RNA二次構造予測における塩基対数最大化アルゴリズムの高速化の検討

柴田, 圭
九州大学大学院システム情報科学府情報理学専攻

馬場, 謙介
九州大学大学院システム情報科学研究院情報理学部門

出版情報：情報処理学会研究報告. 2006-MPS-62(2006-BIO-7)（135), pp.107-110, 2006-12. 情報処理学会
バージョン：

RNA 二次構造予測における塩基対数最大化アルゴリズム高速化の検討

柴田圭 [1], 馬場謙介 [2]
[1] 九州大学大学院システム情報科学府情報理学専攻
[2] 九州大学大学院システム情報科学研究院情報理学部門
〒 819-0395 福岡市西区元岡 744
E-mail:{shibata,baba}@c.csce.kyushu-u.ac.jp

概要 RNA 配列の解析においては，文字列としての単純な並びよりも，塩基対の相互作用による二次構造が重要視されている．本稿では，RNA の二次構造予測の最も基本的なアルゴリズムとして，Nussinov らによる塩基対数最大化アルゴリズムに着目し，高速化の検討を行っている．このアルゴリズムで行われる計算のうち，単純な並列化が適用できない部分について，事前の計算結果から省略できる場合が判別できることを示している．そして，アルゴリズムの高速化へ向けての，この手法の具体的な適用について検討している．

A Note on Speedup of Maximal-pairing Algorithm for RNA Secondary Structure Prediction

Kei Shibata[1] and Kensuke Baba[2]
[1]Graduate School of Information Science and Electrical Engineering, Kyushu University
[2]Faculty of Information Science and Electrical Engineering, Kyushu University
Motooka 744, Nishi-ku, Fukuoka 819-0395, JAPAN
E-mail:{shibata,baba}@c.csce.kyushu-u.ac.jp

Abstract In analysis of the RNA array, the secondary structure based on the base pair is attached to importance from the simple structure as a string. In this paper, the maximal-pairing algorithm by Nussinov *et. al* and a speedup method for it are considered as a basis of secondary structure prediction for RNA. In this algorithm, there exist a calculation which a straightforward parallelism is not applicable. This paper shows the method to distinguish from a prior calculation result the case that the calculation can be omitted. Moreover, an application of our method to speedup the algorithm is discussed concretely.

1 はじめに

あらゆる分野へのコンピュータの普及を背景に，様々な事象を文字列の解析によるアプローチで理解しようとする研究が多く行われている．特に，分子生物学の分野における情報学的な解析の基礎となるのは，DNA 塩基配列やタンパク質アミノ酸配列などの文字列としての処理である．これらの配列では文字の並びが重要視されているのに対し，RNA 配列においては塩基対の相互作用による二次構造が注目されており，解析がより困難になっている．

我々の研究の目的は，既存のコンピュータでは現実的な時間で解くことができない問題を，新しい技術による高性能なコンピュータを用いて解くことである．そのために，ハードウェアの実現可能性を考慮した上で，それに適したアルゴリズムの変更を検討する．本稿では，対象とする問題を RNA の二次構造予測とし，そのための最も基本的なアルゴリズムとして，Nussinov ら [2] による塩基対数最大化アルゴリズムに着目する．ここでは，具体的な計算時間よりも，アルゴリズム中で行われる計算に対してどういった高速化が有効かを明らかにすることを目標とし，この結果を基に，より複雑な RNA 二次構造予測アルゴリズムの効率化を検討する予定である．

Nussinov アルゴリズムは，1 本の RNA 配列に対して，塩基対数が最大となる二次構造を求めるものである．このアルゴリズムは動的計画法に基づいており，行列の値を再帰的に求めるものである．本稿では，アルゴリズムが実行されるコンピュータとして，ワードサイズが大きなものを想定し，実行時間について効率のよいアルゴリズムへの拡張を検討する．具体的には，文字列の編集距離を計算するアルゴリズムについての高速化手法である Myers [1] によるビットパラレル手法の，Nussinov アルゴリズムへの適用を試みる．

2 RNA 配列における二次構造予測

2.1 RNA と二次構造

RNA 配列は，４種類の塩基アデニン，シトシン，グアニン，ウラシルを表す文字 A，C，G，U から構成される文字列である．いくつかの RNA は高次の構造を持たない単純な文字列として説明される一方，意味のある３次元構造を持つ RNA が数多く発見されている．４種類の塩基のうち，G と C や，A や U は水素結合の塩基対を作る．この塩基対構造は RNA の**二次構造**と呼ばれ，塩基対の安定を表すスコア等を導入することで生物学的にもっともらしい構造の予測が行われている．二次構造は典型的には図１に示すような平面図で表現される．

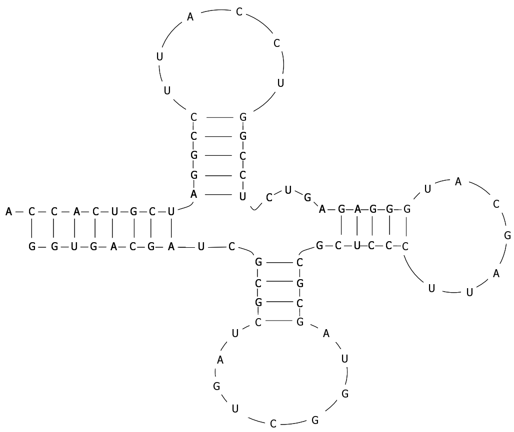

図 1: RNA の二次構造の１例

RNA 二次構造では，塩基対はほぼ常に入れ子になって現れる．入れ子でない塩基対があるとき，**シュードノット**と呼ぶ．シュードノットの場合の二次構造の例を図２に示す．

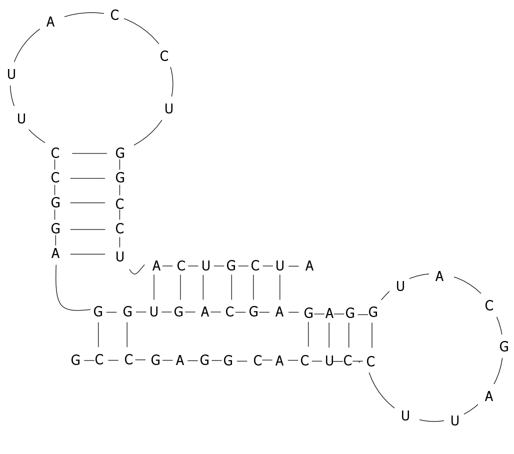

図 2: シュードノットの例

シュードノットの起こる回数は，入れ子になる二次構造の塩基対に比べ非常に少ない．そのため，RNA 相同性のデータベース検索を含む多くの目的の際には，アルゴリズムの効率性のためシュードノットの情報を犠牲にすることがある．

2.2 Nussinov アルゴリズム

Nussinov ら [2] は，シュードノットを考慮しない場合の RNA 配列の塩基対数を最大にする効率的なアルゴリズムを導入した．このアルゴリズムでは，小さな部分配列に対しての最大塩基対数を計算し，より大きな部分配列について計算していく．s を長さ n の入力文字列とし，$1 \leq i \leq n$ について s の i 番目の要素を s_i で，$1 \leq i \leq j \leq n$ について s の部分文字列 $s_i s_{i+1} \cdots s_j$ を $s_{i,j}$ で表す．このとき，$s_{i,j}$ の最適構造は，以下の場合を考えればよい．

- i 非ペア:$s_{i+1,j}$ の最適構造に，塩基対をなさない s_i を加える．
- j 非ペア:$s_{i,j-1}$ の最適構造に，塩基対をなさない s_j を加える．
- i,j ペア:$s_{i+1,j-1}$ の最適構造に，塩基対をなす s_i と s_j 加える．
- 二股枝分かれ:$s_{i,k}$ の最適構造と $s_{k+1,j}$ の最適構造を結合する．

上記４つの状態を図３に示す．

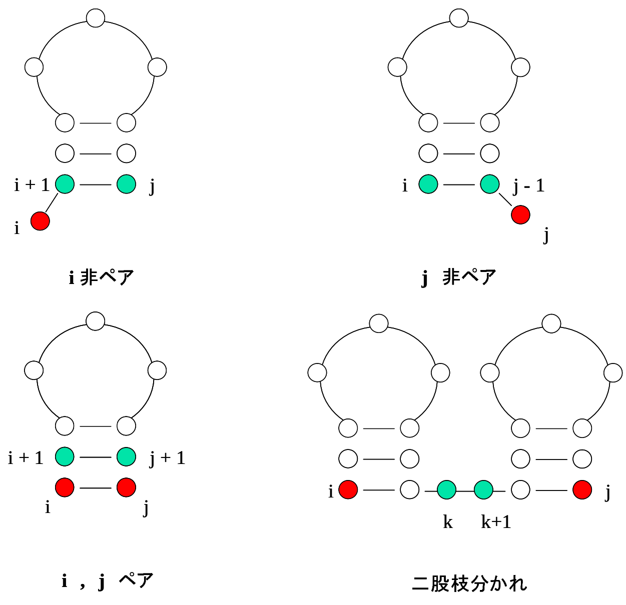

図 3: 最適構造を１段階小さい最適構造から得る方法

今，$\delta(i,j)$ を，s_i と s_j が塩基対をなすとき 1，そうでないとき 0 とすると，$s_{i,j}$ の最大塩基対数 $\gamma(i,j)$ は以下の式で求められる．

$$\gamma(i,j) = \max \begin{cases} \gamma(i+1,j) \\ \gamma(i,j-1) \\ \gamma(i+1,j-1)+\delta(i,j) \\ \max_{i<k<j}\{\gamma(i,k)+\gamma(k+1,j) \end{cases} \tag{1}$$

ただし，初期値として，$1 \leq i \leq n$ について $\gamma(i,i) = 0$，$2 \leq i \leq n$ について $\gamma(i,i-1) = 0$ である．

図 4 にその行列の例を示す．

	G	G	G	A	A	A	U	C	C
G	0	0	0	0	0	0	1	2	3
G	0	0	0	0	0	0	1	2	3
G		0	0	0	0	0	1	2	2
A			0	0	0	0	1	1	1
A				0	0	0	1	1	1
A					0	0	1	1	1
U						0	0	0	0
C							0	0	0
C								0	0

図 4: 行列の例

以上によって求められる $\gamma(1,n)$ が入力文字列の最大塩基対数となる．また，ここで得られる行列の値を $\gamma(1,n)$ からトレースバックすることで，塩基対数についてのある最適構造を見つけることができるが，本稿では $\gamma(1,n)$ の値を求めるまでの計算を考える．このアルゴリズムの領域複雑さは $O(n^2)$，時間複雑さは $O(n^3)$ である．

3 高速化の検討

3.1 ビットパラレル手法

Myers [1] は，動的計画法に基づく近似文字列照合アルゴリズムとして，コンピュータのワード長に応じた高速化が可能な並列化手法を提案した．ここで用いられているビットパラレルと呼ばれる手法は，文字列間の類似度を二値の論理式の演算によって表すことで，求める行列の成分のうちワード長個の値を並行して計算するものである．

このアルゴリズムの詳細な説明は省略するが，ここで用いられている高速化のアイデアのひとつは，求める行列の隣り合う成分の差が高々1であることを利用して，1 列分の差をビット列で表現し，次の列についての値をビット演算により一度に求める点である．この際，論理演算による表現を可能にするために，動的計画法の帰納的な計算の事前の結果と次の結果との関係の詳しい解析を行っている．

3.2 計算の省略

Nussinov アルゴリズムの高速化を行う際，式 (1) の第 3 項までの計算については，例えば，行列の斜めの線に沿った単純な並列化が可能である．我々のアプローチは，第 4 項の計算を省略できる場合をできる限り事前の計算から判別することである．この判別は，ビットパラレル手法のアイデアを適用することで，帰納的計算の事前の結果から行われる．

Nussinov アルゴリズムで求める行列について，縦と横に隣り合う成分の差を

$$\begin{aligned} \Delta v(i,j) &= \gamma(i,j) - \gamma(i+1,j) \\ \Delta h(i,j) &= \gamma(i,j) - \gamma(i,j-1) \end{aligned}$$

とすると，以下が成り立つ．

補題 1 $1 \leq i < j \leq n$ について，$\Delta v(i,j), \Delta h(i,j) \in \{0,1\}$ である．

証明 式 (1) より，$1 \leq i < j \leq n$ について，$\Delta v(i,j) \geq 0$ かつ $\Delta h(i,j) \geq 0$ は明らかである．以下，$\Delta v(i,j) \leq 1$ を示す．$\Delta h(i,j)$ については同様に証明することができる．

$\Delta v(p,q) \geq 2$ となる p と q が存在するとき，式 (1) より，

(a) $\gamma(p,q-1) - \gamma(p+1,q) \geq 2$,

(b) $\gamma(p+1,q-1) - \gamma(p+1,q) \geq 1$，または

(c) $\max_{p<k<q}\{\gamma(p,k)+\gamma(k+1,q)\} - \gamma(p+1,q) \geq 2$

が成り立つ．(a) のとき，$1 \leq i < j \leq n$ について $\Delta v(i,j) \geq 0$ であることから，$\gamma(p,q-1)-\gamma(p+1,q-1) \geq 2$ となるが，$\Delta v(p,q-1) \geq 2$ となり矛盾．(b) のとき，$\Delta h(p+1,q) \leq -1$ となり矛盾．(c) のとき，$\gamma(p,k)+\gamma(k+1,q)-\gamma(p+1,q) \geq 2$ となる k が存在する．式 (1) より，その k について $\gamma(p+1,q) \geq \gamma(p+1,k)+\gamma(k+1,q)$ が成り立つので，$\gamma(p,k)-\gamma(p+1,k) = \Delta v(p,k) \geq 2$ となり矛盾．□

これによって，以下の場合は式 (1) の第 4 項を計算しなくて良いことが分かる．

定理 1 $1 \leq i < j \leq n$ について，

(i) $\Delta v(i+1,j) = 0$ かつ $\Delta h(i,j-1) = 1$,

(ii) $\Delta v(i+1,j) = 1$ かつ $\Delta h(i,j-1) = 0$，または，

(iii) $\Delta v(i+1,j) = \Delta h(i,j-1) = 0$ かつ $\delta(i,j) = 1$

ならば，$\gamma(i,j) = \max\{\gamma(i+1,j), \gamma(i,j-1), \gamma(i+1,j-1) + \delta(i,j)\}$.

証明 定義より，$\Delta v(i,j-1) + \Delta h(i,j) = \Delta h(i+1,j) + \Delta v(i,j)$ である．

(i) のとき，補題 1 より，$\Delta v(i,j) = 1$ かつ $\Delta h(i,j) = 0$ である．よって，$\gamma(i,j) = \gamma(i,j-1) \geq \max\{\gamma(i+1,j), \gamma(i+1,j-1) + \delta(i,j), \max_{i<k<j}\{\gamma(i,k) + \gamma(k+1,j)\}\}$ である．

同様に，(ii) のとき，$\Delta v(i,j) = 0$ かつ $\Delta h(i,j) = 1$ であり，$\gamma(i,j) = \gamma(i+1,j) \geq \max\{\gamma(i,j-1), \gamma(i+1,j-1) + \delta(i,j), \max_{i<k<j}\{\gamma(i,k) + \gamma(k+1,j)\}\}$ である．

(iii) のとき，$\gamma(i+1,j-1) = \gamma(i+1,j) = \gamma(i,j-1)$ かつ $\gamma(i,j) \geq \gamma(i+1,j-1) + 1$ である．よって，補題 1 より，$\gamma(i,j) = \gamma(i+1,j-1) + 1 \geq \max\{\gamma(i+1,j), \gamma(i,j-1), \max_{i<k<j}\{\gamma(i,k) + \gamma(k+1,j)\}\}$ である．□

この証明により，ある行列成分を計算する時，上記の条件を満たしているならば，式 (1) による計算を行わなくとも，行列成分を求めることができる．よってこの方法を実装することで Nussinov らのアルゴリズムが高速化できるのではないかと考えている．

4 おわりに

RNA 二次構造予測の最も基本的なアルゴリズムとして，Nussinov らによる塩基対数最大化アルゴリズムの高速化について検討を行った．このアルゴリズムで行われる計算のうち，単純な並列化が困難であるものについて，事前の計算結果から省略できる場合を判別する手法を示した．また，この結果を利用したさらなる高速化の可能性の検討を行った．今後の課題として，提案する手法に基づくアルゴリズムを実装し，定量的な評価を行う予定である．

謝辞

本文をまとめるにあたり，共にご議論いただいた九州大学システム LSI 研究センターならびに安浦・村上・松永・井上研究室の皆様に感謝いたします．なお，本研究は，一部平成 18 年度科学研究費補助金若手研究 B（課題番号 17700020）による．

参考文献

[1] Gene Myers, "A Fast Bit-Vector Algorithm for Approximate String Matching Based on Dynamic Programming", Journal of the ACM, pp.395–415, 1999.

[2] Ruth Nussinov, George Pieczenik, Jerrold R. Griggs, and Daniel J. Kleitman, "Algorithms for Loop Matching", SIAP vol.35, issue 1, pp.68–82, 1978.