Exploring Folding Landscapes with Motion Planning Techniques

Bonnie Kirkpatrick\(^2\), Xinyu Tang\(^1\), Shawna Thomas\(^1\), Dr. Nancy Amato\(^1\)

\(^1\) Texas A&M University
\(^2\) Montana State University
Outline

- Motivation: Biopolymers
- Goal: Folding Landscapes
- Method: Motion Planning
- Application: RNA Folding
Outline

- Motivation: Biopolymers
- Goal: Folding Landscapes
- Method: Motion Planning
- Application: RNA Folding
What is Ribonucleic Acid (RNA)?

- Is composed of a sequence of nucleotides
- Folds in 3D into energetically optimal conformations
- Is essential to the process of carrying out a gene functions in cells.
- Performs specific functions including protein synthesis, acting as catalysts, and splicing introns, and regulating activities
- The folding behaviors of the molecule tell us much about their structure and function.
Ribonucleic Acid (RNA) Molecules

- **Primary Structure**
  - Sequence of bases
  - Each base is one of:
    - \{A, C, G, U\}
  - e.g. **ACGUGCCAUCG**
  - Obtained by experiment

- **Secondary Structure**
  - A 2D, planar representation

- **Tertiary Structure**
  - The sequence loops back on itself and **folds** in 3D.
RNA Molecules

- **Primary Structure**
  - Sequence of bases
  - Bases: A, C, G, U
  - e.g. `ACGUGCCAUCG`
  - Obtained by experiment

- **Secondary Structure**
  - A 2D, planar representation
  - Base Pair:
    - A-U
    - G-C
    - G-U

- **Tertiary Structure**
  - The sequence loops back on itself and **folds** in 3D.
RNA Conformations

- Chemical **bonds** (or **contacts**) form between complementary residues in close proximity.

- There are many possible **conformations** of the primary sequence.
  - e.g. CACAGAGUGU

- **Potential energy** calculations based on number and types of bonds are used to classify conformations.
  - The stable, low-energy conformation is known as the **native structure**.
  - Conformations with few bonds and high energy are referred to as **unfolded**.
Planar Representations

- Bonds between base pairs are lines or parentheses

All representations are equivalent
Planar Representations

- Bonds between base pairs are lines or parentheses

All representations are equivalent
Representations (cont.)

- Contact Map
- A dot is placed in the $i^{th}$ row and $j^{th}$ column of a triangular array to represent the intra-chain contact $[i, j]$
Secondary Structure Formalized

- A secondary structure conformation is specified by a set of intra-chain contacts (bonds between base pairs) that follow certain rules.

- Given any two intra-chain contacts \([i, j]\) with \(i < j\) and \([k, l]\) with \(k < l\), then:
  1) If \(i = k\), then \(j = l\)
     - Each base can appear in only one contact pair
  2) If \(k < j\), then \(i < k < l < j\)
     - No pseudo-knots

Violates criteria (1)

Violates criteria (2)
Secondary Structure Summarized

- 2D representation of the tertiary structure
- Planar representation
- Nested pairs
- Sufficient structural information

- Pseudo knots are considered a tertiary structure, rather than a secondary structure
Outline

- Motivation: Biopolymers
- Goal: Folding Landscapes
- Method: Motion Planning
- Application: RNA Folding
Folding Landscapes

- A “grand challenge” problem in biology
- Study the **kinetics** of folding
- Each RNA has a unique folding landscape
- Assumption:
  Native state $\leftrightarrow$ Lowest energy conformation
- Different from the structure prediction problem
  - Prediction of the native conformation
The Folding Process (a.k.a. the black box)

Unfolded Conformation (high energy)

AGGCUCACUUGGGAGCCUUCUCC

Physical Laws cause folding

Native Conformation (low energy)
Folding Landscapes

- Description of the “black box”
- A space in which every point corresponds to a conformation (or set of conformations) and its associated potential energy value (\textit{C-space}).
- A \textit{complete} folding landscape contains a point for every possible conformation of a given sequence.

\textbf{Tetrahymena Ribozyme Landscape}
[Russell, Zhuang, Babcock, Millett, Doniach, Chu, and Herschlag, 2002]
Folding Landscapes (cont.)

- **Conformational changes** describe how a molecule changes physically to fold from one conformation to another
  - Discrete
    - RNA Folding Model
    - Bonds either exist or do not exist
Features of Folding Landscapes

- **Folding pathways** consist of the set of conformational changes a molecule is likely to fold though when moving from one conformation to another.
  - N to X to Y

- **Energy barriers** are areas of the landscape with high energy that separate groups of conformations.
  - Y is separated from X and N

- **Intermediate states** are conformations lying on the folding pathway represent local minimums of potential.
  - Y and X

Mutant α mRNA fragment [Chen and Dill, 2000]
An RNA Folding Pathway

Phenylalanine tRNA [Hofacker, 1998]
Mapping Folding Landscapes

- Existing techniques for mapping landscapes are limited to relatively short sequences (~200 nucleotides).

- A robotics motion planning technique called PRM has successfully been applied to protein folding.
Outline

- Motivation: Biopolymers
- Goal: Folding Landscapes
- Method: Motion Planning
- Application: RNA Folding
(Basic) Motion Planning (in a nutshell):

Given a movable object, find a sequence of valid configurations that moves the object from the start to the goal.

Motion Planning for Foldable Objects:

Given a foldable object, find a valid folding sequence that transforms the object from one folded state to another.
Probabilistic Roadmap Method (PRM): Robotics
Native state

Construct the roadmap:
1. Generate nodes.
2. Connect to form roadmap

The Roadmap is like a net being laid down on the RNA’s potential landscape.

Now the roadmap can be used:
1. To extract multiple paths
2. To compute population kinetics
Outline

- Motivation: Biopolymers
- Goal: Folding Landscapes
- Method: Motion Planning
- Application: RNA Folding
Outline

- Conformation Space
- Node Generation
- Node Evaluation
- Node Connection
- Edge Weights
PRM: Conformation Space

- C is less than the set of every possible combination of contact pairs.
  \[ |U| \leq 2^n \]
  Where \( n \) is the number of possible contact pairs

- C contains only valid secondary structures.

- C-space is very large.
  - Sequence: (ACGU)\(_{10}\)
  - Length: 40 nucleotides
  - C-Space: 1.6x10\(^8\) structures
  - Smaller than the conformation space for protein folding.

- **Purpose:** generate a roadmap in C-space that describe the space without covering it
PRM: Node Generation

- Enumeration of C-space
  - Only feasible for small RNA

- Enumeration of stack-based conformations
  - A stack is any two base pairs that occur sequentially in the secondary structure:

    . ( ( . . . . . ) )

- Maximal pair random sampling
  - Every generated node has the maximal number of contacts possible
PRM: Node Generation

- Random Node Generation Algorithm
  - Starting with an empty configuration, $c$, random contacts are added to $c$ one at a time.
  - Each step preserves the condition that $c$ contains a valid set of base pair contacts.
  - Contacts are added until there are no more contacts that do not conflict with the contact set of $c$.

- Every node generated has valid secondary structure and is a member of $C$-space.

- Since every generated node has the maximal number of contacts, the sampling is biased toward the area of $C$-space near the native state.
PRM: Collision Detection

- Evaluation of Nodes
  - Potential energy determines how good a node is.
  - Only add a node to the roadmap if it has a low energy.
  - Probability of a node $q$ being added to the roadmap:

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

$E_{\min}$ and $E_{\max}$ are parameters of the method.
PRM: Node Connection

- Choosing nodes to connect
  - K-closest
  - Fixed Radius

- Distance Metric: Base Pair Distance
  - The number of contact pairs that differ between the two conformations
  - This is the number of contacts that must be opened or closed to move from one conformation to the next

```
.. ( . . ( . ) . . ) ..
.. ( . ( ( . ) ) . ) ..
```

```
.. ( . ( ( . ) ) . ) ..
```
Given two nodes in C-Space, $C_1$ and $C_2$, find a path between them consisting of a sequence of nodes:
\[
\{ C_1 = S_1, S_2, \ldots, S_{n-1}, S_n = C_2 \}
\]

The path must have the property that for each $i, 1 < i < n$, the set of contact pair of $S_i$ differs from that of $S_{i-1}$ by the application of one transformation operation:

1. open or
2. close

a single contact pair.

Discrete move-set.
There exists a path between any two nodes in C-Space.

Not just any path will do; we want a good one.

Bad paths have high energy nodes in them.

How do we find the **lowest energy** path?
Connection Algorithm

- more contacts $\Leftrightarrow$ less potential energy
- **Heuristic:** if a contact is opened by the transition from one node to another, try to close a contact in the next transition

<table>
<thead>
<tr>
<th>Transition</th>
<th>Bases Involved</th>
</tr>
</thead>
<tbody>
<tr>
<td>c1 = s1:  (\ldots((\ldots))\ldots)</td>
<td>open</td>
</tr>
<tr>
<td>s2: (\ldots(\ldots)\ldots)</td>
<td>close</td>
</tr>
<tr>
<td>s3: (\ldots((\ldots))\ldots)</td>
<td>open</td>
</tr>
<tr>
<td>s4: (\ldots(\ldots)\ldots)</td>
<td>close</td>
</tr>
<tr>
<td>C2 = s5: (\ldots((\ldots))\ldots)</td>
<td></td>
</tr>
</tbody>
</table>

Bases involved in the change are marked in red.
Connection Algorithm

Start
\[ c_1: \ldots (\ldots) \ldots \]

End
\[ c_2: \ldots (\ldots) \ldots \]

(a)

Dependency Graph

Open
\[ 0: \]

Close
\[ L: \]

(b)
Edge Weight

- Weights indicate the energetic feasibility of the edge.

- $\Delta E_i = E(s_{i+1}) - E(s_i)$

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

\[
w(q_1, q_2) = \sum_{i=0}^{n-1} -\log(P_i),
\]
Roadmap Analysis

- What does the roadmap tell us?
- Folding Pathways
Roadmap Analysis

- Population Kinetics
Roadmap Analysis

- Population Kinetics
Roadmap Analysis

- Population Kinetics

A conformation
Roadmap Analysis

- Population Kinetics
Roadmap Analysis

- Population Kinetics
Roadmap Analysis

- Population Kinetics
Roadmap Analysis

- Population Kinetics
  - Solved using a differential equation
Future Work

- **Validation**
  - Compare other statistical mechanical models
  - Compare with experimental results
  - Compare our sampled landscapes with complete landscapes
- **Explore the limits of our model**
- **Try different sampling methods**
- **Experiment with different distance metrics**
References

Ivo L. Hofacker Jan Cupal and Peter F. Stadler. Dynamic programming algorithm for the density
L. Kavraki, P. Svestka, J. C. Latombe, and M. Overmars. Probabilistic roadmaps for path
580, August 1996.
R. Li and C. Woodward. The hydrogen exchange core and protein folding. Protein Sci., 8:1571-
1591, 1999.
John S. McCaskill. The equilibrium partition function and base pair binding probabilities for rna
Ruth Nussinov, George Pieczenik, Jerrold R. Griggs, and Danel J. Kleitman. Algorithms for loop
D. Sanko and J.B. Kruskal. Time warps, string edits and macromolecules: the theory and
A.P. Singh, J.C. Latombe, and D.L. Brutlag. A motion planning approach to exible ligand
binding. In 7th Int. Conf. on Intelligent Systems for Molecular Biology (ISMB), pages 252-
261, 1999.