, it should be estimated.

Zuker and Sankoff [27] developed a close estimation of $|\mathcal{C}|$ using a stochastic approach to account for the effect of the specific sequence. Given an RNA sequence of length n, they calculate the probabilities p_A, p_C, p_G, and p_U of the occurrence of each nucleotide, i.e., the percentage of that nucleotide in the sequence. They then use $p = 2(p_A p_U + p_C p_G)$ as the probability of two bases making a contact and obtain the approximation $|\mathcal{C}| \approx hn^{\frac{3}{2}}\alpha^n$, where $\alpha = (\frac{1+\sqrt{1+4\sqrt{p}}}{2})^2$ and $h = \frac{\alpha(1+4\sqrt{p})^{1/4}}{2\sqrt{\pi}p^{3/4}}$.

Unfortunately, however, the Zuker and Sankoff estimate doesn't fit our model because they do not consider the wobble pair G-U or the restriction of the minimal hairpin size to 5. We modified this formula to fit our model by including the wobble pair in the probability $p' = 2(p_A p_U + p_C p_G + p_U p_G)$, and then scaling the probability p' to $p = p' \cdot (n-3)(n-4)/n^2$ to restrict the minimal hairpin size to 5. Our revised estimate results from substituting the new p in the equations for α and h.

As can be seen in Table 2, our estimate can be a significantly better estimate of $|\mathcal{C}|$ for our model than the estimate used in [27]. Our exact enumeration results match Cupal [9]. It can also be seen that $|\mathcal{C}|$ grows exponentially with sequence length, and hence it becomes impractical to enumerate all conformations when the sequence length exceeds 40 nucleotides [10] and thus some type of sampling must be used instead.

2.2 PRMs and Protein Folding

Our approach to RNA folding is based on the *probabilistic roadmap* (PRM) technique for motion planning [12]. Motion planning determines valid paths to move objects from one configuration to another. PRMs build graphs (roadmaps) that ideally approximate the topology of the feasible planning space, and can be used to answer many, varied queries quickly. Briefly, PRMs work by sampling points from the movable object's conformation space (C-space) and retaining those that satisfy feasibility requirements (e.g., collision-free). The movable object's C-space is the set of all positions and orientations of the movable object, feasible or not [13]. Next, the retained points are connected to form a graph, or roadmap, using some simple local planning method (e.g., a straight line) to connect nearby points. During query processing, paths connecting the start and goal conformations are extracted from the roadmap using standard graph search techniques (see Figure 2).

In previous work, we have used PRMs to study protein folding when the native structure is known [19, 2, 1, 20]. Here, the moving object is the protein, and the main difference from the usual PRM application is that the collision-

[1] As we only consider secondary structure in our method, we will usually omit this qualification when referring to conformations and conformation space.

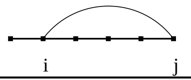
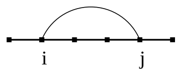
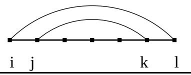
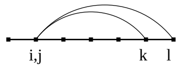
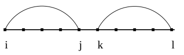
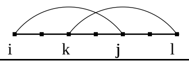
Description	Valid Contact	Invalid Contact
Case 1: (Separation) Bases of each pair must be separated by at least 3 other residues, i.e., $\|i-j\| > 3$		
Case 2: (Multiplicity) Each base can be paired to only one other, i.e., $i = k$ if and only if $j = l$		
Case 3: (Planarity) The contacts must be planar (no pseudoknots), i.e., if $i < k < j$, then $i < k < l < j$		

Table 1: Definition of valid secondary structure for any two contacts $[i,\, j]$ and $[k,\, l]$ with $i < j$ and $k < l$.

Sequence	# nucl	Exact $\|\mathcal{C}\|$	Estimation	
			Zuker [27]	Ours
(ACGU)2	8	5	22	6
(ACGU)3	12	35	206	47
ACUGAUCGUAGUCAC	15	1.4×10^2	1.0×10^3	2.4×10^2
GGCGUAAGGAUUACCUAUGCC	21	8.6×10^3	6.2×10^5	1.3×10^4
(ACGU)10	40	1.7×10^8	1.6×10^{10}	3.3×10^9

Table 2: Estimated and actual sizes of C-space for several RNA sequences.

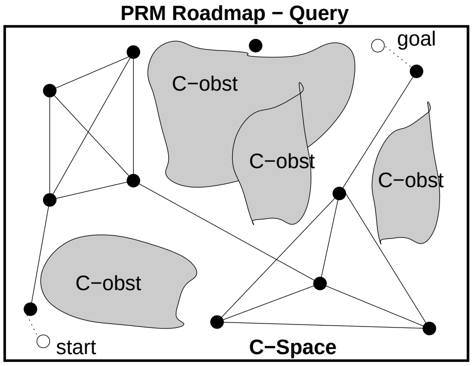


Figure 2: A PRM roadmap in C-space and a query.

detection feasibility test is replaced by a preference for low energy conformations. We have obtained promising results that were validated with experimental data for several moderate sized proteins, e.g., we were able to observe the subtle folding differences in the structurally similar proteins G and L [20].

2.3 Related Work

Research on RNA folding falls into two categories: structure prediction and the study of folding kinetics. 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 [15]. Zuker and Stiegler formulated an algorithm to address the minimum energy problem. Today, Zuker's MFOLD algorithm is widely used for structure prediction [22]. McCaskill's algorithm [14] uses dynamic programming to calculate the partition function $Q = \sum_s exp(-\Delta G(s)/kT)$ over all secondary structures s, while Chen [4] uses matrices for approximation. As described in [4], the partition function is "the sum of Boltzmann factors over all possible branching patterns in which the chain can be arranged into helices and intervening regions". As we will see, the partition function can also be used to study folding kinetics. Wuchty extended this algorithm to compute the density of states at a predefined energy resolution [24]. The ViennaRNA package is based on this work, and implements Zuker and McCaskill's algorithms as well as some energy functions [9].

Several approaches other than thermodynamics have been used to investigate RNA kinetics, e.g., Gillespie [8] used a Monte Carlo algorithm to find folding pathways, and Shapiro et al.[3] used a genetic algorithm to study RNA folding pathways.

Several methods have been proposed that involve computations on the folding landscape. Dill [4] used matrices to compute the partition function over all possible secondary structures. Complete folding landscapes are approximated by this method. Wuchty modified Zuker's algorithm [24] to generate all the secondary structures within some given energy range of the native structure. Flammy and Wolfinger [7, 23] extended this algorithm to find local minima within some energy difference of the native state, then to connect them via energy barriers. The uses energy barrier tree to represent the energy landscape.

The master equation can be used to compute the population kinetics of the folding landscape. It uses a matrix of differential equations to represent the probability of transition between different conformations. Once solved, the dominate modes of the solution describe the general folding kinetics [16, 17, 11, 25].

3. RNA FOLDING WITH PRMS

In this section, we discuss how to apply PRMs to study RNA folding. There are two main steps in our approach: constructing the roadmap and analyzing it. Constructing the roadmap requires sampling a set of RNA conformations and computing their energies. Next, we determine which conformations we should attempt to connect using a "local planner", i.e., a simple method to find transitions between

RNA conformations. One difference from the protein folding application is that our C-space is not continuous but discrete, and hence our options for making the local connections are more restricted. The local planner also assigns weights to the transitions to reflect their energetic feasibility. This results in a roadmap (graph) of conformations (nodes) connected by transitions (edges) that represents the energy landscape and where each pathway is a sequence of conformational changes the RNA molecule goes through as it transforms from one conformation to another.

After the roadmap is built, we perform some analysis on it to study the population kinetics and provide insight into the folding process. We can identify transitional conformations where the folding process could be trapped or delayed, the folding rate, and representative folding pathways. In this work, we analyze the landscape via folding pathways and the master equation.

Energy computations are required to measure conformation feasibility and to calculate the roadmap edge weights (as discussed in Section *3.1.2*). Our current implementation uses a third-party energy function relying on the Turner rules to determine the validity of a point in C-space. This energy function is part of the ViennaRNA package [9].

3.1 Roadmap Construction

The goal of roadmap construction is to build an approximation of the energy landscape that captures its important features. The quality of our approximation depends on our node sampling and connection methods.

3.1.1 Node Generation

Our framework currently has three methods for generating RNA conformations: complete base-pair enumeration (for small RNA), stack-pair enumeration and maximal-contact sampling.

Complete Base-Pair Enumeration (BPE). Our discrete RNA C-space makes it possible to enumerate all the conformations for small RNA molecules. However, this is not feasible for molecules with more than around 40 nucleotides [10]. Let $\mathcal{S}$ be the set of all possible base-pair contacts. To generate a valid conformation, we first select one contact in $\mathcal{S}$. Then we remove all contacts from $\mathcal{S}$ that would violate the criteria in Table 1 if combined with the already selected contacts. The process of iteratively selecting a valid contact from $\mathcal{S}$ and then removing invalid contacts from $\mathcal{S}$ continues until $\mathcal{S}$ is empty. Each time we select a new contact, we define a new secondary structure. To enumerate the entire space, we enumerate all possible combinations of a valid set of contacts from $\mathcal{S}$ as above. Figure 3 shows the complete enumeration for the RNA sequence ACGUCACGU.

Stack-Pair Enumeration (SPE). This enumeration contains only those conformations containing stack-pair contacts. A *stack-pair contact* is a set of adjacent base-pair contacts, i.e., no contacts are isolated from the others. More formally, if a stack-pair contact has a contact $[i,\ j]$, where $i < j$, then it must also have at least one of the contacts $[i-1, j+1]$ or $[i+1, j-1]$. For example, the contacts in Figure 3(c) form a stack, but the contacts in Figure 3(f) do not because they are not adjacent. A conformation is a valid *stack-pair conformation* if it only has stack-pair contacts, i.e., if there are no isolated base pairs. The conformations in Figure 3 (a), (c), (d), (h), and (j) are the enumeration of stack-pair conformations for RNA sequence ACGUCACGU. We favor these conformations because isolated base pairs are unstable. This simplification has been used in [25]. We can study larger RNA molecules with this method than is possible with complete enumeration because we can enumerate all stack-pair conformations without enumerating all conformations. The stack-pair enumeration is implemented similarly to the base-pair enumeration except that $\mathcal{S}$ contains stacks instead of base-contact pairs.

Maximal-Contact Sampling (MCS). In this method, nodes are generated in a more 'random' fashion. To get lower energy conformations, we only generate conformations with maximal contacts, i.e., no more contacts can be added to those conformations without causing a violation (Table 1). First, we create a conformation c without any contacts. Then, single contacts are successively added until it is not possible to add a contact and maintain a valid conformation. This method biases the node distribution toward the areas of C-space with more contacts. Since more contacts usually means more stability for the conformation, the energy of these conformations is usually lower. Each time a contact is added, it is randomly selected from all currently feasible contacts, and the set of feasible contacts is updated. This continues until no more contacts can possibly be added. In Figure 3, (a), (d), (e), (g) and (h) are the maximal-contact conformations.

3.1.2 Node Connection

After node generation, it would be expensive, and generally not necessary, to make all $\theta(n^2)$ connections. Here, we restrict our attention to connecting nearby conformations. This requires distance metrics to identify nearby conformations for connection and techniques for connecting them.

Distance Metrics. The distance metric defines which nodes are close to each other and which are far apart. Here we use base-pair distance (the number of contact pairs that differ between two conformations). This denotes the number of base pairs that have to be opened or closed to transform one conformation into another. Our approach can utilize other distance metrics such as string edit distance or tree edit distance [18], but we found that base-pair distances perform the best on the RNA we have studied.

Identifying Nodes for Connection. Neighboring roadmap nodes are connected using a local planner. We use two different strategies for determining neighbors. One strategy attempts to connect a node with the k closest nodes and the other attempts to connect a node with all nodes within a fixed radius r.

Generating Transitional Conformations. Once the neighbors are determined, the local planner connects each pair of nodes by generating transitions between them. To generate a transition from conformation c_1 to conformation c_2, we first identify the set $\mathcal{O}$ of contacts to be opened (i.e., contacts in c_1 and not in c_2) and the set $\mathcal{L}$ of contacts to be closed (i.e., contacts in c_2 but not in c_1). See Figure 4 (a): contacts q_1 and q_2 are in $\mathcal{O}$ and contacts p_1 and p_2 are in $\mathcal{L}$. To ensure that transitional conformations do not violate our planarity constraint, we construct a *conflict graph* G between $\mathcal{O}$ and $\mathcal{L}$. G describes which contact pairs cannot exist together in a valid conformation. If one contact $p \in \mathcal{L}$ conflicts with another contact $q \in \mathcal{O}$, then p cannot be closed until q is opened, and we have an edge from q to p in G. See Figure 4(b): q_1 and q_2 conflict with p_1; q_2 conflicts with p_2.

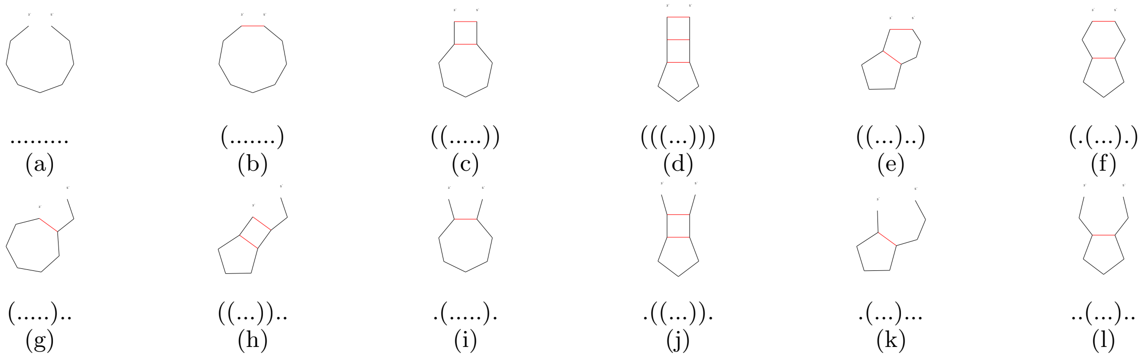


Figure 3: Complete enumeration of all conformations for RNA sequence ACGUCACGU. Conformations (a), (c), (d), (h) and (j) are the stack-pair conformations. Conformations (a), (d), (e), (g) and (h) are maximal-contact conformations.

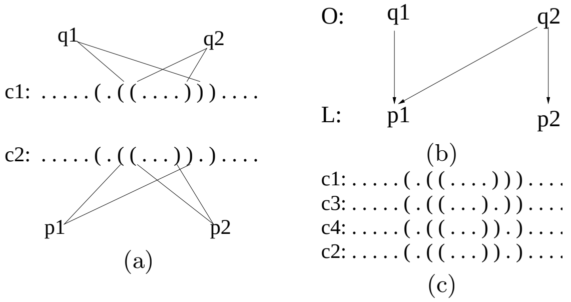


Figure 4: Transitional node generation. (a) Start and goal conformations and contact pairs to be opened and closed: q_1, q_2 are in $\mathcal{O}$; p_1, p_2 are in $\mathcal{L}$. (b) Conflict graph: q_1 and q_2 conflict with p_1, q_2 conflicts with p_2. (c) Sequences generated: First open q_2 and close p_2, then open q_1 and close p_1. c_3 and c_4 are the two transitional conformations to connect c_1 and c_2, here c_4 happens to be identical to c_2.

A valid transition is a sequence of transitional conformations that doesn't violate G.

Our framework can use any strategy to determine the order to open contacts in $\mathcal{O}$ and close contacts in $\mathcal{L}$. The most naive method is to first open all the contacts in $\mathcal{O}$ and then to close all the contacts in $\mathcal{L}$. This does not violate G, but it produces high energy transitional conformations. To find low energy transitions, we want to produce conformations with as many contacts as possible since they usually have lower energy. So, once we open a contact, we close all contacts in $\mathcal{L}$ that do not violate G. We use a greedy strategy to determine the order for opening the contacts. In particular, we sort the contacts in $\mathcal{L}$ according to the number of contacts in $\mathcal{O}$ they conflict with (given by their indegree in L). We select the contact in $\mathcal{L}$ with the smallest number of conflicts and open all the contacts in $\mathcal{O}$ that conflict with it. We then close all the contacts in $\mathcal{L}$ that have no conflicts. See Figure 4(c): c_3, c_4 are the two transitional conformations generated for the connection. This is repeated until both $\mathcal{O}$ and $\mathcal{L}$ are empty. This strategy works well for the RNA we have studied.

Edge Weights. Edge weights are assigned to reflect the transition rate between neighboring conformations, i.e., the probability the molecule folds from one conformation to the other. Thus, edge weights reflect the energetic feasibility for the folding process on this edge.

When an edge (q_1, q_2) is added to the roadmap, it is assigned a weight that depends on the sequence of transitional conformations $\{q_1 = c_0, c_1, c_2, \ldots, c_{n-1}, c_n = q_2\}$ determined by the local planner. For each pair of consecutive conformations c_i and c_{i+1}, the probability P_i of moving from c_i to c_{i+1} is

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

where $\Delta E_i = E(c_{i+1}) - E(c_i)$, k is the Boltzmann constant, and T is the temperature of folding. For a detailed discussion of different rules to calculate the transition probabilities, please refer to [6]. The edge weight $w(q_1, q_2)$ is calculated as

$$w(q_1, q_2) = \sum_{i=0}^{n-1} -log(P_i). \qquad (2)$$

(Negative logs are used since each $0 \leq P_i \leq 1$.) By assigning the weights in this manner, we can find the most energetically feasible path in our roadmap when performing queries. This is the same method used in our previous work on protein folding.

4. ROADMAP ANALYSIS

The roadmap is an approximation of the folding landscape, and it can be used to study individual folding pathways as well as the global folding kinetics.

A folding pathway is a sequence of transitional conformations the RNA molecules goes through during the folding process from any unfolded conformation to the native conformation. Similar to our previous work on protein folding [1], we can extract folding pathways and compute the free-energy profile, energy barriers, and important states of the folding process. From all the folding pathways to the native conformation, we extract the pathway with minimum total weight because this corresponds to the most energetically feasible path *in our roadmap.* Individual pathway results are provided for one case study in Section 5.3.

Beyond the study of specific folding pathways, we are interested in the global properties of the energy landscape. For example, how does the population of conformations in the landscape vary as a function of time, i.e., the population kinetics. Folding rates and transition states are also of great interest. These can all be studied using the master equation.

4.1 Folding Kinetics and the Master Equation

Master equation formalism has been developed for folding kinetics in a number of earlier studies [11, 25]. The stochastic process of folding 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 the following master equation:

$$dP_i(t)/dt = \sum_{i \neq j}^{n} (k_{ji} P_j(t) - k_{ij} P_i(t)) \tag{3}$$

where k_{ij} denotes the transition rate from state i to state j. Thus the change in population $P_i(t)$ is the difference between transitions *to* state i and transitions *from* state i. The transition rates are computed from the edge weight: $K_{ij} = K_0 e^{-W_{ij}}$. K_0 is the constant coefficient 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

$$\begin{cases} M_{ij} = k_{ji} & i \neq j \\ M_{ii} = -\sum_{i \neq j} k_{ij} & i \neq j \end{cases} \tag{4}$$

The master equation can be represented in matrix form:

$$d\mathbf{p}(t)/dt = M\mathbf{p}(t). \tag{5}$$

The solution to the master equation is:

$$P_i(t) = \sum_k \sum_j N_{ik} e^{\lambda_k t} N_{kj}^{-1} P_j(0) \tag{6}$$

where N is the matrix of eigenvectors N_i for the matrix M in equation 4 and Λ is the diagonal matrix of its eigenvalues λ_i. $P_j(0)$ is the initial population of conformation j.

From equation 6, 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 as time goes by. 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 Boltzmann distribution of all conformations in equilibrium. To validate our implementation, we compared our master equation results to the Boltzmann distribution, and they match exactly.

For the same reason, we see that the large magnitude eigenvalues correspond to the fast folding modes, that is, those modes which fold in a burst. Their contribution to the population will die away quickly. Similarly, the smaller the magnitude of the eigenvalue is, the more influence its corresponding eigenvector has on the global folding process. Thus, the global folding rates are determined by the slow modes.

For some folders (2-state folders), their folding rate is dominated by only one non-zero slowest mode. If we sort the eigen spectrum by ascending magnitude, there will be one other eigenvalue λ_1 in addition to eigenvalue λ_0, that is significantly smaller in magnitude than all other eigenvalues. This λ_1 corresponds to the folding mode which determines the global folding rate. We will refer it as the *master folding mode*. Its corresponding eigenvector denotes its contribution to the population of each state. Hence, the large magnitude components of the eigenvector correspond to the states whose populations are most impacted by the master folding mode. These states are the transition states [16, 17].

5. RESULTS AND DISCUSSION

With our kinetics analysis tools, we are able to evaluate our roadmap-based approximation of the energy landscape.

Generally, the best way to evaluate an approximation is to compare it to the exact method. Thus, ideally, we should compare the “exact” full base-pair enumeration (BPE) roadmap with the “approximate” stack-pair enumeration (SPE) and the maximal contact sampling (MCS) roadmaps. As previously mentioned, we can afford to do a full enumeration for RNA with up to around 40 nucleotides [10]. Also, there is currently a limit on the size of the master equation we can accurately handle (due to a limitation in our current implementation). For these reasons, we performed the full comparison on a 15 nucleotide sequence and a 21 nucleotide RNA hairpin sequence. As shown in Table 3, we used all three sampling strategies to generate nodes for the roadmap. For the maximal-contact sampling, we tried to generate about twice as many nodes we got with the stack-pair enumeration. We have also processed a 41 nucleotide RNA using stack-pair enumeration.

RNA	Connection	BPE	SPE	MCS
15nt	k-closest	10	10	10
	radius	1	10	20
21nt	k-closest	40	40	40
	radius	1	20	40

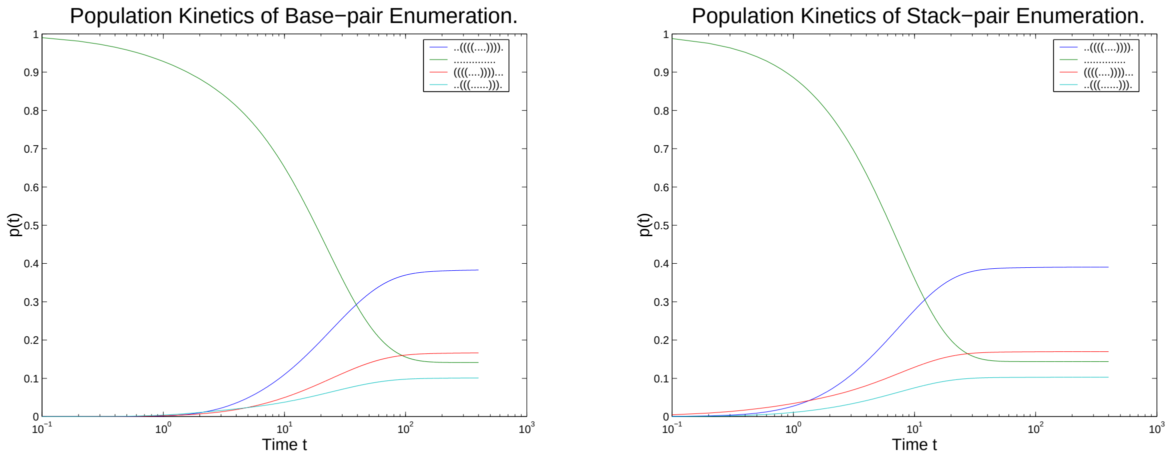
(a)

# Bases	RNA	# Nodes		
		BPE	SPE	MCS
15nt	ACUGAUCGUAGUCAC	142	15	33
21nt	UAUAUAUCGACACGAUAUAUA	5353	250	602

(b)

Table 3: In the tables, BPE, SPE, and MCS denote base-pair enumeration, stack-pair enumeration, and maximal contact sampling, respectively. (a) Parameters used for connection. (b) Number of roadmap nodes for RNA sequences studied.

Each of the roadmaps was connected using both the k-closest and the radius connection strategies described in Section *3.1.2*. The parameters for these methods need to be carefully selected. The roadmap generated using complete base-pair enumeration and radius connections with $r = 1$ corresponds to a fine mesh on the energy landscape. Recall that our distance metric is the base-pair distance, therefore setting $r = 1$ creates transitions between all pairs of nodes differing by a single contact. This roadmap is used as a basis for comparison to determine appropriate k and r values for the other node generation methods. If we increase k or r, the connections will be more complete and more expensive. We want these parameters to be as small as possible yet still large enough to capture the important transitions.



(a) (b)

Figure 5: The population kinetics of the 15 nucleotide hairpin sequence UAUAUAUCGACACGAUAUAUA with the native structure ..((((....)))). Figures (a) and (b) give a comparison the folding kinetics of the base-pair enumeration roadmap (142 conformations) to the stack-pair enumeration roadmap (15 conformations).

For the complete base-pair enumeration, we tested $k = 5$, 10, 15, 20, 30, and 40. We found the smallest values that closely matched the complete landscape for the 15nt and 21nt RNA, respectively. For the other sampling strategies, the distance between conformations is usually greater than 1, thus, we must use larger values for r and k. To determine the appropriate parameters, we compared the kinetics results using $r = 1$, 2, 5, 10, and 20, and $k = 5$, 10, 15, 20, 30, and 40. For the 21nt RNA, $k = 40$ always generated a close approximation to the complete energy landscape. Table 3 gives the parameters used in the results presented here. We only show results for the k-closest-connected roadmaps.

5.1 15nt RNA

Here we provide detailed results for the 15nt RNA. Figure 5 shows the population kinetics of the four most significant conformations calculated using the base-pair and stack-pair roadmaps. These conformations have the largest population during or after the folding process, so their existence is more likely to be observed in experiments. As illustrated in the two figures, their population kinetics are very similar to each other during the folding process. Hence, the stack-pair roadmap is a good approximation of the complete energy landscape. It preserves the main characteristics of the energy landscape while using notably fewer nodes (15 vs. 142). Figures 6 and 7 demonstrate the similarities between the kinetics of the three maps. Most significant is the discovery that the eigenvalues for the base-pair enumeration and the stack-pair enumeration are approximately the same (Figure 6). In addition, the components of the eigenvectors (not shown due to space constraints) are nearly identical. Also in Figure 6, we see that the folding rates of the maximal-contact sampling method are farther from the completely enumerated kinetics than the stack-pair kinetics are. We expected this because the stack-pair method encourages the formation of energetically stable conformations with stacks. The maximal-contact sampling is more random than the stack-pair method and does not attempt to capture the stability inherent in stacking pairs.

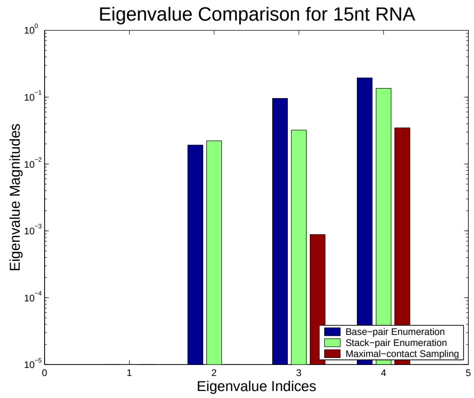
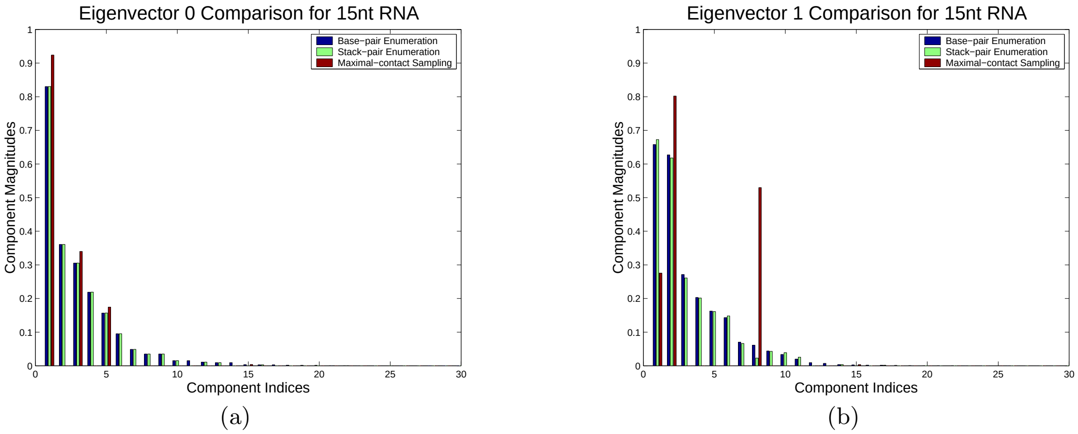


Figure 6: Comparison of the eigenvalues of the 15 nucleotide hairpin sequence ACUGAUCGUAGUCAC with the native structure ..((((....)))). and a C-space size of 142. It illustrates the differences in the eigenvalues and overall folding rates for base-pair enumeration, stack-pair enumeration, and maximal-contact sampling.

Figure 7 compares the folding kinetics of the base-pair enumeration, stack-pair enumeration and maximal-contact sampling. Figure 7(a) shows the equilibrium solutions of the three folding landscapes. They all match very well with the Boltzmann distribution for this molecule. Figure 7(b) illustrates the small differences in magnitude of the components of the second eigenvector for all three folding landscapes. Although the maximal-contact sampling varies more from the complete base-pair enumeration in eigenvector N_1 than in N_0, the differences in magnitude are still relatively small. These results indicate that given some specific conformations, it's possible to examine the folding kinetics of these conformations by computing the folding landscape of that set combined with some additional random sampling. This combination will approximate the slow mode eigenvectors.

5.2 21nt RNA

Figure 8 shows the similarities between the four smallest



(a)

(b)

Figure 7: The folding kinetics of the 15 nucleotide hairpin sequence ACUGAUCGUAGUCAC with the native structure ..((((....)))). and a C-space of 142 conformations. Graph (a) compares the biggest 30 components of eigenvector N_0 and (b) compares the 30 biggest components of eigenvector N_1 for base-pair enumeration, stack-pair enumeration, and maximal-contact sampling.

eigenvalues of the three roadmaps. Different from the results on the 15nt RNA, we found that except for the zero eigenvalue, the other three eigenvalues are comparable to each other. This means its folding behavior is different from the 15nt RNA examined above. As described in section 4.1, all three small modes have a non-negligible influence on the global folding rate. Hence, the contributions of their corresponding eigenvectors to the transition states should not be ignored.

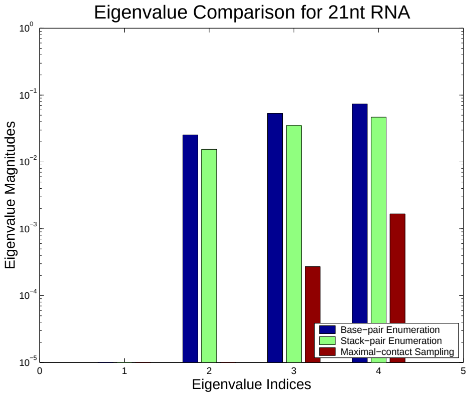


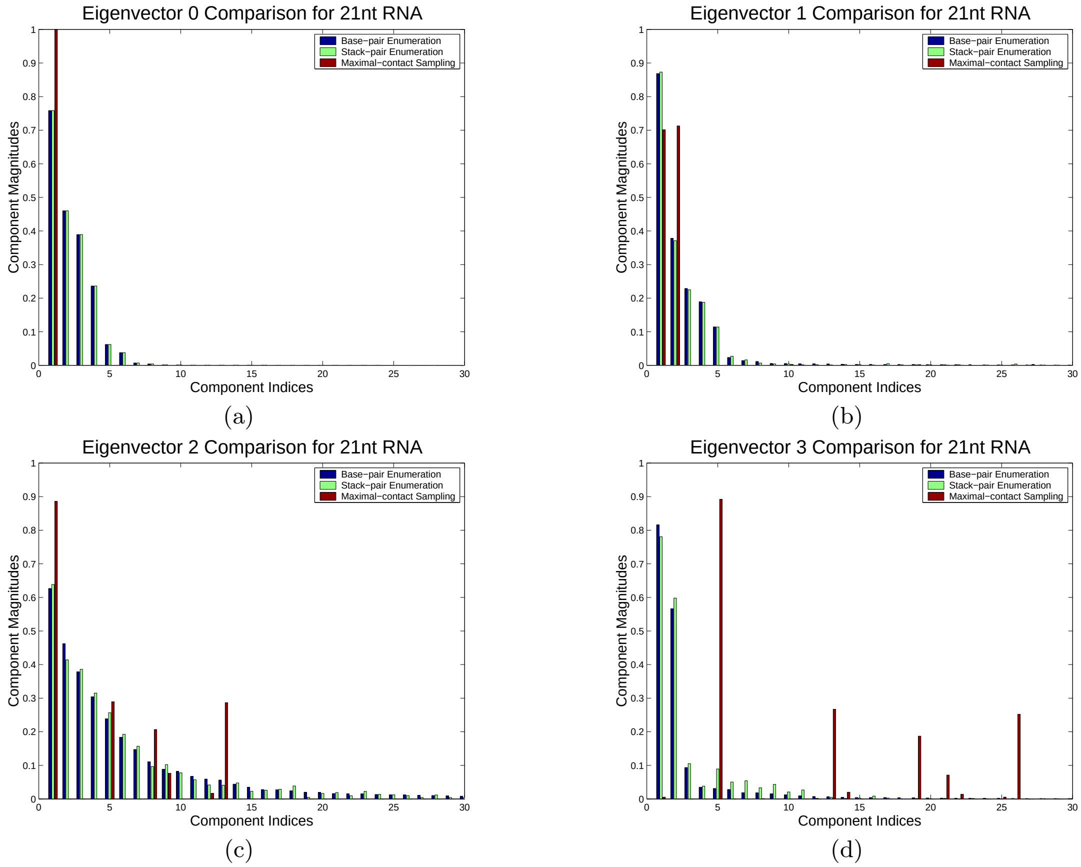
Figure 8: Comparison of the eigenvalues of the 21 nucleotide hairpin sequence UAUAUAUCGACACGAUAUAUA and a C-space of 5353 conformations. It illustrates the differences in the eigenvalues and overall folding rates for base-pair enumeration, the stack-pair enumeration, and the maximal-contact sampling.

In Figure 9, we compared the corresponding eigenvectors of the four small eigenvalues for the three roadmaps. Figure 9(a) shows the equilibrium of distributions, while (b), (c), and (d) show the contribution of the three eigenmodes on the transitional conformations. We found that in Figure 9 (a), (b), and (c), the eigenvectors for base-pair and stack-pair roadmaps are very similar to each other. The random sampling for the maximal-contact conformations, however, missed some important conformations. Yet, for those sampled conformations, the values are similar in magnitude to the corresponding components in stack-pair and base-pair roadmaps. This means that although the random maximal-contact sampling is not accurate enough, it does capture some global properties of the folding process. Also we found that we can easily increase our approximation accuracy by connecting more conformations.

As illustrated above, using the roadmaps generated by three different strategies, we compared the kinetics analysis of two RNA molecules which have different folding behaviors. The roadmap generated by enumeration of base-pairs is the most accurate representation. However, it is not feasible to enumerate any RNA with more than 40 nucleotides. The stack-pair roadmap is still generated by a form of enumeration, but, clearly, this much smaller subset of the full enumeration can effectively approximate the energy landscape, even for RNA with different folding behaviors. The maximal-contact roadmap does not require enumeration and can be of any size we desire. Although its approximation is slightly inferior to the stack-pair roadmap, our preliminary work indicates that this approximation can be improved. Most important, our work demonstrates that we can effectively characterize the energy landscape using notably fewer conformations than exist in the complete enumeration. These results signify that more work into improving our sampling strategies will yield more concise and efficient representations of the energy landscape. Our method is applicable to much longer RNA sequences.

5.3 Folding Pathways Results

Similar to our previous work on protein folding [1], we extract folding pathways and compute the free-energy profile, energy barriers, and important states of the folding process. From all the folding pathways to the native conformation, we extract the pathway with minimum total weight because this corresponds to the most energetically feasible path *in our roadmap*. For a given pathway, its energy profile shows the energy of each transitional conformation and it is easy for us to find the local minima and energy barriers on the pathway. These profiles provide an informal visualization of the folding process.



(a)

(b)

(c)

(d)

Figure 9: The folding kinetics of the 21 nucleotide hairpin sequence UAUAUAUCGACACGAUAUAUA with the native structure (((((((((...))))))))) and a C-space of 5353 conformations. The figures compare the biggest 30 components of eigenvectors (a) N_0, (b) N_1, (c) N_2, and (d) N_3 for base-pair enumeration, stack-pair enumeration and maximal-contact sampling.

An example is given in Figure 10. It shows the energy profile and folding pathway for a 21nt nucleotide RNA (GGCG-UAAGGAUUACCUAUGCC) from a misfolded conformation to the native state. It had to overcome a high energy barrier to reach the native conformation as shown in its energy profile in Figure 10 (a).

6. CONCLUSION

We have demonstrated that the PRM method is a promising technique for studying RNA folding kinetics. PRMs allow us to efficiently characterize the folding landscape using small road-maps, and moreover, our roadmaps were suitable for computing the folding kinetics for the RNA we have studied so far. Our results also indicate that further work on more sophisticated generation and connection methods will yield better results, and this is the subject of current work.

7. 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] W. K. J. C. W. Bruce A. Shapiro, David Bengali. Rna folding pathway functional intermediates: Their prediction and analysis. *JMB*, 312:27–44, 2001.

[4] S.-J. Chen and K. A. Dill. Rna folding energy landscapes. *Proc. Natl. Acad. Sci. USA*, 97:646–651, 2000.

[5] J. Cupal, I. L. Hofacker, and P. F. Stadler. Dynamic programming algorithm for the density of states of rna secondry structures. *Computer Science and Biology 96*, 96:184–186, 1996.

[6] 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.

[7] C. Flamm. Kinetic folding of rna. *Dissertation*, 1998.

[8] D. Gillespie. Exact stochastic simulation of coupled chemical reactions. *JPC*, 81:2340–2361, 1977.

[9] I. L. Hofacker. Rna secondary structures: A tractable model of biopolymer folding. *J.Theor.Biol.*, 212:35–46, 1998.

[10] A. R. P. S. J. Cupal, C. Flamm. Density of states,

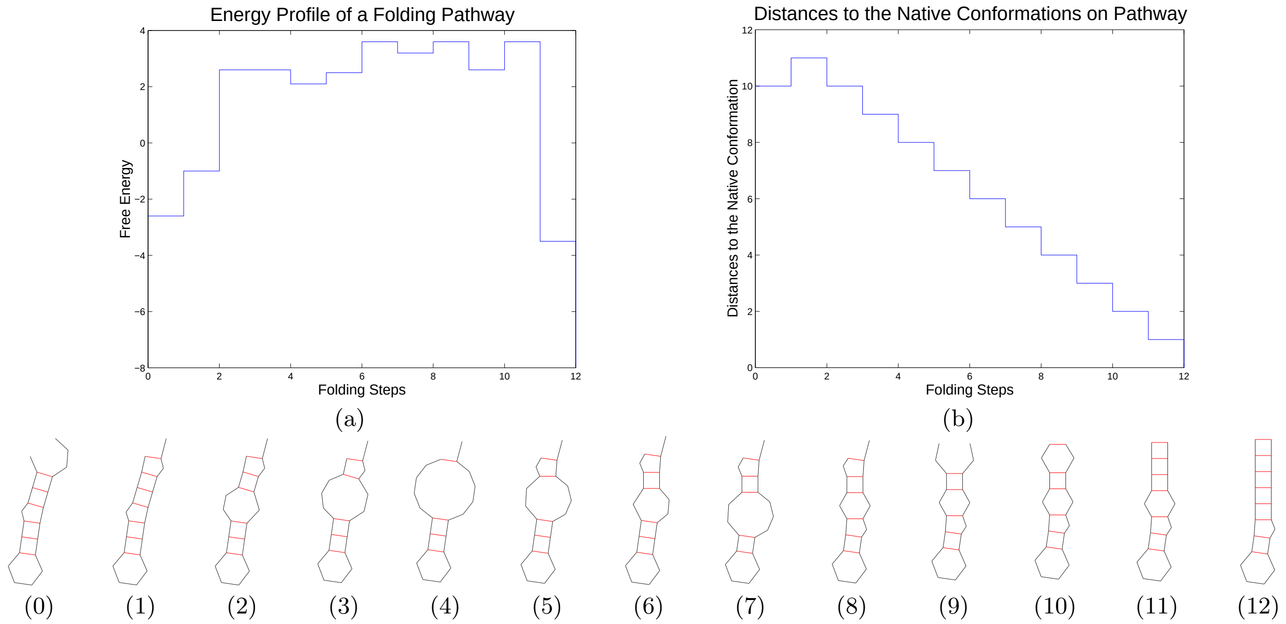


Figure 10: A folding pathway for RNA sequence GGCGUAAGGAUUACCUAUGCC from a misfolded conformation to the native conformation. (a) The energy profile of transitional conformations. (b) The distance of each transitional conformation to the native conformation. (0)-(12) Each transitional conformation numbered according to its position on the pathway.

metastable states, and saddle points exploring the energy landscape of an rna molecule. *Proceedings, Fifth International Conference on Intelligent Systems for Molecular Biology*, pages 88–91, 1997.

[11] N. V. Kampen. Stochastic processes in physics and chemistry. *North-Holland Personal Library*, 1992.

[12] L. Kavraki, P. Svestka, J. C. Latombe, and M. Overmars. Probabilistic roadmaps for path planning in high-dimensional configuration spaces. *IEEE Trans. Robot. Automat.*, 12(4):566–580, August 1996.

[13] J. C. Latombe. *Robot Motion Planning.* Kluwer Academic Publishers, Boston, MA, 1991.

[14] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for rna secondary structure. *Biopolymers*, 29:1105–1119, 1990.

[15] R. Nussinov, G. Piecznik, J. R. Griggs, and D. J. Kleitman. Algorithms for loop matching. *SIAM J. Appl. Math.*, 35:68–82, 1972.

[16] I. B. S. Banu Ozkan, Ken A. Dill. Fast-folding protein kinetics, hidden intermediates, and the seuential stabilization model. 11:1958–1970, 2002.

[17] I. B. S. Banu Ozkan, Ken A. Dill. Computing the transition state population in simple protein models. 68:35–46, 2003.

[18] D. Sankoff and J. Kruskal. Time warps, string edits and macromolecules: the theory and practice of sequence comparison. *Addison Wesley, London*, 1983.

[19] G. Song and N. M. Amato. Using motion planning to study protein folding pathways. In *Proc. Int. Conf. Comput. Molecular Biology (RECOMB)*, pages 287–296, 2001.

[20] 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.

[21] I. Tinoco and C. Bustamante. How rna folds. *J. Mol. Biol.*, 293:271–281, 1999.

[22] A. Walter, D. Turner, J. Kim, M. Lyttle, P. Muller, D. 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.

[23] M. Wolfinger. The energy landscape of rna folding. 2001.

[24] S. Wuchty. Suboptimal secondary structures of rna. *Master Thesis*, 1998.

[25] W. Zhang and S. Chen. Rna hairpin-folding kinetics. *Proc. Natl. Acad. Sci. USA*, 99:1931–1936, 2002.

[26] M. Zuker, D. Mathews, and D. Turner. Algorithms and thermodynamics for rna secondary structure prediction: A practical guide. In J. B. . B. Clark, editor, *RNA Biochemistry and Biotechnology*, NATO ASI Series. Kluwer Academic Publishers, 1999.

[27] M. Zuker and D. Sankoff. Rna secondary structure and their prediction. *Bulletin of Mathematical Biology*, 46:591–621, 1984.