Modeling RNA Folding Landscapes with Probabilistic Roadmap Methods
Parasol Lab Technical Report #TR03-004

Bonnie B. Kirkpatrick
Undergraduate
Montana State University

Xinyu Tang
Graduate Student Mentor

Shawna Thomas
Graduate Student Mentor

Nancy M. Amato
Faculty Mentor

PARASOL Lab., Department of Computer Science
Texas A&M University
bkirk@cns.montana.edu, {xinyut,sthomas,amato}@cs.tamu.edu

Bonnie Kirkpatrick
765 W Park St
Dillon, MT 59725

Abstract

Knowledge of the energy landscape of an RNA molecule is necessary to understand its folding kinetics and final structure. To efficiently analyze the energy landscape of RNA folding without enumeration, we apply the probabilistic roadmap *(*PRM*) motion planning method which has been successfully applied to the study of protein folding landscapes. In addition, our model deals with secondary structure and connects conformations using methods involving the base-pair distance measurement. The* PRM*-based algorithms for roadmap construction using conformation generation and connection strategies are discussed in this paper. We determine the population kinetics using a statistical mechanical model and predict the folding rates, intermediate states, and folding pathways. Preliminary results are given in this paper.*

1 Introduction

1.1 RNA Folding Problem

RNA performs diverse and important functions including synthesizing protein, catalyzing reactions, splicing introns, and regulating activities [8]. The goal of research on folding kinetics is to further the understanding of these functions and the structure of the RNA that performs them. RNA molecules fold into energetically optimal conformations referred to as native states. For our research, we presuppose a known native structure, and we use this to help us efficiently explore the funnel-shaped folding landscape. The landscape contains all possible conformations of the molecule and their corresponding potential energies. Properties of the landscape are folding pathways, intermediate states, transition rates, and population kinetics.

A complete folding landscape containing every possible tertiary conformation is too large to compute for most sequences. In an effort to produce a more useful model, two simplifications are made. The first is to consider only secondary structures since they provide sufficiently useful structural information. However, the number of the secondary structure conformations increases exponentially with the length of the RNA sequence [12], leaving it too large to compute for sequences over 80 nucleotides. The second simplification we use is the *probabilistic roadmap* (PRM) approach to sample the characteristic points in the landscape without full enumeration. This method grew out of robotics research for planning the motion of an object through an environment. Here our environment is the folding landscape, and the object navigating the environment is the RNA molecule. In this paper, we describe the PRMmethod explain roadmap analysis, and give results for several small RNA sequences.

1.2 Related Work

Previous work falls into three categories: energy calculations, structure prediction, and the study of folding kinetics. Energy calculations are the essential component to all research on RNA folding. One commonly used energy function is the Turner or nearest-neighbor rules. 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 [11].

Structure prediction, the second area of study, is commonly solved with dynamic programming. Zuker and Stiegler formulated the popular MFOLD algorithm to address the minimum energy problem [9]. Based on this algorithm, the Theoretical Biochemistry Group at the University of Vienna developed the ViennaRNA package which implemented Zuker's algorithms and some energy functions [2].

Several methods have been proposed that involve computations on the folding landscape. One method generates all the secondary structures within some given energy range of the native structure. Sequences of 80 nucleotides or shorter can be handled by this method [1]. Another method for computing the population kinetics of the folding landscape is to solve the rate equation. This approach uses the master equation and examines the dominate modes of the solution [10].

2 RNA Folding with PRM Methods

In this section, we explain our motion planning approach to exploring folding landscapes. After introducing the *probabilistic roadmap* method, we explain how it applies to RNA folding landscapes. We formally define our RNA model of secondary structure, energy calculations, and distance metrics. Finally, we give our roadmap construction method and discuss ways to sample from the roadmap.

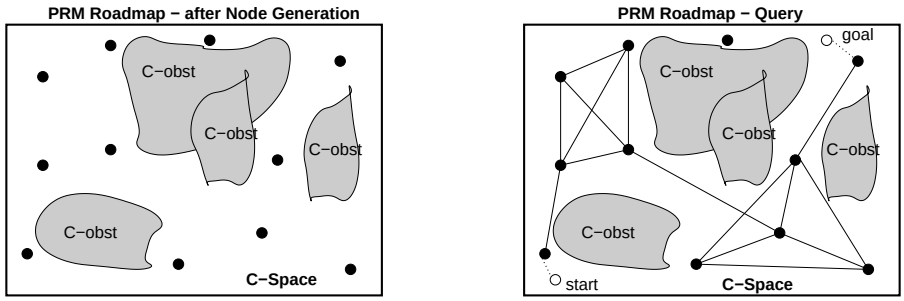
2.1 The Probabilistic Roadmap Method (PRM)

Our approach to RNA folding is based on the *probabilistic roadmap* (PRM) approach for motion planning [3]. The motion planning problem is one of finding appropriate paths for locomotion through a given environment. Typically, PRMs are used to approximate the environment's feasible regions by constructing a map. This map, or roadmap, can be used subsequently to answer many, varied queries for practical paths through the environment. Briefly, PRMs work by sampling points 'randomly' from the movable object's configuration space,* and retaining those that satisfy certain feasibility requirements (e.g., collision-free configurations, see Figure 1(a)). Then, these points are connected to form a roadmap, using a simple local planning method to connect nearby points (see Figure 1(b)). During query processing, paths connecting the start and goal configurations are extracted from the roadmap using standard graph search techniques (see Figure 1(b).)

A major strength of PRMs is that they are quite simple to apply, even for problems with high-dimensional configuration spaces, requiring only the ability to randomly generate points in C-space, and then test them for feasibility.

In previous work, we proposed the PRM framework as a methodology for studying protein folding when the native structure is known [7]. The main difference from the usual PRM application is that the collision detection feasibility test is replaced by

*The movable object's *configuration space*, or C-space, is the set of all positions and orientations of the movable object, feasible or not [4].



(a) (b)

Figure 1: A PRM roadmap in C-space. A PRM roadmap: (a) after node generation, and (b) after the connection phase and being used to solve a query.

a preference for low energy conformations. We obtained very promising results for several small proteins (e.g., proteins G and L, both with approximately 60 residues), and in particular, we showed that the pathways extracted from our roadmaps are in agreement with known experimental results [5].

Our PRM-based technique for RNA folding differs from protein folding in that the C-space is not continuous but discrete. However, even those discrete points can always be connected by interpolating two conformations. Therefore we can sample nodes in the space and connect them to build the roadmap.

2.2 Model of RNA Conformation

Each RNA molecule is a sequence of nucleotides that differs from others in its bases. There are 4 bases: adenine (A), cytosine (C), guanine (G) and uracil (U). Hydrogen bonds form between complementary Watson-Crick base pairs C-G and A-U. The other strong base-pairing considered in our model occurs in the wobble pair G-U. Each RNA secondary structure conformation is denoted by a set of base pairs. Valid secondary structures must meet three criteria. For any two contacts $[i, j]$ and $[k, l]$ with $i < j$ and $k < l$, then:

1. Both contacts must be valid base-pairs
2. Each base must be paired to only one other:
   $i = k$ if and only if $j = l$
3. No pseudo-knots are allowed:
   if $i < k < j$, then $i < k < l < j$

The complete folding landscape corresponding to the *conformation space* (C-space) consists of all valid secondary structures. These points are refered to interchangably as conformations and roadmap nodes.

2.2.1 C-Space

The size of the conformation space for this model is dependent on the sequence of the RNA. Here we formalize the discussion of size. Let $\mathcal{U}$ be the set of every possible

combination of base pairs, including invalid combinations. From this definition, it is clear that the size of $\mathcal{U}$ depends only on the number of all possible contact pairs, m, and is related to the sequence length, n, as follows:

$$|\mathcal{U}| = \sum_{k=0}^{m} \binom{m}{k} = 2^m, \quad \text{where} \quad m \leq \frac{n^2}{4} \tag{1}$$

Let $\mathcal{C}$, the conformation space, be the valid combinations of base pairs making $\mathcal{C} \subset \mathcal{U}$. The size of $\mathcal{C}$ depends on the specific sequence, as well as the length of the sequence. Zuker and Sankoff [12] have developed a much closer estimate of the size of $\mathcal{C}$. This approximation calculates only valid conformations and uses a stochastic approach to account for the effect of a specific sequence on the size of the conformation space. Given an RNA sequence of length n, we calculate the probabilities $p(A), p(C), p(G), p(U)$ of the occurence of each nucleotide in the sequence. Let

$$p = 2(p(A)p(U) + p(C)p(G) + p(G)pG(U)) \tag{2}$$

be the probability of two bases forming a contact. Then the approximation is:

$$|\mathcal{C}| \sim hn^{-3/2}\alpha^n \tag{3}$$

where

$$\alpha = (\frac{1+\sqrt{1+4\sqrt{p}}}{2})^2, \quad \text{and} \quad h = \frac{\alpha(1+4\sqrt{p})^{1/4}}{2\sqrt{\pi}p^{3/4}} \tag{4}$$

For example, the actual C-space for the hairpin sequence ACUGAUCGUAGUCAC with the native structure ..((((....)))). is 142. For this sequence, $|\mathcal{U}| \approx 1.7 \times 10^7$. The approximate $|\mathcal{C}| \approx 1067$.

2.2.2 C-Space Measurments

The definition of C-space considers only whether a set of contacts is valid, and says nothing of its energetic feasiblablity. We use the ViennaRNA package energy function which relies on the Turner rules to ascertain the viability of a conformation. This *energy calculation* also determines the edge weight calculation.

To find the neighbors of a particular conformation we need a metric to measure distances in C-space. Here we use the *base-pair distance metric*. This denotes the number of base pairs that have to be opened or closed to transform one conformation into another. We chose the base-pair distance because it is commonly used to study folding landscapes.

2.3 Roadmap Construction

Our goal is to avoid enumerating the entire landscape due to its size. We do this by choosing some set of points from $\mathcal{C}$ that adequately characterize the features of the landscape. Thus, our node sampling method is crucial in applying PRMs to

RNA folding. Our choice of sampling methods will determine the effectiveness of our method.

We have designed our implementation to exploit the size of C-space in testing the effectiveness of our sampling methods. The first step involves the creation of the complete roadmap which enumerates all valid conformations in the C-space. We compare this roadmap with other subsets of conformations in our Results section (3.2) to determine which sampling and connection methods better represent the landscape. These methods will be useful for exploring the landscapes of other RNA whose sequences are too long to completely enumerate.

2.3.1 Node Generation

Taking advantage of the size and discreteness of C-space, stack we have implemented several methods for node generation including complete enumeration, stack-based enumeration, and maximal-contact sampling. The *complete enumeration* is done in a straight forward manner that enumerates all the possible conformations in the entire space. Only conformations with at least 3 unpaired bases in a hairpin loop are kept in order to satisfy the energy calculation. This generation method is only feasible for short RNA.

Stack-based enumeration is a subset of nodes from the complete enumeration with the addition of the stack requirement. Every stack-based node must contain only those base pairs that occur in a stack. Formally, for any contact $[i, j]$, $i < j$, there must exist another contact $[k, l]$, $k < l$ such that:

1. Together they form a valid secondary structure
2. Either (a) $i - 1 = k$ and $j + 1 = l$; or (b) $i + 1 = k$ and $j - 1 = l$

This encourages the formation of helices in the enumerated conformations and contributes to the energetic stability of the folded molecule.

In the *maximal-contact sampling* nodes are generated in a 'random' fashion. The first step is to create a conformation c without any base pair contacts. Then a single valid contact is added to the secondary structure of c in each iteration of the loop. The loop finishes when we have generated a maximal-contact conformation—meaning that valid pairs are added until no more can be added while maintaining the validity of the secondary structure. This method has the effect of biasing the distribution of nodes toward the areas of C-space with more contacts. This typically yields lower energy conformations.

2.3.2 Node Connection

Neighboring roadmap nodes are connected using a local planner. A conformation is a neighbor if it is one of the k-closest nodes to a particular node The local planner then connects each pair of neighbors by generating intermediate nodes on a path between them.

Each path is a sequence of intermediate conformational changes the RNA molecule folds through as when transition between the neighbors. Each of the intermediate

nodes should have as many contacts as possible to exclude high energy differences on the path. Once we open a base pair, we will close all base pairs whose pairing does not introduce invalid secondary structure.

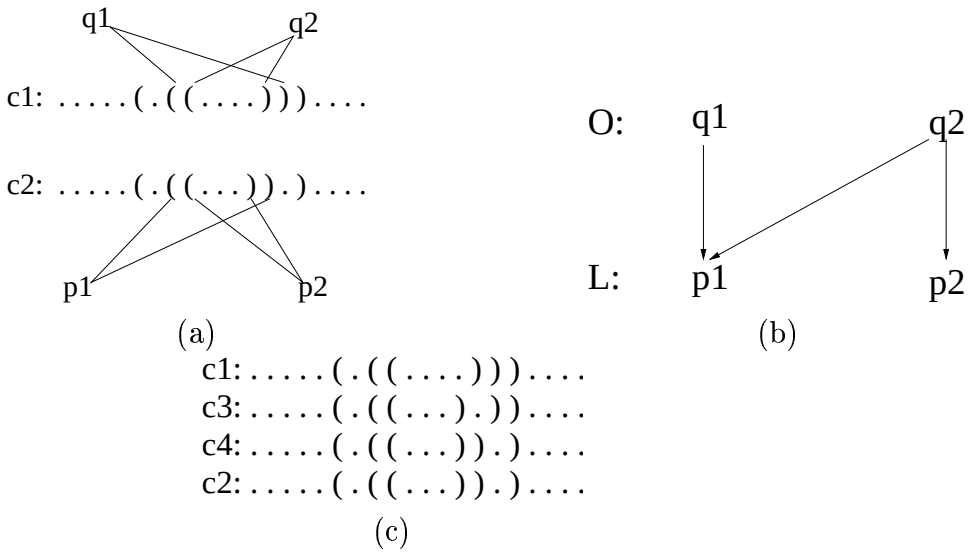


Figure 2: Intermediate node generation. (a) Start and goal conformations and contact pairs to be opened and closed: $q1$, $q1$ are in $\mathcal{O}$; $p1$, $p2$ are in $\mathcal{L}$. (b) Dependency graph: $p1$ depends on $q1$ and $q2$, $p2$ depends on $q2$. (c) Sequences generated: $c3$ and $c4$ are the two intermediate conformations to connect $c1$ and $c2$, here $c4$ happens to be identical to $c2$.

To select the intermediate nodes, we first find the sets $\mathcal{O}$ and $\mathcal{L}$ of base pairs to be opened and closed respectively. In Figure 2 (a) the pairs to be opened are $q1$, $q2$ and $p1$, $p2$ need to be closed. Next, we construct the dependency graph between sets $\mathcal{O}$ and $\mathcal{L}$ of the base pairs that cannot exist together in a valid conformation. In the example, in Figure 2 (b), $p1$ depends on $q1$ and $q2$ while $p2$ depends on $q2$. Finally, we use a heuristic to choose which order is best for opening the base pairs. We open pairs that will allow us to close other contacts. In the Figure 2 (c), $c3$, $c4$ are the two intermediate conformations generated.

Each edge connecting neighbors is assigned an edge weight that reflects the transition rate between neighboring states. This is the probability of an RNA folding from one to the other of the conformations. Hence, the edge weight reflects the energy barrier for the folding process on this edge

When an edge (q_1, q_2) is added to the roadmap. Each edge (q_1, q_2) is assigned a weight that depends on the sequence of intermediate conformations
$\{q_1 = c_0, c_1, c_2, \ldots, c_{n-1}, c_n = q_2\}$ connecting q_1 and q_2. 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 a function of the difference between their potential energies, $\Delta E_i = E(c_{i+1}) - E(c_i)$. In our work, we adopt the Metropolis rule:

$$P_i = \begin{cases} e^{\frac{-\Delta E_i}{kT}} & \text{if } \Delta E_i > 0 \\ 1 & \text{if } \Delta E_i \leq 0 \end{cases} \tag{5}$$

This keeps the detailed balance between two adjacent conformations, and enables the weight of an edge to be computed by the sum of the logarithms of all n probabilities:

$$w(q_1, q_2) = \sum_{i=0}^{n-1} -log(P_i), \tag{6}$$

Negative logs are used since $0 \leq P_i \leq 1, \forall i$. By assigning the weights in this manner, we can find the most energetically feasible path in our roadmap when performing queries. A similar weight function, with different probabilities, was used by Latombe [6] and in our previous work on protein folding [7].

3 Roadmap Analysis

3.1 Master Equation

The master equation formalism has been developed for folding kinetics in a number of earlier studies. The stochastic process of folding is represented as a set of transition rates among n conformational states. $P(t)$ denotes the time evolution of the populations of all the conformational states. For a single conformation, the population change, $P_i(t)$, is described by the following master equation:

$$dP_i(t)/dt = \sum_{i \neq j}^{n} (k_{ji} P_j(t) - k_{ij} P_i(t)) \tag{7}$$

Here k_{ij} denotes the transition rate from state i to state j. These transition rates must satisfy the detailed balance assumption. This assumption causes the system to converge to the equilibrium Boltzmann distribution.

If we use an n-dimensional column vector $\vec{p}(t) = (P_1(t), P_2(t), ..., P_n(t))^T$ to denote the population of all n conformational states. Then we can construct an $n \times n$ matrix M, where

$$\begin{cases} M_{ij} = k_{ji} & i \neq j \\ M_{ii} = -\sum k_{ij} & i \neq j \end{cases} \tag{8}$$

Now the master equation can be writen in the matrix form:

$$d\vec{p}(t)/dt = M\vec{p}(t) \tag{9}$$

We can solve the matrix N of eigen vectors n_i for the matrix in equation 8, and the diagnal matrix Λ of its eigen values λi.

$$\vec{p}(t) = Ne^{\Lambda t} N^{-1} \vec{p}(0) \tag{10}$$

This allows us to write:

$$P_i(t) = \sum_{k=1}^{n} N_{ik} e^{\lambda_k t} C_k \quad \text{where} \quad C_k = \sum_j N_{kj}^{-1} X_j(0) \quad \text{and} \quad X(t) = N^{-1}\vec{p}(t) \tag{11}$$

From this, we can see that the population kinetics for each conformation is actually the sum of the contributions of all n independent kinetic eigen modes. All eigenvalues are negative, with the smallest magnitude eigenvalue $|\lambda_0| = 0$. This eigenvector, N_0, is the equilibrium solution and is the Boltzmann distribution.

The population kinetics of the system are dominated by the slow mode eigenvalues. These small magnitude, non-zero eigenvalues correspond to slow folding and dominate the folding rate. Consequently, we number the eigenvalues (and corresponding eigenvectors) by their sorted order with $|\lambda_0| = 0 < |\lambda_1| < ... < |\lambda_n|$. For 2-state folders, λ_1 will be several orders smaller than than the other eigenvalues. This folding mode determines the global folding rate. Its eigenvector denotes this folding mode's contribution to the equilibrium population of each conformation.

The crutial assumption for the validity of the population kinetic solution is the detailed balance of the transition rates. As we discussed in section 2.3.2, the edgeweight already ecodes the information for transition rates. Therefore, the rate constants can be computed by edgeweight:

$$K_{ij} = k_0 e^{-W_{ij}} \tag{12}$$

The rate constant is scaled with a coefficient k_0 to match the rate constant determined in experimental results.

3.2 Results

We studied a 14 nucleotide RNA hairpin sequence using our PRM method. The sequence ACUGAUCGUAGUCAC has a base-pair enumeration of 142 conformations and a stack-based enumeration of 15 conformations. We also generated a random roadmap of 28 conformations using the maximal-contact sampling method. Each of these roadmaps was connected using the radius connection strategy, but with different radii. The base-pair enumeration had $r = 2$, while for the stack-based enumeration $r = 10$, and the maximal-contact sampling method $r = 20$.

Figure 3.2 shows several graphs demonstrating 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-based enumeration are approximately the same in graph (c). In addition the components of the eigenvectors (not shown due to space constraits) are nearly identical. Also in graph (c), we see that the folding rates of the maximal-contact sampling method are farther from the completely enumerated kinetics than the stack-based kinetics are. We expected this, because the stack-based pairs encourage the formation energetically stable conformations with helices. The maximal-contact sampling is more random than the stack-based pairs, and does not attempt to capture the stability inherent in stacking pairs.

Also in Figure 3.2 the first two graphs compare the folding kinetics of the base-pair enumeration and the maximal-contact sampling. The first graph (a) shows the equilibrium solutions of the two folding landscapes. They both match very well with the Boltzmann distribution for this molecule. The second graph illustrates the small differences in magnitude of the components of the second eigenvector for both fold-

ing landscapes. Although the maximal-contact sampling eigenvector N_1 varies more from the kinetics of the complete enumeration than the equilibrium solution, N_0, the differences are still very small. These results indicate that given some specific conformations, we can examine the folding kinetics of these conformations by computing the folding landscape of that set combined with some additional random sampling. This is sufficient to compute the equilibrium solution and approximate the slow mode eigenvectors.

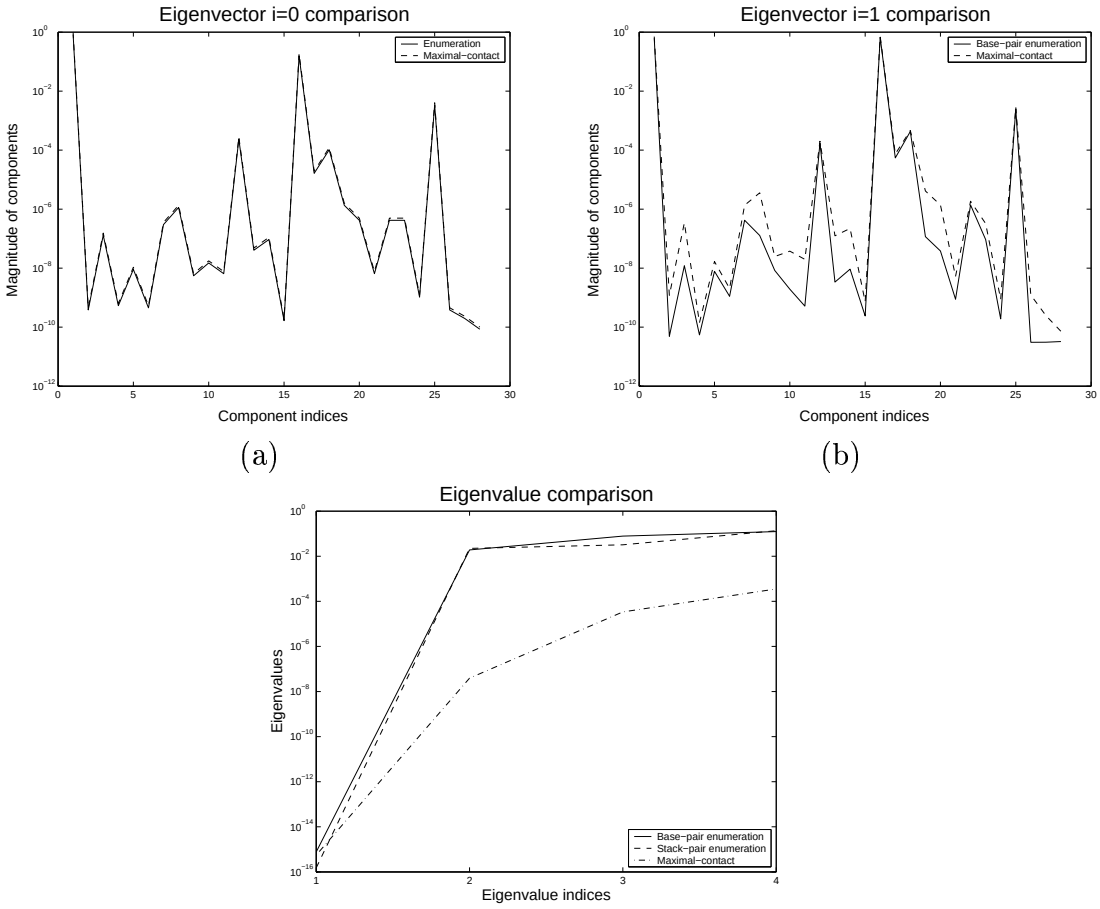


Figure 3: The folding kinetics of the 14 nucleotide hairpin sequence ACUGAUCGUAGUCAC with the native structure ..((((....)))). and a C-space of 142 conformations. The first two graphs compare the folding kinetics of the base-pair enumeration roadmap to the maximal-contact sampling roadmap of 28 conformations. Graph (a) shows the components of the eigenvector N_0 for each roadmap, while graph (b) shows the eigenvector N_1 for each roadmap. The final graph (c) illustrates the differences in the eigenvalues and overall folding rates for base-pair enumeration, the stack-based enumeration, and the maximal-contact sampling.

4 Conclusion

We have demonstrated that the PRM method is quite applicable to the problem of studying RNA folding kinetics. Prababilistic methods allow us to efficiently characterize the folding landscape. Our roadmaps model the complete statistical mechanical model of folding landscapes and allow us to well approximate the folding kinetics. We have presented preliminary results that demonstrate the effectiveness of our approach.

References

[1] Jan Cupal, Ivo L. Hofacker, and Peter F. Stadler. Dynamic programming algorithm for the density of states of rna secondry structures. *Computer Science and Biology 96*, 96:184–186, 1996.

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

[3] 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.

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

[5] R. Li and C. Woodward. The hydrogen exchange core and protein folding. *Protein Sci.*, 8(8):1571–1591, 1999.

[6] A.P. Singh, J.C. Latombe, and D.L. Brutlag. A motion planning approach to flexible ligand binding. In *7th Int. Conf. on Intelligent Systems for Molecular Biology (ISMB)*, pages 252–261, 1999.

[7] 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.

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

[9] 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.

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

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

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

Acknowledgements

This 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, by the DOE ASCI ASAP program, and by the Texas Higher Education Coordinating Board grant ATP-000512-0261-2001.

Biography

Bonnie Kirkpatrick grew up in Montana, and she is now a Computer Science major at Montana State University (MSU). The summer of 2001 found her doing database research at MSU with Dr. Gwen Jacobs at the Center for Computational Biology. Also the same summer, she began working with MSU computer science professor Dr. Brendan Mumey on a protein structure prediction project. For the summers of 2002 and 2003, Bonnie has been at Texas A&M University working under the direction of Dr. Nancy Amato to study protein and RNA folding landscapes.