Efficient and Approximate Simulation Algorithm of Kinetic Folding of an RNA Molecule

Takumi Tanigawa and Satoshi Kobayashi
Department of Computer Scinece
University of Electro-Communications
1-5-1, Chofugaoka, Chofu, Tokyo 182-8585, Japan

Abstract— *Recently it is recognized as a very important research topic to simulate kinetic folding of an RNA molecule in order to understand its functionality in vivo. In this paper, we will propose a new approach to simulating kinetic folding of an RNA molecule based on a new idea of "enumerating secondary structures by a graph." Although most of the previous works try to reduce the conformation space of a given RNA molecule in order to escape from the combinatorial explosion problem, the present paper gives us an efficient and approximate simulation methodology for hairpin formation with* keeping the conformation space completely. *As far as the authors' knowledge, this is the first polynomial update time simulation algorithm for kinetic folding analysis of an RNA molecule which has a nice theoretical property that the convergence point of its simulation always* exactly coincides with the equilibrium distribution *of secondary structures of the RNA molecule. We evaluated the time efficiency and the accuracy of the proposed method against the exhaustive method which numerically simulates the master equation by completely generating all secondary structures. The results show that the proposed method is much faster than the exhaustive method and that the proposed method gives us well approximated simulation results.*

Keywords: Kinetic Folding, Simulation, RNA, Equilibrium Computation

1. Introduction

RNA secondary structure plays important role in the biological function of many RNAs. Thus, the prediction of RNA structure is an important research topic in bioinformatics. One of the most effective method for such prediction is to use dynamic programming (DP) to obtain a minimum free energy (MFE, for short) structure ([15] [28] [21]). DP method is extendedly applied also to the calculation of equilibrium structure ensembles of RNA secondary structures([11]). These algorithms, however, can deal with only thermodynamical equilibrium, and not with kinetic effects on secondary structures (for instance, during the synthesis of RNA molecules). Furthermore, although stacking free energy of 5 base pairs is around 10 kcal/mol at 300 K, thermal energy kT is only 0.6 kcal/mol at 300 K, which implies that a native RNA may easily be trapped into a suboptimal structure. Thus, the analysis of kinetic folding process of RNA molecules is very important for understanding their biological functions([13]).

A kinetical approach to RNA secondary structure prediction was introduced by Martinez ([10]), where folding kinetics is modeled by a Monte Carlo construction of secondary structures based on rate constants for iterative addition of complete helical regions, called helices, to some already existing structure. Modeling structure change by addition or deletion of helices is effective in reducing the conformation space of the RNA, thus, there are many works on RNA folding kinetics based on this formulation ([12], [1], [5], [13], [4], [7], [25]). However, the physical relevance of such moves seems debatable, because they cause large structural change per time step([3]). Furthermore, for a longer RNA sequence, we can not escape from the combinatorial explosion of conformation space even if we use helix based formulation.

Schmitz and Steger proposed a simulation method for kinetic folding of RNA secondary structures by using a Monte Carlo method based on rate constants for adding or removing a single base pair to some already existing structure([16]). The proposed move set is much more accurate than the helix based move set and was supported by many researchers ([3], [26], [18], [23], [14]), but, the combinatorial explosion problem of conformation space is more severe than the helix based approach.

In this paper, we will give a novel approach to simulating kinetic folding of an RNA molecule based on an elegant new idea of "enumerating secondary structures by a graph." Although most of the previous works try to reduce the conformation space of a given RNA molecule in order to escape from the combinatorial explosion problem, the present paper will provide us with an efficient and approximate simulation methodology for hairpin formation with *keeping the conformation space completely*.

As far as the authors' knowledge, this is the first polynomial update time approximate simulation algorithm for kinetic folding of an RNA molecule which has a nice theoretical property that the convergence point of its simulation always *exactly coincides with the equilibrium distribution* of secondary structures of the RNA molecule. Although we focus on secondary structures which are pseudoknot-free and multiloop-free, the developed system is of great importance since the folding of complex RNA tertiary structures often involves the conformational change of hairpin structures ([2] [20] [24] [19] [17] [22]) and the detailed kinetical analysis of hairpin formation still have many research topics to be studied([27]).

2. Problem Definition

2.1 Secondary Structures

Let $X = x_1 x_2 \cdots x_n$ be an RNA sequence with each letter x_i being an element of $\Sigma = \{A, C, G, U\}$, ordered from $5'$ to $3'$ direction. It is known that every pair of bases in $WC = \{(A, U), (U, A), (C, G), (G, C), (G, U), (U, G)\}$ may form a hydrogen bond, resulting in a stable structure, called secondary structure. A secondary structure of X is a finite set S of pairs (i, j) of integers such that $1 \leq i < j \leq n$ and $(x_i, x_j) \in WC$ hold and for any $bp_1 = (i_1, j_1)$ and $bp_2 = (i_2, j_2)$ in S either $i_1 = i_2$ or $j_1 = j_2$ implies $bp_1 = bp_2$. A secondary structure S is said to be *pseudoknotted* if there exist base pairs (i, j) and (k, l) such that $i < k < j < l$ (Figure 1 (a)). A secondary structure S is said to be *pseudoknot-free* if it is not pseudoknotted. Although there are some experimental reports on structural roles of pseudoknotted structures in biological functions, the computational analysis of secondary structures including them is time consuming ([21]) and thus it is often the case that we focus on pseudoknot-free structures. Furthermore, from the view point of RNA folding kinetics theory, the detailed study and analysis on the hairpin formation is still of great importance([26], [27]). Thus, in this paper, we will focus on the class of secondary structures which are pseudoknot-free and multiloop-free, where a structure is said to contain a multiloop if there exist base pairs (i, j), (i_1, j_1) and (i_2, j_2) such that $i < i_1 < j_1 < i_2 < j_2 < j$ (Figure 1 (b)). A typical example of pseudoknot-free and multiloop-free structures are given in Figure 1 (c). As is shown in Figure 1 (c), such a structure can be explained as a sequence of linear structures concatenated in parallel, where by a linear structure, we mean a secondary structure consisting of a sequence of base pairs $(i_1, j_1), ..., (i_k, j_k)$ such that $i_p < i_{p+1} < j_{p+1} < j_p$ holds for every $p = 1, ..., k-1$.

2.2 Move Set

Let $C(X) = \{S_1, ..., S_m\}$ be the set of all secondary structures of X, i.e., a conformation space of X. As a first step of a novel efficient simulation methodology of kinetic folding of an RNA molecule, we do restrict our attention to multiloop-free and pseudoknot-free structures. Structures of $C(X)$ are related by a relation, called a "move", which defines a transition path of kinetic folding of the RNA sequence X. Two kinds of moves, *Add* and *Delete* are considered in this paper. The former modifies a secondary structure by adding a new base pair in compliance with no-multiple-loop and no-pseudoknot restrictions(Figure 2 (a)), and the latter removes a base pair in the structure (Figure 2 (b)). At every moment, a structure $S_i \in C(X)$ will change its structure to another one by choosing a move according to some probability distribution from a pool of acceptable moves. Successive choices of such elementary moves generates a folding process(Figure 2 (c)). For a structure S_i, by $Nbr(S_i)$, we denote the set of structures which can be obtained by applying an elementary move, *Add* or *Delete*, to S_i.

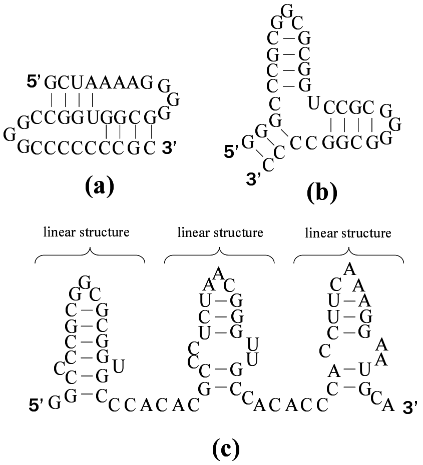


Fig. 1: Secondary Structures

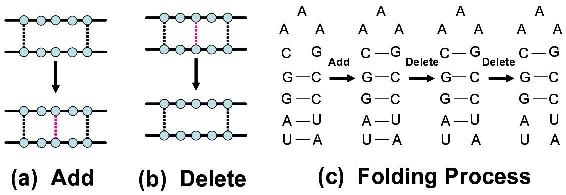


Fig. 2: Elementary Moves

2.3 Folding Kinetics

By G^0_i, we denote the Gibbs free energy change when the sequence folds into the structure S_i from its random chain structure. Then, the rate constant $k_{i,j}$ of the transition from S_i to S_j is given by the following Metropolis rule:

$$k_{i,j} = \begin{cases} e^{-\frac{G^0_j - G^0_i}{RT}} & \text{if } G^0_j - G^0_i > 0 \text{ and } S_j \in Nbr(S_i) \\ 1 & \text{if } G^0_j - G^0_i \leq 0 \text{ and } S_j \in Nbr(S_i) \\ 0 & \text{if } S_j \notin Nbr(S_i). \end{cases}$$

Let $P_i(t)$ be the fraction (or probability) of the structure S_i at time t. Then, the population dynamics of the folding process of an RNA sequence follows the master equation:

$$\frac{\mathrm{d}P_i(t)}{\mathrm{d}t} = k_{cal} \sum_{j=1}^{m} (P_j(t) k_{j,i} - P_i(t) k_{i,j}), \qquad (1)$$

where k_{cal} is a calibration constant for adjusting simulation results with experimental results. We use the value $k_{cal} = 3.34 \times 10^6$, which was used in [16].

Most of the previous works have tried to reduce the conformation space in order to escape from its combinatorial explosion problem. In this work, however, we *completely* keep the space $C(X)$ and try to numerically simulate the master equation *efficiently and approximately* with the theoretical guarantee that the results will always reach to the *exact equilibria*. In order to achieve this goal, we will apply our previous theoretical work on the equilibrium analysis of chemical reaction systems in which molecules are interacting in various ways to generate tremendously many structures([8], [9]). Although the present paper deals with unimolecular reaction, the theory applies since the unimolecular reaction can be treated as a special case of the framework. In the next section, we will review the theory specialized to unimolecular reaction systems.

3. Enumeration Approach to Equilibria Analysis

Let X be a molecule and $C(X)$ be the conformation space, i.e., the set of structures, of X. Free energy of a conformation S of X is given by $F(S)$.

Assume that we have a *directed* graph $G = (V, Eg)$ with a finite set V of vertices and a finite set Eg of *directed* edges. For a vertex $v \in V$, by v_{in} and v_{out}, we denote the set of edges coming into v and going out from v, respectively. A vertex v with $v_{in} = \emptyset$ (with $v_{out} = \emptyset$, respectively) is called an *initial vertex* (a *final vertex*, respectively). By V_0 and V_f, we denote the set of initial and final vertices of G, respectively. A *simple* path of G is a path with each vertex appearing at most once. By $PT(G)$, we denote the set of simple paths starting from some vertex in V_0 and reaching to some vertex in V_f.

The essential part of the theory depends on the existence of a special *one-to-one* mapping ψ from $PT(G)$ to $C(X)$ satisfying the conditions explained bellow. After constructing such a mapping, the theory reduces the problem of computing equilibrium state to a convex optimization problem with respect to a set unknown variables whose size is $|Eg|$. Note that the cardinality of $PT(G)$ could be exponential with respect to $|Eg|$. Thus, the theory enables us to escape from the combinatorial explosion problem of the conformation space $C(X)$.

The requirement for the mapping ψ is *very simple* as follows. We ask the existence of a weight function ϵ on the edge set Eg such that for every path $\gamma \in PT(G)$, $F(\psi(\gamma)) = \sum_{e \in Eg \text{ s.t. } e \in \gamma} \epsilon(e)$ holds. This condition means that for every $\gamma \in PT(G)$, the sum of weight of edges appearing in γ equals to the free energy of the corresponding structure $\psi(\gamma)$ of the path γ. Intuitively speaking, every edge in the graph G corresponds to some local structure of conformation space, and its weight is just the free energy of the corresponding local structure. In case of equilibrium analysis of an RNA molecule at the secondary structure level, it would be expected that we can construct a graph whose edge would correspond to local structures, such as hairpin loops, bulge loops, internal loops, etc. An example of such enumeration graphs will be given in the next section 4.

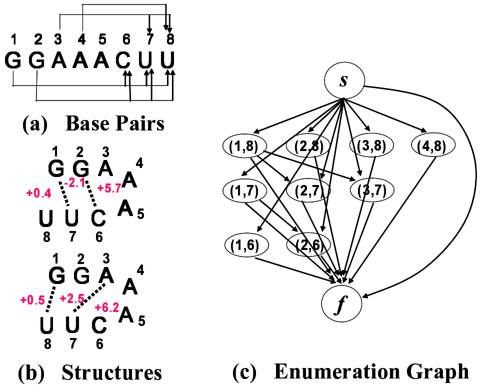
4. Enumerating Secondary Structures of an RNA

We will give an example of graphs by which we can enumerate all linear secondary structures of an RNA sequence.

Let $X = x_1 \cdots x_n$ be an RNA sequence. Then, we prepare a set of vertices corresponding to base pairs which may form in the sequence X. Moreover, we use two additional special vertices: an initial vertex s and a final vertex f. The construction of edge set is as follows. We draw an edge from a base pair (i, j) to a base pair (k, l) if and only if $i < k < l < j$ holds. Furthermore, for every base pair bp, we put an edge from s to bp and an edge from bp to f. Formally, we can define a graph $G = (V, Eg)$ for the sequence X:

$$\begin{aligned}
BP &= \{(i,j) \mid 1 \le i < j \le n,\ (x_i, x_j) \in WC\}, \\
V &= \{s, f\} \cup BP, \\
Eg &= \{(s, bp) \mid bp \in BP\} \cup \{(bp, f) \mid bp \in BP\} \cup \\
&\quad \{((i,j),(k,l)) \mid (i,j),(k,l) \in BP,\ i < k < l < j\}.
\end{aligned}$$

A path in $PT(G)$ for G defined above naturally corresponds to a linear secondary structure consisting of base pairs contained in it. An example of graphs for enumerating secondary structures of the sequence $X =$ GGAAACUU is given in Figure 3.



(a) Base Pairs

(b) Structures

(c) Enumeration Graph

Fig. 3: An Example of Enumeration Graphs

Figure 3 (a) illustrates all possible base pairs of the sequence X. Figure 3 (c) shows an enumeration graph for the sequence X. A path $s \to (1,7) \to (2,6) \to f$ corresponds to the upper secondary structure in Figure 3 (b). A path $s \to$

$(1,8) \to (3,7) \to f$ corresponds to the lower secondary structure in Figure 3 (b). In this way, we can enumerate all linear secondary structures of X. The mapping from a path to its corresponding secondary structure is denoted by ψ.

As is clear from the above example, an edge between the base pairs (i,j) and (k,l) in the graph corresponds to a local loop structure (either of stacked base pairs, a bulge, or an internal loop) surrounded by (i,j) and (k,l). An edge between s (f) and a base pair (i,j) corresponds to a free end loop outside (a hairpin loop closed by) the base pair (i,j). Thus, the weight $\epsilon(e)$ of an edge e is defined as the free energy of the corresponding local secondary structure of e. For instance, the free energy values of local secondary structures are given as real values in Figure 3 (c). Thus, the weight of edges $s \to (1,7)$, $(1,7) \to (2,6)$, $(2,6) \to f$, $s \to (1,8)$, $(1,8) \to (3,7)$, $(3,7) \to f$ are given by $+0.4$, -2.1, $+5,7$, $+0.5$, $+2.5$, $+6.2$, respectively.

5. Efficiently Computing Equilibria by Convex Programming

Let X be a molecule and $C(X)$ be a conformation space of X. An *equilibrium distribution* of $C(X)$ is a probability distribution $[\,]$ over $C(X)$ such that for any conformations S_1 and S_2 in $C(X)$, the following equilibrium equation holds:

$$\frac{[S_2]}{[S_1]} = e^{-\frac{F(S_2)-F(S_1)}{RT}}.$$

When we succeed in constructing an enumeration graph G for a conformation space $C(X)$ of a molecule X satisfying the conditions explained in section 3, following a general theory developed by the second author of this paper([8], [9]), we can efficiently compute an equilibrium distribution by solving a minimization problem explained bellow.

First, we will introduce an unknown variable w_e for each edge e of the graph G. The variable w_e takes a real value between 0 and 1, and represents a probability of the local substructure corresponding to e existing in the current probability distribution over $C(X)$.

For convenience, for every $v \in V - V_0 - V_f$, we define $w_v = \sum_{e \in v_{out}} w_e$. Consider the following minimization problem:

Minimization Problem P1
minimize :

$$FE((w_e \mid e \in Eg)) \overset{def}{\equiv} \sum_{e \in Eg} \frac{\epsilon(e)}{RT} \cdot w_e + \sum_{e \in Eg} w_e(\log w_e - 1) - \sum_{v \in V - V_0 - V_f} w_v(\log w_v - 1)$$

subject to :

$$\begin{aligned} \sum_{v \in V_0} \sum_{e \in v_{out}} w_e &= 1, & \\ \sum_{e \in v_{in}} w_e &= \sum_{e \in v_{out}} w_e, & (\forall v \in V - V_0 - V_f\}) \\ w_e &\geq 0, & (\forall e \in Eg) \end{aligned}$$

where unknown variables are w_e's $(e \in Eg)$ and recall that w_v's are sums of variables w_e's.

Then, the following theorem was proved in [8]:

Theorem 1: Consider a minimizer $(w_e \mid e \in Eg)$ of the above minimization problem *P1*. Then, an equilibrium distribution is given by: for any $S \in C(X)$,

$$[S] = \frac{\displaystyle\prod_{e \in Eg \text{ s.t. } e \in \psi^{-1}(S)} w_e}{\displaystyle\prod_{v \in V - V_0 - V_f \text{ s.t. } v \in \psi^{-1}(S)} w_v}, \qquad (2)$$

In order to obtain an equilibrium distribution of an RNA molecule at the secondary structure level, we should first obtain a minimizer of the optimization problem *P1* based on the graph G given in section 4. This is achieved efficiently since the objective function of the problem *P1* is convex as shown in [8], and thus, we can apply a convex programming method to obtain a minimizer. Equilibrium distribution is then obtained by the expression (2).

6. The Objective of This Work

In this way, we will be able to efficiently compute an equilibrium distribution of an RNA molecule. This is not, however, the main purpose of this paper. In this work, we aim at efficiently simulating kinetic folding process specified by the master equation (1). Thus, applying convex programming method might not lead us to the goal of this paper. We need to carefully choose a decending direction of the objective function of the optimization problem *P1*. Such a careful choice of the direction will be proposed in section 7, and the validity of the choice will be shown theoretically in section 8. This theoretical argument guarantees that the proposed simulation algorithm will always converge to an equilibrium distribution. Our method is distinguished from the others in the *convergence property to equilibria*.

7. Algorithm

In this sectin, we will give an algorithm for efficiently and approximately simulating the kinetic folding process of an RNA molecule at the secondary structure level. The algorithm is presented with intuitive explanation of the reason why we will obtain the algorithm. The key idea behind the algorithm is to *locally interpret in the graph representation the kinetic moves of Add and Delete*.

We first explain how to interpret the move *Add* in view of enumeration graph (Figure 4). The move *Add* inserts a

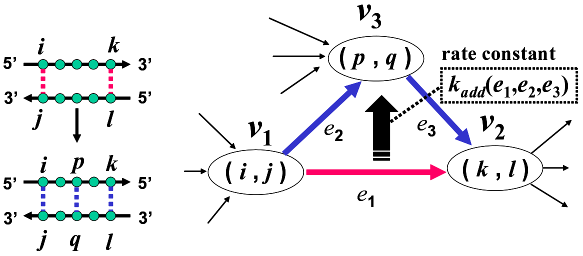


Fig. 4: Local Interpretation of *Add* Move

new base pair to the current conformation. Consider an *Add* move which inserts a base pair (p,q) between two base pairs (i,j) and (k,l). Note that $i < p < k < l < q < j$ holds. This move can be interpreted in the enumeration graph representation as a move from a path containing the edge $(i,j) \to (k,l)$ to another path containing the edges $(i,j) \to (p,q)$ and $(p,q) \to (k,l)$ keeping the probabilities of the other edges unchanged (See Figure 4). Based on this observation, it is very natural to interpret the *Add* operation locally as the change of probabilities of the three edges, $(i,j) \to (k,l)$, $(i,j) \to (p,q)$, and $(p,q) \to (k,l)$, as follows:

$$\begin{aligned} \Delta w_{e_1} &= -k_{add}(e_1,e_2,e_3)\cdot w_{e_1}\Delta t, && (3)\\ \Delta w_{e_2} &= k_{add}(e_1,e_2,e_3)\cdot w_{e_1}\Delta t, && (4)\\ \Delta w_{e_3} &= k_{add}(e_1,e_2,e_3)\cdot w_{e_1}\Delta t, && (5) \end{aligned}$$

where $k_{add}(e_1,e_2,e_3)$ is a rate constant for this local reaction which causes the change of probabilities of the edges e_1, e_2 and e_3, which is defined by:

$$k_{add}(e_1,e_2,e_3) = \begin{cases} e^{-\frac{\epsilon(e_2)+\epsilon(e_3)-\epsilon(e_1)}{RT}} & \text{if } \epsilon(e_2)+\epsilon(e_3)-\epsilon(e_1) \ge 0 \\ 1 & \text{otherwise} \end{cases}$$

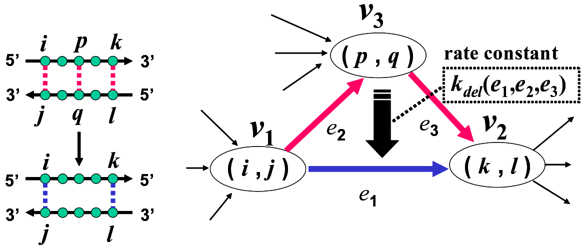


Fig. 5: Local Interpretation of *Delete* Move

Next we consider the case of the move *Delete*. The move *Delete* removes a base pair from the current conformation. Consider a *Delete* move which removes a base pair (p,q) from between two base pairs (i,j) and (k,l). Note that $i < p < k < l < q < j$ holds also in this case. This move can be interpreted in the enumeration graph representation as a move from a path containing the edges $(i,j) \to (p,q)$ and $(p,q) \to (k,l)$ to another path containing the edge $(i,j) \to (k,l)$ keeping the probabilities of the other edges unchanged (See Figure 5). Thus, it is natural to interpret the *Delete* operation locally as the change of probabilities of the three edges, $(i,j) \to (k,l)$, $(i,j) \to (p,q)$, and $(p,q) \to (k,l)$, as follows:

$$\begin{aligned} \Delta w_{e_1} &= k_{del}(e_1,e_2,e_3)\cdot w(e_2,e_3)\Delta t, && (6)\\ \Delta w_{e_2} &= -k_{del}(e_1,e_2,e_3)\cdot w(e_2,e_3)\Delta t, && (7)\\ \Delta w_{e_3} &= -k_{del}(e_1,e_2,e_3)\cdot w(e_2,e_3)\Delta t, && (8) \end{aligned}$$

where $w(e_2,e_3)$ is the probability of the paths passing through both of the edges e_2 and e_3, and $k_{del}(e_1,e_2,e_3)$ is a rate constant for this local reaction which causes the change of probabilities of the edges e_1, e_2 and e_3, which is defined by:

$$k_{del}(e_1,e_2,e_3) = \begin{cases} e^{-\frac{\epsilon(e_1)-\epsilon(e_2)-\epsilon(e_3)}{RT}} & \text{if } \epsilon(e_1)-\epsilon(e_2)-\epsilon(e_3) \ge 0 \\ 1 & \text{otherwise} \end{cases}$$

Note that we still have difficulty in this local interpretation of the move *Delete*, because we do not have any information about the probability $w(e_2,e_3)$. We only know about probabilities of edges e_2 and e_3 as w_{e_2} and w_{e_3}, respectively. So, we should approximately guess the probability $w(e_2,e_3)$ of paths passing through both of edges e_2 and e_3. We will propose to use the following estimate:

$$w(e_2,e_3) = \begin{cases} \frac{w_{e_2}\cdot w_{e_3}}{w_{v_3}}, & \text{if } w_{v_3} > 0 \\ 0 & \text{otherwise} \end{cases}$$

where v_3 is the vertex corresponding to the base pair (p,q). This expression intuitively means that every path coming to the vertex v_3 is splitted into all directions from v_3 proportionally to the probability distribution of edges going out from v_3. This estimate theoretically guarantees that the proposed method will always reach to an equilibrium distribution at the convergence point, as will be shown in the next section 8.

The proposed method will apply the above rule of local probability change to every triple of edges located in a triangular form as illustrated in Figure 4 and Figure 5. It is clear that the time complexity of the update of probabilities of all w_e's ($e \in Eg$) is bounded by a polynomial function with respect to the length of the sequence X.

8. Theoretical Analysis of the Algorithm

Theorem 2: The direction specified by the expressions (3)-(8) is a decending direction of the optimization problem *P1*. *Proof:* For every triangle consisting of edges e_1, e_2, e_3 and vertices v_1, v_2, v_3 as illustrated in Figure 4 and

5, we have:

$$\begin{aligned}\frac{\partial FE((w_e \mid e \in Eg))}{\partial w_{e_1}} &= \frac{\epsilon(e_1)}{RT} + \log w_{e_1} - \log w_{v_1}, \\ \frac{\partial FE((w_e \mid e \in Eg))}{\partial w_{e_2}} &= \frac{\epsilon(e_2)}{RT} + \log w_{e_2} - \log w_{v_1}, \\ \frac{\partial FE((w_e \mid e \in Eg))}{\partial w_{e_3}} &= \frac{\epsilon(e_3)}{RT} + \log w_{e_3} - \log w_{v_3},\end{aligned}$$

and the sum of the moves *Add* and *Delete*, denoted by (d_1, d_2, d_3), is given by:

$$\begin{aligned}d_1 &= -k_{add}(e_1, e_2, e_3) w_{e_1} + k_{del}(e_1, e_2, e_3) \frac{w_{e_2} w_{e_3}}{w_{v_1}}, \\ d_2 &= k_{add}(e_1, e_2, e_3) w_{e_1} - k_{del}(e_1, e_2, e_3) \frac{w_{e_2} w_{e_3}}{w_{v_1}}, \\ d_3 &= k_{add}(e_1, e_2, e_3) w_{e_1} - k_{del}(e_1, e_2, e_3) \frac{w_{e_2} w_{e_3}}{w_{v_1}}.\end{aligned}$$

Then, we have:

$$\begin{aligned}\sum_{i=1}^{3} \frac{\partial FE((w_e \mid e \in Eg))}{\partial w_{e_i}} \cdot d_i &= \\ & k_{add}(e_1, e_2, e_3)\, w_{e_1} (1 - \frac{k_{del}(e_1, e_2, e_3) w_{e_2} w_{e_3}}{k_{add}(e_1, e_2, e_3) w_{e_1} w_{v_3}}) \times \\ & (\log \frac{w_{e_2} w_{e_3}}{w_{e_1} w_{v_3}} e^{\frac{\epsilon(e_2)+\epsilon(e_3)-\epsilon(e_1)}{RT}}) = \\ & k_{add}(e_1, e_2, e_3)\, w_{e_1} (1 - \frac{w_{e_2} w_{e_3}}{w_{e_1} w_{v_3}} e^{\frac{\epsilon(e_2)+\epsilon(e_3)-\epsilon(e_1)}{RT}}) \times \\ & (\log \frac{w_{e_2} w_{e_3}}{w_{e_1} w_{v_3}} e^{\frac{\epsilon(e_2)+\epsilon(e_3)-\epsilon(e_1)}{RT}}) \le 0,\end{aligned}$$

where we use:

$$\frac{k_{del}(e_1, e_2, e_3)}{k_{add}(e_1, e_2, e_3)} = e^{\frac{\epsilon(e_2)+\epsilon(e_3)-\epsilon(e_1)}{RT}},$$

completing the proof.

Since the objective function of *P1* is convex, by Theorem 2, we can conclude that the simulation by the proposed method will reach to an equilibrium distribution.

9. Simulation Results

We have done two kinds of computational experiments. The first one is for evaluating the time efficiency of the proposed method against the exhaustive method, in which for an input sequence X, we generated all the secondary structures in $C(X)$, and simulated the folding kinetics of X based on the master equation (1). Simulations of both methods start from a random chain structure.

The other experiment is for showing that the proposed method gives us a well approximated simulation result for structures which are dominant at equilibrium.

In this section, we will give some computational experimental results which will show that the proposed method is very time efficient compared to the exhaustive method and it gives us a fairly well approximated kinetic folding pahts. The tested sequences are listed in Table 1.

No.	length	Sequence
1	10	AGCCGUUUCC
2	12	AACCCUACCCUU
3	14	GGGCGAAACGCCCU
4	16	GCCGCGAAACGCGGCC
5	18	CGGGCCGAAAUGGGCCCU
6	20	CGGGCGCGAAAUUCGCGCCC

Table 1: RNA Sequences

No.	N_{str}	T_E	T_P
1	15	0.27s	0.09s
2	14	0.27s	0.08s
3	200	7.34s	0.84s
4	322	12.99s	1.13s
5	832	38.57s	2.98s
6	3293	2m58.29s	9.40s

Table 2: Time Efficiency Result

For a given sequence X, we did kinetic simulations starting from a random chain structure by using the exhaustive method and the proposed method up to $1,000$ time steps, where we use $\Delta t = 1.0 \times 10^{-8}$ sec. The time for executing $1,000$ step simulation is given in Table 2, where T_E is for the exhaustive method and T_P for the proposed method. The number of structures in $C(X)$ is given in the column N_{str}.

In Fig.6 and Fig.7, we simulated dominant structures of the sequence ACGUGCACAAAAGUGCACGU of length 20. The optimal strcuture is ((((((((....)))))))) (-12.0 kcal/mol) and its simulation result is shown in Fig.6. Suboptimal structures are St1= (((((((......))))))) (-10.0 kcal/mol) and St2= ..((((((....)))))).. (-9.9 kcal/mol), and their simulation results are shown in Fig.7. In both of the figures, the lines specified by "E" and "P" represent the simulation results by the *E*xhaustive and the *P*roposed methods, respectively. Simulations by the proposed method give us well approximated results compared to the exhaustive (i.e., exact) simulations.

Concerning rare structures, the time step Δt should be carefully chosen as small values enough, since concentrations of rare structures are very sensitive to large Δt, which result in incorrect simulations in both of the exhaustive and the proposed methods. The topic on the choice of appropriate Δt would be a future research topic.

10. Conclusion

We proposed a novel method for efficiently and approximately numerically simulating kinetic folding process of an RNA molecule based on the idea of "enumerating conformations by a graph." The proposed method has a very nice theoretical property that the convergence point of simulation results exactly coincides with the equilibrium. Time efficiency, the accuracy and the effectiveness of the method were shown by computational experiments.

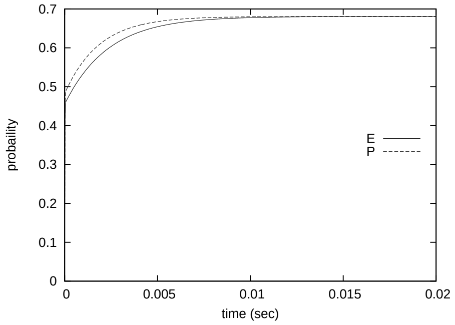


Fig. 6: Simulations of Optimal Structure

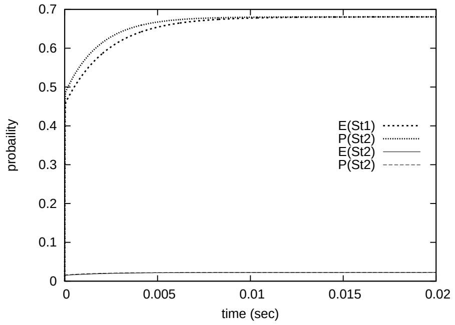


Fig. 7: Simulations of Suboptimal Structures

The current implementation is restricted only to the class of linear secondary structures, i.e. structures which do not contain branches. But, the proposed method can be extended to a more broader class of secondary structures if we prepare an appropriate enumeration graph for the extended structure class. Thus, it is an important future research topic to find such an enumeration scheme for a broader class of secondary structures.

In this paper, we evaluated the accuracy of the proposed simulation method only by computational experiments. Theoretical analysis of the accuracy of the method is also an important open problem. Furthermore, based on this kind of theoretical analysis, it might be interesting to improve the method in order to achieve a better accuracy.

Acknowledgement

The second author is supported in part by Grant-in-Aid for Scientific Research (C) No.22500010, Japan Society for the Promotion of Science.

References

[1] Abrahams, J.P., van den Berg, M., van Batenburg, E., and Pleij, C. (1990), Prediction of RNA secondary structure, including pseudoknotting, by computer simulation, *Nucleic Acids Research*, **18**, 3035-3044.

[2] Bartley, L.E., Zhuang, X., Das, R., Chu, S., and Herschlag, D. (2003), Exploration of the transition state for tertiary structure formation between an RNA helix and a large structure RNA, *J. Mol. Biol.*, **328**, 1011-1026.

[3] Flamm, C., Fontana, W., and Hofacker, I.L. (2000), RNA folding at elementary step resolution, *RNA*, **6**, 325-338.

[4] Galzitskaya, O.V., and Finkelstein, A.V. (1996), Computer simulation of secondary structure folding of random and "edited" RNA chains, *J. Chem. Phys.*, **105**, 319-325.

[5] Gultyaev, A.P., van Batenburg, F.H.D., and Pleij, C. (1995), The Computer Simulation of RNA Folding Pathways Using a Genetic Algorithm, *J. Mol. Biol.*, **250**, 37-51.

[6] Hofacker, I.L., Fontana, W, Stadler, P.F., Bonhoeffer, L.S., Tacker, M., Schuster, P. (1994), Fast folding and comparison of RNA secondary structures (the Vienna RNA package), *Monatshefte für Chemie*, **125**, 167-188.

[7] Isambert, H., and Siggia, E.D. (2000), Modeling RNA folding paths with pseudoknots: Application to hepatitis delta virus ribozyme, *PNAS*, **97**, 6515-6520.

[8] Kobayashi, S. (2007), A new approach to computing equilibrium state of combinatorial hybridization reaction systems, in *Proc. of Computing and Communications from Biological Systems: Theory and Applications*, Budapest, Hungary, CD-ROM, paper2376. (Extended full version is available at http://comp.cs.uec.ac.jp/˜satoshi/TR_CS0801rev.pdf)

[9] Kobayashi, S. (2008), A software tool for analyzing combinatorial hybridization reaction systems, in *Proc. of 14th International Meeting on DNA Based Computer*, Track B, oral presentation.

[10] Martinez, H.M. (1984), An RNA folding rule, *Nucleic Acids Research*, **12**, 323-334.

[11] McCaskill, J.S. (1990), The equilibrium partition function and base pair binding probabilities for RNA secondary structure, *Biopolymers*, **29**, 1105-1119.

[12] Mirnov, A.A., Dyakonova, L.P., and Kister, A.E. (1985), A kinetic approach to the prediction of RNA secondary structures, *Journal of Biomolecular Structure and Dynamics*, **2**, 953-962.

[13] Morgan, S.R., and Higgs, P.G. (1996), Evidence for kinetic effects in the folding of large RNA molecules, *J. Chem. Phys.*, **105**, 7152-7157.

[14] Ndifon, W. (2005), A complex adaptive systems approach to the kinetic folding of RNA, *BioSystems*, **82**, 257-265.

[15] Nussinov, R., Pieczenik, G., Griggs, J.R., and Kleitman, D.J. (1978), Algorithms for Loop Matchings, *SIAM J. Appl. Math.*, **35**, 68-82.

[16] Schmitz, M., and Steger, G. (1996), Description of RNA folding by "Simulated Annealing", *J. Mol. Biol.*, **255**, 254-266.

[17] Sosnick, T.R., and Pan, T. (2003), RNA folding:models and perspectives, *Curr. Opin. Struct. Biol.*, **13**, 309-316.

[18] Tang, X., Kirkpatrick, B., Thomas, S., Song, G., and Amato, N.M. (2004), Using motion planning to study RNA folding kinetics, in *Proc. of 8th Annual International Conference on Research in Computational Molecular Biology (RECOMB'04)*, 252-261.

[19] Thirumalai, D., Lee, N., Woodson, S.A., Klimov, D.K. (2001), Early events in RNA folding, *Annu. Rev. Phys. Chem.*, **52**, 751-762.

[20] Treiber, D.K., and Williamson, J.R. (2001), Beyond kinetic traps in RNA folding, *Curr. Opin. Struct. Biol.*, **11**, 309-314.

[21] Uemura, Y., Hasegawa, A., Kobayashi, S., and Yokomori, T. (1999), Tree Adjoining Grammars for RNA Structure Prediction, *Theoretical Computer Science*, **210**, 277-303.

[22] Uhlenbeck, O.C. (1990), Nucleic-acid structure — tetraloops and RNA folding, *Nature*, **346**, 613-614.

[23] Wolfinger, M.T., Svrek-Seiler, W.A., Flamm, C., Hofacker, I.L., and Stadler, P.F. (2004), Efficient computation of RNA folding dynamics, *J. Phys. A: Math. Gen.*, **37**, 4731-4741.

[24] Woodson, S.A. (2000), Recent insights on RNA folding mechanisms from catalytic RNA, *Cell. Mol. Life Sci.*, **57**, 796-808.

[25] Xayaphoummine, A., Bucher, T., and Isambert, H. (2005), Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots, *Nucleic Acids Research*, **33**, Web Serve issue, W605-W610.

[26] Zhang, W., and Chen, S.-J. (2002), RNA hairpin-folding kinetics, *PNAS*, **99**, 1931-1936.

[27] Zhang, W., and Chen, S.-J. (2006), Exploring the complex folding kinetics of RNA hairpins: I.: General folding kinetics analysis, *Biophysical Journal*, **90**, 765-777.

[28] Zuker, M., and Steigler, P. (1981), Optimal Computer Folding of Large RNA Sequences using Thermodynamics and Auxiliary Information, *Nucleic Acids Research*, **9**, 133-148.