Using Motion Planning to Map Protein Folding Landscapes and Analyze Folding Kinetics of Known Native Structures*

Bonnie B. Kirkpatrick†	**Guang Song**‡	**Xinyu Tang**‡	**Nancy M. Amato**‡
Montana State University	Texas A&M University	Texas A&M University	Texas A&M University
USRG/DMP Participant	Graduate Student Mentor	Graduate Student Mentor	Faculty Mentor
`bkirk@cns.montana.edu`	`gsong@cs.tamu.edu`	`xinyut@cs.tamu.edu`	`amato@cs.tamu.edu`

Abstract

Knowledge of the energy landscape of a biopolymer molecule is necessary to understand its folding kinetics and final structure. To analyze the energy landscape of RNA folding, we apply the PRM (Probabilistic Roadmap Method) method which has been successful in the study of protein folding landscapes. Although RNA conformation spaces (secondary structures) are not continuous as are those of proteins, they can nevertheless be connected by intermediate conformations, and thus fit well the PRM framework. PRM-based algorithms for RNA conformation generation and connection are described in this paper.

1 Introduction

Features of RNA folding landscapes can reveal much about the kinetics and thermodynamics of RNA. Energy barriers, folding routes, conformational changes, and intermediate states can all be found in the features of the landscape. Thus, finding these important features has become a central goal of research in RNA folding.

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 usually made. The first is to consider only secondary structures since they exist in two dimensions, rather than in three, and still provide sufficient structural information. This significantly reduces the size of the folding landscape, but still it remains too large for complete enumeration of sequences longer than about 80 nucleotides [2]. The second simplification is to use methods that sample the characteristic points in the landscape without full enumeration.

Related work. 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. For short RNA chains, it is possible to exhaustively enumerate all the secondary structures. Sequences of 80 nucleotides or shorter can be handled by this method. However, the number of the secondary structures increases exponentially with the length of the sequence [3].

Some researchers use dynamic programming to overcome the sequence length limitation. Nussinov came up with a dynamic programming solution to find the structure with the maximum number of base pairs [8]. Zuker and Stiegler formulated an algorithm to address the minimum energy problem. John McCaskill's algorithm can calculate the partition function. As described in [Chen & Dill], the partition function is the sum of Boltzmann factors over all possible ways and branching patterns in which the chain can be arranged into helices and intervening regions. Each Boltzmann factor accounts for the base pairing and stacking free energies for that particular configuration. Their

*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.

†Research supported in part by the CRA Distributed Mentor Project. Department of Computer Science, Montana State University, Bozeman, MT 59717.

‡Department of Computer Science, Texas A&M University, College Station, TX 77843-3112.

algorithm recursively calculates the partition function over all secondary structures by dynamic programming [7]. Wuchty extended this algorithm to compute the density of states at a predefined energy resolution [12]. Based on this algorithm, they developed the ViennaRNA package which implemented Zuker and McCaskill's algorithms as well as some energy functions [2].

Some recent theoretical and experimental advances are beginning to shed some light on the full energy landscape. Matrices have been used to compute the partition function over all possible secondary structures. Complete folding landscapes can be computed by this method for sequences shorter than 200 nucleotides [1].

1.1 Our contribution.

In this paper we present an application of *probabilistic roadmap* methods (PRM) to RNA folding landscapes. By using a motion planning approach, we are able to quickly characterize the most significant features of the folding landscape. These features are found by an approximation of the folding landscape, rather than a full enumeration. In addition, PRM has an advantage over other methods in that it uses both global and local properties to characterize the landscape.

1.2 Outline

First, an introduction to PRM methods will be presented. Next, in Section 2.1, we describe how RNA folding kinetics are modeled using PRM. Algorithms for node generation and node connection are presented in Sections 2.2 and 2.3.

2 A Probabilistic Roadmap Method for RNA Folding

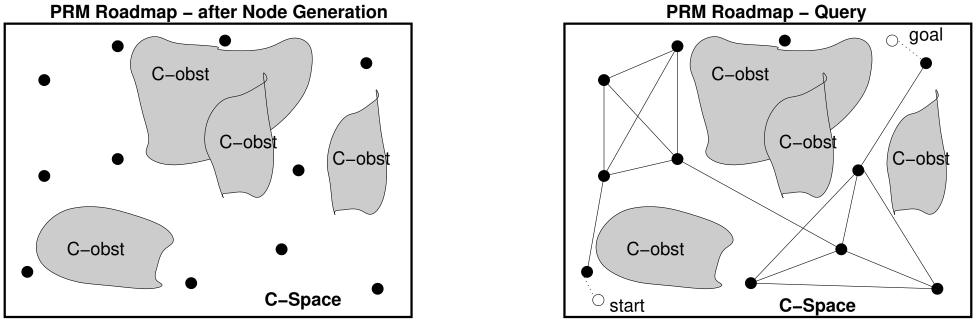
Our approach to RNA folding is based on the *probabilistic roadmap* (PRM) approach for motion planning [4]. Typically, PRMs are used to construct a map of the feasible regions of the environment which can be used subsequently to answer many, varied motion planning queries. Briefly, PRMs work by sampling points 'randomly' from the movable object's configuration space,[1]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 graph, or roadmap, using some 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 (local connection can often be performed using multiple applications of the feasibility test).

In previous work, we proposed the PRM framework as a methodology for studying protein folding when the native structure is known [11]. The main difference from the usual PRM application is that the collision detection feasibility test is replaced by a preference for low energy conformations. We obtained very promising results for several small proteins (e.g., proteins A and GB1, both with approximately 60 residues), and in particular, we showed that the pathways extracted from our roadmaps seemed to be in agreement with known experimental results [6].

Our PRM-based technique for RNA folding differs from protein folding in that the configuration space is not continuous but is discrete instead. However, even those discrete points can always be connected by interpolating between two configurations. Therefore we can sample nodes in the

[1]The movable object's *configuration space*, or C-space, is the set of all positions and orientations of the movable object, feasible or not [5].



(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.

space and connect them to build the roadmap. Again, as with protein folding, an evaluation of conformation energy replaces collision detection testing.

2.1 Modeling RNA Secondary Structures (C-space)

Each RNA molecule is a sequence composed of nucleotides, which differ from each other in their bases. There are 4 types of bases: adenine (A), cytosine (C), guanine (G) and uracil (U). The complementary bases, C-G and A-U, form stable contact pairs with each other through the creation of hydrogen bonds between donor and acceptor sites on the bases. These types of base pairs are known as Watson-Crick pairs. In addition, we consider the weaker G-U contact pairs, which are known as Wobble pairs. For an RNA sequence, each secondary structure configuration can be denoted by a set of contact pairs. The complete folding landscape, which corresponds to the *configuration space* (C-space), consists of all combinations of Watson-Crick and wobble base pairs. Each point in the space represents a set of contacts, which specifies a set of secondary structures.

A valid set of base pair contacts 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 either:
   (a) Watson-Crick pairs: [A, U] or [G, C]; or
   (b) Wobble pairs: [G, U]
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 $i < k < l < j$

We give some examples in Fig. 2 which violate rule (1) and (2).

(a) (b)

Figure 2: Illegal contact pairs: (a) violating rule (2), and (b) violating rule (3), the pseudo knot.

To formalize this discussion, let $\mathcal{U}$ be the set of every possible combination of base pairs. From this definition, it is clear that the size of $\mathcal{U}$ depends only on the number of all possible contact pairs, n, and is quite large.

$$|\mathcal{U}| = \sum_{k=1}^{n} \binom{n}{k} = \sum_{k=1}^{n} \frac{n!}{k!(n-k)!} = O(2^n) \tag{1}$$

Let $\mathcal{C}$ be the valid combinations of base pairs:

$$\mathcal{C} = \{c \in \mathcal{U} : c \text{ has valid contacts}\} \tag{2}$$

The size of $\mathcal{C}$ depends on the specific sequence, as well as the length of the sequence. Clearly, the size of $\mathcal{C}$ is bounded above by the size of $\mathcal{U}$. For example, the C-space for the sequence $(\mathrm{ACGU})_2$ has size 5. This is much smaller than $\mathcal{U}$ for the same sequence.

The definition of $\mathcal{C}$ considers only whether a set of contacts is valid, and says nothing on whether it is feasible. A third-party energy function, which is part of the ViennaRNA package, is used to determine the validity of a point in C-space.

2.2 Node generation

Our goal is to avoid completely enumerating the landscape or C-space by choosing some set of points from $\mathcal{C}$ that adequately describe the features of the landscape. This makes our node sampling method a crucial step in applying PRM to RNA folding.

2.2.1 Sampling strategy

Nodes are generated in a 'random' fashion. The first step is to create a configuration c without any contact pairs. Then a single contact is added to c in each iteration of a loop. Each step preserves the condition that c contains a valid set of base pair contacts. Each time we generate a maximal configuration so that contacts are added until no more can be added— meaning that there are no remaining contacts that do not conflict with the contact set of c.

Each time a contact is randomly sampled, it is selected from a contact matrix M that contains all the remaining possible contacts. That is, adding the contact to c does not contradict the three validity requirements. An important component of our sampling strategy is to maintain M so it contains only contacts that do not conflict with the current contacts of c. The contacts in M may conflict with each other, but not with c.

At each step, a contact is randomly chosen from M, added to c, and M is updated accordingly. This continues until M is empty and no more contacts can possibly be added to c.

Once a node is generated, its potential energy $E(q)$ is evaluated and it is added to the roadmap with the following probability:

$$P(\text{accept} \quad q) = \begin{cases} 1 & \text{if } E(q) < E_{\min} \\ \frac{E_{\max}-E(q)}{E_{\max}-E_{\min}} & \text{if } E_{\min} \leq E(q) \leq E_{\max} \\ 0 & \text{if } E(q) > E_{\max} \end{cases}$$

The values $E_{\min}$ and $E_{\max}$ are parameters of the method.

2.2.2 Distribution of nodes

The distribution of nodes is biased toward the area of C-space near the native conformation. Since every node generated has a maximal number of contacts, the nodes generated are more clustered around the native state than they would be if fewer contacts were made.

2.3 Connecting the roadmap

The second step of constructing the roadmap is to connect the generated nodes. To obtain a good roadmap, we should build representative, low energy edges between nodes. For each of the roadmap nodes, we first determine its k nearest neighbors, for some specified constant k. Then, we connect them with a local planner, which is a simple method that operates on a local scale.

2.3.1 Distance metrics

To find the nearest neighbors of a node, we need a metric to measure the distance between two nodes (RNA conformations). Here we use base-pair distance (the number of contact pairs that differ between two conformations), which actually denotes the number of base pairs that have to be opened or closed to transform one conformation into another.

There are many other distance metrics that could be used. For example, since the RNA secondary structures can be represented by strings or trees, we can use string edit distance or tree edit distance [9] to measure the dissimilarity between two secondary structures.

2.3.2 Generate intermediate nodes

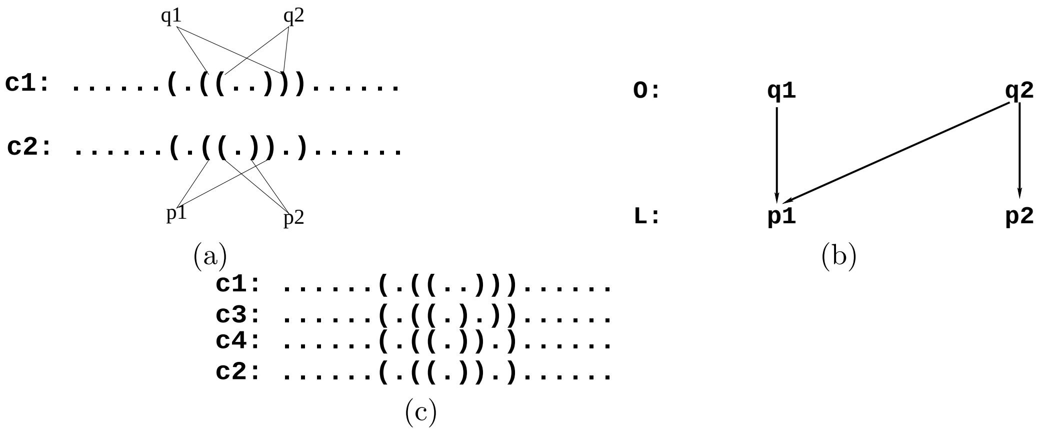


Figure 3: Intermediate node generation. (a) Start and goal configurations and contact pairs to be opened and closed: q1,q1 are in O; p1,p2 are in L. (b) Dependency graph: p1 depends on q1 and q2, p2 depends on q2. (c) Sequences generated: c3 and c4 are the two intermediate configurations to connect c1 and c2, here c4 happens to be identical to c2.

Each path is a sequence of conformational changes the RNA molecule goes through as it folds from one conformation to another. Each roadmap node on the path represents one such intermediate conformation in this folding process. Each roadmap edge on the path also represents a conformational change and is itself a path which consists of a number of intermediate configurations. Finding a path involves finding these intermediate nodes. The objective is to find those intermediate nodes that have a low energy.

Each of the intermediate nodes should have as many contacts as possible to keep its energy low. Once we open a base pair, we will close all base pairs whose contact does not introduce a conflict with the three validity criteria.

First, we find the set $\mathcal{O}$ of base pairs to be opened and the set $\mathcal{L}$ of base pairs to be closed. These are found directly from the base-pair distance. If a contact pair exists in the source conformation, c_1, and not in the goal conformation, c_2, then it is an element of $\mathcal{O}$. If a contact exists in c_2 and not in c_1, then it is an element of $\mathcal{L}$. See Fig. 3 (a): q1,q2 are in O and p1, p2 are in L.

Second, we construct the dependency graph, G, between these two sets. Since each node in the path must have a valid secondary structure, G describes which contact pairs cannot exist together in a valid configuration. If one base pair contact $p \in \mathcal{L}$ conflicts with another base pair $q \in \mathcal{O}$, then p cannot be closed until q is opened. This means that p depends on q. Thus, we construct a bipartite dependency graph between the two sets. See Fig. 3 (b): p1 depends on q1, q2 while p2 depends on q2.

Finally, we must choose which order is best for opening the contacts. This is a crucial point, because our goal is to produce configurations with as many contacts as possible. It is normally better to open a contact that allows two contacts to close than to open one that allows only one contact to close. To accomplish this, we use a heuristic that roughly mirrors the physical properties of RNA folding. The elements in the set $\mathcal{L}$ are sorted according to the number of dependencies they have. We first choose the element with the smallest number of dependencies, open all the contacts it depends on, and close it together with all the other base pairs in $\mathcal{L}$ that can be closed. See Fig. 3 (c): c3, c4 are the two intermediate configurations generated for connection. This is repeated until both $\mathcal{O}$ and $\mathcal{L}$ are empty.

2.3.3 Edge weight evaluation

When two nodes q_1 and q_2 are connected by the local planner, the corresponding 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 conformations $\{q_1 = c_0, c_1, c_2, \ldots, c_{n-1}, c_n = q_2\}$ on the straight line in $\mathcal{C}$ 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} depends on the difference between their potential energies $\Delta E_i = E(c_{i+1}) - E(c_i)$.

$$P_i = \begin{cases} e^{\frac{-\Delta E_i}{kT}} & \text{if } \Delta E_i > 0 \\ 1 & \text{if } \Delta E_i \le 0 \end{cases} \tag{3}$$

This keeps the detailed balance between two adjacent states, and enables the weight of an edge to be computed by summing the logarithms of the probabilities for consecutive pairs of conformations in the sequence. (Negative logs are used since each $0 \le P_i \le 1$.)

$$w(q_1, q_2) = \sum_{i=0}^{n-1} -log(P_i), \tag{4}$$

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 in [10].

3 Conclusion

A new method to analyze the RNA folding energy landscape is proposed. It's a method named PRM which was originated in robotics and has been successful in protein folding. Here, first we discussed its feasibility. Then prototype methods for its implementation were described. Currently, we have implemented these methods and are now investigating methods to analyze the roadmap (RNA landscapes) we produce.

4 About the Authors

Bonnie B. Kirkpatrick is a junior at Montana State University, where she is majoring in Computer Science and minoring in French and Math. Last summer (2001), she divided her time between two research projects. The first was applications of semistructured databases to data obtained from experiments on crickets at the Center for Computational Biology with Dr. Gwen Jacobs. This work lead to her attending Supercomputing 2001, where she learned about the DMP program. The second project she worked on that summer was on protein structure determination with Dr. Brendan Mumey. This research involved taking data they had obtained from epitopes on the protein's surface and mapping it to the one dimensional residue sequence. She continued the protein structure research during the school year. The result is that she became a coauthor on a paper and enjoyed the humidity of Texas.

Guang Song and Xinyu Tang are PhD students in the Department of Computer Science at Texas A&M University, and Nancy Amato is an associate professor in the Department of Computer Science at Texas A&M University.

References

[1] Shi-Jie Chen and Ken A. Dill. Rna folding energy landscapes. *PNAS*, 97:646–651, 2000.

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

[3] Ivo L. Hofacker Jan Cupal 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.

[4] 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.

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

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

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

[8] Ruth Nussinov, George Piecznik, Jerrold R. Griggs, and Danel J. Kleitman. Algorithms for loop matching. *SIAM J. Appl. Math.*, 35:68–82, 1972.

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

[10] 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.

[11] 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.

[12] Stefan Wuchty. Suboptimal secondary structures of rna. *Master Thesis*, 1998.