社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2009－MPS－73 (24)
2009／3／5

RNA フォールディングシミュレーションのための新しいアルゴリズム

谷川 拓己　小林 聡
電気通信大学大学院 電気通信学研究科 情報工学専攻

RNA 分子のフォールディングシミュレーションを高速かつ精確に行うことの重要性が近年高まっている．本論文では，RNA 二次構造をグラフを用いて数え上げるという新しい発想を用いて，二次構造レベルで RNA のフォールディングシミュレーションを近似的かつ高速に行うアルゴリズムを提案する．従来の研究では，RNA 分子の二次構造に制限を与えることにより，構造空間の組合せ爆発を抑えようとするものがほとんどであるが，本研究では，構造空間を全く削減することなく効率良く近似的に RNA のヘアピン形成シミュレーションを行うためのアルゴリズムを与える．このアルゴリズムは，シミュレーションの収束点が平衡状態と一致するという理論的な性質を備えた最初の多項式更新時間シミュレーションアルゴリズムである．

A New Algorithm for Simulating Kinetic Folding of an RNA Molecule

Takumi Tanigawa and Satoshi Kobayashi
Dept. of Computer Science, Graduate School of Electro-Communications,
Univ. of Electro-Communications
e-mail:takumi1219@comp.cs.uec.ac.jp,satoshi@cs.uec.ac.jp

The need for efficiently and accurately simulating kinetic folding of an RNA molecule has become increasingly apparent over the decades. In this paper, we will give a novel approach to simulating kinetic folding of an RNA molecule at the secondary structure level 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 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.

1 はじめに

RNA 分子のフォールディングシミュレーションを二次構造レベルで高速かつ正確に行うことの重要性が近年高まっている．しかしながら，この問題には，構造空間の組合せ爆発のため効率良くシミュレーションすることは非常に難しい．本論文では，RNA 二次構造をグラフを用いて数え上げるという新しい発想を用いて，フォールディングシミュレーションを近似的かつ高速に行うアルゴリズムを提案する．

従来の研究としては，ヘリックスとよばれる連続した塩基対領域を単位として，モンテカルロ法により確率的に構造変化を行いながらシミュレーションする手法 ([4, 6]) や，1 つの塩基対を単位として構造変化を行うアプローチ ([1, 5]) がある．前者は，構造の空間をある程度削減することができるが，構造空間の組合せ爆発の問題を解決できるわけではないし，ヘリックスという大きな構造を 1 単位時間で追加・削除するため，精度に問題があるといわれている ([1])．後者は精確なシミュレーション結果を与えてくれるが，組合せ爆発の問題は前者より深刻である．

本研究では，構造空間として，シュードノットとマルチループを含まない構造を考える．つまり，線形な構造が並列に連接されることが許されるような二次構造のクラスである．このように線形構造しか含まない二次構造のクラスは，応用上限定的であると考えられるが，複雑な三次構造のフォールディング過程にはヘアピンの構造変化が含まれているケースも多く，物理化学的に詳細な研究が進められている重要な研究対象である．

2 RNA フォールディング

塩基対 (i,j) および (k,l) が $i<k<l<j$ を満たすとき $(i,j) \prec (k,l)$ と書く．二次構造 S に含まれる任意の異なる塩基対 bp_1 と bp_2 に対して $bp_1 \prec bp_2$ または

$bp_2 \prec bp_1$ が成り立つとき，S は線形な二次構造であるという．S に含まれる 3 つの塩基対 (i,j), (i_1,j_1) および (i_2,j_2) が $i < i_1 < j_1 < i_2 < j_2 < j$ を満たすとき，S はマルチループを含むという．S が $i < k < j < l$ を満たす塩基対 (i,j) および (k,l) をもつとき，S はシュードノットを含むという．

RNA 分子 X のシュードノットとマルチループを含まない二次構造の集合を $C(X) = \{S_1, ..., S_m\}$ とし，$C(X)$ のことを X の構造空間とよぶことにする．$C(X)$ の各構造は，**Add** と **Delete** という 2 つの操作によって別の構造に遷移する．**Add** は現在の構造に新たに 1 つの塩基対を追加する操作である．**Delete** は現在の構造から 1 つの塩基対を取り除く操作である．構造 S_i に **Add** または **Delete** のいずれか 1 つの操作を 1 回適用することにより到達できる構造の集合を $Nbr(S_i)$ で表すことにする．

ある時点において，X が二次構造 S_i をもつとき，S_i はその隣接構造 $S_j \in Nbr(S_i)$ のうちの 1 つに自由エネルギーに基づいて定められる確率分布に基づいて遷移する．以下にその仕組みを説明する．

塩基対を一切もたない二次構造をランダムチェインとよぶことにする．ランダムチェインから構造 S_i に X がフォールディングするときのギブス自由エネルギーの変化を G_i^0 で表す．本研究では，構造 S_i から S_j に変化する速度定数は以下の Metropolis rule で与えられるものとする:

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

時刻 t に X が二次構造 $S_i \in C(X)$ をとる確率を $P_i(t)$ で表す．すると RNA 分子のフォールディングの運動方程式は以下のマスター方程式で与えられる：

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

ここで，k_{cal} は実験結果との調整をするための補整パラメータである．本研究では，[5] と同様に $k_{cal} = 3.34 \times 10^6$ という値を用いる．

X の長さに関して $C(X)$ の要素数は組合せ爆発を起こす．従って，従来の研究のほとんどは構造空間 $C(X)$ のサイズを削減するアプローチをとっている．本研究では $C(X)$ をそのまま厳密に保持したままで，効率良くかつ近似的にマスター方程式に基づいてシミュレーションをするアルゴリズムを提案する．このアルゴリズムはシミュレーションの収束点が平衡状態に一致するという理論的な性質をもつ．

本論文の第二著者は，分子がさまざまな複合体を形成する化学反応系の平衡状態を求めるための一般的な枠組みを提案している ([2])．平衡状態は反応の収束点であるから，平衡状態計算と反応のシミュレーション問題は密接な関係があるはずである．本研究では，この発想に基づいて新しい効率の良いシミュレーションアルゴリズムを提案する．次節では，一分子反応系に限定した場合における [2] の平衡状態計算の枠組みを紹介する．

3 構造列挙による平衡状態計算

[2] の枠組みでは，以下のようにグラフのパス集合と構造空間 $C(X)$ を対応させることにより，平衡状態計算に必要な変数の個数を大幅に削減することに成功している．

分子 X の構造空間を $C(X)$ とし，$S \in C(X)$ のギブス自由エネルギーを $F(S)$ で表す．アサイクリックな有向グラフ $G = (V, Eg)$ を考える．頂点 $v \in V$ に対して，v_{in} (v_{out}) によって v に入る (v から出る) 辺の集合を表す．条件 $v_{in} = \emptyset$ と $v_{out} = \emptyset$ を満たす頂点 v をそれぞれ初期頂点および最終頂点とよぶ．初期頂点の集合を V_0，最終頂点の集合を V_f で表す．G における初期頂点から最終頂点にいたるすべてのパスの集合を $PT(G)$ で表す．

[2] の枠組みで最も本質的なのは，後述する条件を満たすような $PT(G)$ から $C(X)$ への 1 対 1 の関数 ψ を考えるところである．つまり，関数 ψ によって $C(X)$ に含まれる構造をすべて列挙する．このようなグラフと ψ の組合せが準備できると，反応系の平衡状態を計算する問題は，G の辺の個数 $|Eg|$ だけの変数をもつ目的関数を最小化する問題に帰着できることが示されている ([2])．ここで，構造の個数はパスの個数 $|PT(G)|$ 存在することに注目されたい．つまり非常に莫大な個数の構造体が生成されるような反応系の平衡状態が，わずかな個数 ($|Eg|$) しかない変数の最適化問題に帰着できるわけである．しかも，この目的関数は凸関数であり凸計画法が適用できる．

関数 ψ に対する条件は，以下に記述するように非常に単純なものである．つまり，辺集合 Eg に対する重み関数 $\epsilon : Eg \to \mathbf{R}$ が存在して以下の式がすべての $\gamma \in PT(G)$ に対して成り立つことが要求される：

$$F(\psi(\gamma)) = \sum_{e \in Eg \text{ s.t. } e \in \gamma} \epsilon(e).$$

これは，γ に現れる辺の重みの総和が γ に対応する構造 $\psi(\gamma)$ の自由エネルギーに等しいことを意味する．直観的には，各辺 e は局所的なある構造に対応しており，その局所的な構造の自由エネルギーが $\epsilon(e)$ で与えられる．

RNA 分子の平衡状態を計算する場合は，各辺は，ヘアピンループ，内部ループ，バルジループといった局所的なループ構造に対応する．次節に RNA 分子に適用した場合の例を与える．

4 RNA 分子の二次構造の列挙グラフ

線形構造を列挙するグラフを与える．シュードノットとマルチループを含まない構造クラスへの拡張は，紙面の都合上省略する．

RNA 分子 $X = x_1 \cdots x_n$ を考える．頂点集合 V は，X が形成し得るすべての塩基対と，初期頂点 s および最終頂点 f からなるものとする．塩基対 (i,j) と (k,l) の間に $(i,j) \prec (k,l)$ が成り立つとき頂点 (i,j) から頂点 (k,l) へ辺をひく．また，初期頂点からすべての塩基対へ辺を追加し，すべての塩基対から最終頂点へ辺を追加する．

RNA 配列 X = GGAAACUU に対するグラフを図 1 (c) に与える．図 1 (a) は X が形成し得るすべての塩基対を表す．パス $s \to (1,7) \to (2,6) \to f$ は，図 1 (b) の上の線形構造に対応し，パス $s \to (1,8) \to (3,7) \to f$ は図 1 (b) の下の線形構造に対応する．このように，このグラフ G を用いてすべての線形な二次構造を列挙することができる．このようなパスから線形構造への対応を表す写像が ψ である．

この例からもわかるように，塩基対 (i,j) から (k,l) への辺は，この 2 つの塩基対によって囲まれる局所ループ構造 (スタック塩基対，バルジループ，内部ループ) に対応する．また，初期頂点 s から塩基対 (i,j) への辺は，(i,j) の外側のフリーエンドループに対応する．最後に，塩基対 (i,j) から最終頂点 f への辺は，(i,j) によって閉じられるヘアピンループに対応する．よって，辺 e の重み $\epsilon(e)$ は対応する局所的なループ構造の自由エネルギーを与えればよい．前述の例では，各部分構造の自由エネルギーは図 1 (b) の実数値として与えられている．従って，辺 $s \to (1,7)$, $(1,7) \to (2,6)$, $(2,6) \to f$, $s \to (1,8)$, $(1,8) \to (3,7)$, $(3,7) \to f$ の重みはそれぞれ，+0.4, −2.1, +5.7, +0.5, +2.5, +6.2, で与えられる．

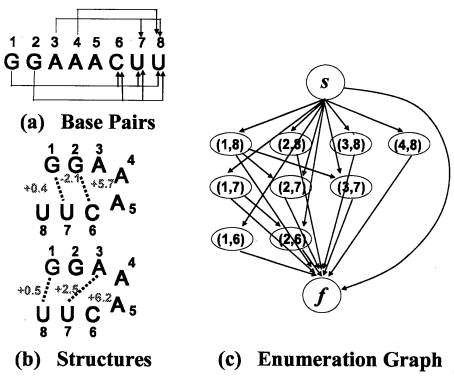

図 1: 列挙グラフの例

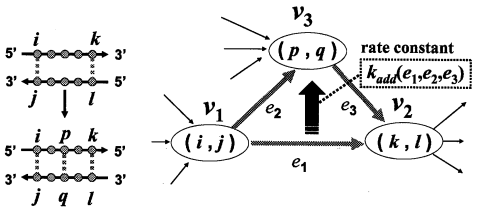

図 2: **Add** 操作の局所的解釈

5 提案するアルゴリズム

この節では，4 節で導入したグラフを用いて効率良くシミュレーションするアルゴリズムを提案する．このアルゴリズムの本質は，**Add** および **Delete** という操作を**グラフ上で局所的に解釈する**という発想にある．そこで，各辺 e に対して，e に対応する局所的な構造の存在確率を表す変数 w_e を導入する．

ここでは紙面の都合上，**Add** 操作を局所的に解釈する方法のみを示す (図 2)．塩基対 (i,j) と (k,l) の間に新しい塩基対 (p,q) を追加する **Add** 操作を考える．ここで，$i < p < k < l < q < j$ が成り立つものとする．この操作は，局所的に解釈すると，辺 $(i,j) \to (k,l)$ を含むパスが辺 $(i,j) \to (p,q)$ と 辺 $(p,q) \to (k,l)$ を含むパスに変化することに対応する．従って，**Add** 操作は 3 つの辺 $(i,j) \to (k,l)$, $(i,j) \to (p,q)$, および $(p,q) \to (k,l)$ の確率 $w_{e_1}, w_{e_2}, w_{e_3}$ を以下のように増減させることに対応する．

$$\Delta w_{e_1} = -k_{cal} \cdot k_{add}(e_1,e_2,e_3) \cdot w_{e_1} \Delta t, \quad (2)$$

$$\Delta w_{e_2} = k_{cal} \cdot k_{add}(e_1,e_2,e_3) \cdot w_{e_1} \Delta t, \quad (3)$$

$$\Delta w_{e_3} = k_{cal} \cdot k_{add}(e_1,e_2,e_3) \cdot w_{e_1} \Delta t, \quad (4)$$

ここで，$k_{add}(e_1,e_2,e_3)$ はこの局所的な構造変化に対応

する速度定数であり，以下のように定義される．

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

提案するアルゴリズムは，このように局所的に解釈された **Add** と **Delete** 操作を，図 2 のように局所的に配置されている全ての 3 角形の 3 つの辺に対して適用し，各辺 e の存在確率 w_e を更新していく．このとき，各構造 S の存在確率は辺の存在確率から以下のように得る．

$$[S] = \frac{\prod_{e\in Eg \text{ s.t. } e\in\psi^{-1}(S)} w_e}{\prod_{v\in V-V_0-V_f \text{ s.t. } v\in\psi^{-1}(S)} w_v}, \quad (5)$$

上記のような 3 角形は高々 X の長さに関して多項式個しか存在しないので，このアルゴリズムが時間ステップ Δt 分の更新を行うのに必要な計算時間は X の長さに関して多項式時間で抑えられる．また，各 w_e の変化は，[2] の枠組みで用いられている目的関数の降下方向であることが証明できる．この目的関数は凸関数であるから，このアルゴリズムは，その最適値を与える分布，つまり平衡状態に収束する．

構造クラスをシュードノットとマルチループを含まない構造へ拡張する方法は紙面の都合上省略する．

6 実験結果

提案手法の有効性を示すために，すべての構造を列挙してマスター方程式 (1) に基づいてシミュレーションを行った場合との比較を行った．表 1 に用いた RNA 配列を示し，表 2 にその実験結果を示す．実験はすべての配列に対して，$\Delta t = 1.0 \times 10^{-8}$ (sec) として，500,000 ステップのシミュレーションを行った．表 2 において，T_E は全列挙による方法にかかった計算時間であり，T_P は提案手法にかかった計算時間である．N_{str} は各配列の二次構造の個数であり，Ac は提案手法と全列挙手法の間の相対誤差を，収束までの時間と全構造にわたって平均して求めた値である．提案手法が圧倒的に高速であり良い近似を与えていることがわかる．また，すべてのシミュレーションにおいて，提案手法が平衡状態に収束していることを確認した．

7 おわりに

本稿では，RNA 二次構造のフォールディングシミュレーションに対する新しい効率の良いアルゴリズムを与え，その有効性を検証した．マルチループやシュードノットといった二次構造や，複数配列のインタラクションを取扱えるようにすることが今後の課題である．

No.	length	Sequence
1	20	CGUGCUGCUGACUACAGCGA
2	20	AGCUACGUAGCUAGCUAGCU
3	20	CUUAGACAUGUGCGCUAGCG
4	25	GUCAGCGAGCAUCGUAGCUGACUGA

表 1: RNA 配列

No.	N_{str}	T_E	T_P	Ac
1	2601	1m09s	32s	7.26×10^{-3}
2	2660	1m15s	28s	5.21×10^{-2}
3	2881	1m16s	38s	3.03×10^{-3}
4	36190	1h12m52s	3m25s	1.08×10^{-1}

表 2: 計算時間と精度

参考文献

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

[2] 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)

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

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

[6] 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.