完備辞書による snoRNA 修飾領域の一次スクリーニング法

山森　一人[a)]・山本　尭之[b)]・相川　勝[c)]

A Screening Method to Extract snoRNA Modification Domain with Fully Indexable Dictionary

Kunihito YAMAMORI, Takayuki YAMAMOTO, Masaru AIKAWA

Abstract

Small nucleolar RNA (snoRNA) is one of the non-coding RNAs exisiting in nucleolus of cells. SnoRNA is known that they will participate with RNA modifications such as methylation. Since the functions of snoRNAs in detail have not been known still yet, a new method is necessary to make clear the functions of snoRNA effectively. In particular, some researchers want to know where snoRNAs modify target RNAs. In this paper, we propose a screening method that extracts modification domain candidates on the target RNA based on the snoRNA sequence. A part of sequence makes a base pair structure between a snoRNA and a target RNA. Therefore complement of this part on snoRNA is similar to modification domain of target RNA. Our method finds the modification domain candidates by using Fully Indexable Dictionary (FID) built by complements of snoRNA sequences and snoRNA names. We use Trie with Level-Order Unary Degree Sequence (LOUDS) as FID. Trie is applied to full-text search, retrieving dictionary, and so on. LOUDS is a kind of succinct data structure, and it achieves low memory requirement and fast operation. However we have to concern to mutations and gaps in genomic searches with FID. If we allow unlimited number of mutations and gaps, it give us complete search in the dictionary. Therefore we give a mutation and a gap penalty, and terminate search process if the sum of penalty is more than a limit. Simulation results show that sensitivity, specificity and accuracy become 0.9 when maximum penalty is 2.

Keywords: snoRNA, rRNA, trie, LOUDS, succinct data structure

1. はじめに

ゲノム科学の進歩やコンピュータの大幅な進歩により、2003 年にヒトゲノム配列の解読完了が宣言された。これ以降、ゲノム科学の研究は各種ゲノムの解析が主となり、医学やバイオテクノロジーの飛躍的な発展への貢献が期待されている。ゲノムは、タンパク質のアミノ酸配列をコードするコーディング領域と、それ以外のノンコーディング領域に大別することができる。ゲノム解析が始まった当初、ノンコーディング領域は大部分は意味を成さないジャンク領域であると考えられてきた。しかし、現在では遺伝子発現調節を行う機能の他に、生体にとって必要な機能性RNAと呼ばれる遺伝子情報が含まれていることが分かってきており、注目を集めている[1)]。

機能性RNA の１つにsnoRNA[2)] がある。snoRNA は一部のRNA を標的とした化学的修飾に関与していると言われているが、いまだその意義は不明である。現在、異種生物間で同様の働きを持つと同定されたsnoRNA は約200 種類である。[2)] snoRNA遺伝子データベース“snOPY”[3)] ではsnoRNA遺伝子情報やsnoRNAの異種生物間の相同性、標的RNAとの情報などが公開されている。しかし、その解析スピードは十分とは言えない。この理由として、専門家による解析が手作業のため、予想以上に時間がかかることが挙げられる。このことから、snoRNAの解析作業をより効率的に、かつ正確に行う手法が必要である。

本研究では、Google日本語入力などに使われる簡潔データ構造LOUDSと木構造Trieを用いて、snoRNAによる標的RNA修飾領域候補を抽出する一次スクリーニング法を提案する。提案手法では、塩基配列情報に基づき、snoRNA遺伝子の相補的な塩基配列の相補鎖を辞書として用い、標的リボソームRNA（rRNA）塩基配列中の修飾領域を高速に検出する。

2. snoRNA の構造

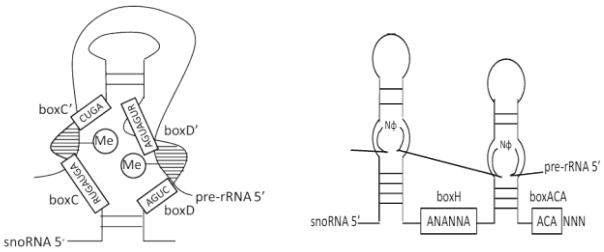
snoRNA は核小体内に存在する 60～250 塩基の比較的小さな RNA で、図 1(a) の boxC/D 型と図 1(b) の boxH/ACA 型の２つに大別され、それぞれ特徴的な２次構造を持つ。

boxC/D 型は概ね 60～100 塩基の長さであり、特徴的な配列 RUGAUGA（boxC）と CUGA（boxD）を持つ。また、ループ内にもう１セット同様な配列を持つ場合もある。

a) 宮崎大学工学教育研究部准教授

b) 宮崎大学工学部情報システム工学科

c) 宮崎大学工学部教育研究支援技術センター技術職員



(a). boxC/D 型. (b). boxH/ACA 型.

図 1. snoRNA の二次構造.

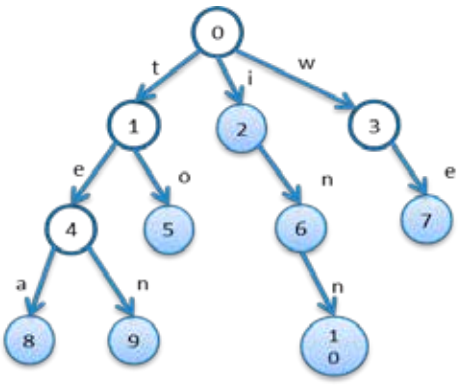


図 2. Trie の例.

boxD の上流には、標的となる RNA と相補的な塩基配列（以下、相補配列と呼ぶ）を 9〜20 塩基の長さで持つ。

boxH/ACA 型は 120〜250 塩基の長さで、中央の２ つのヘアピン構造の間には boxH（ANANNA）が、また 3’ 末端側の上流には boxACA（ACA）が存在する。boxH/ACA 型も boxC/D 型と同様に標的 RNA との相補配列を 6〜20 塩基の長さで持つ。

3. TrieとLOUDS

3.1. Trie

Trieは順序付き木構造の一種で、キー集合を扱う代表的なデータ構造であり、自然言語処理やフィルタリングに用いられる。図2 のように各枝に文字が付随し、根から葉へのパス上の枝に付随する文字をつないだ文字列がキーに対応する。色が付いているノードは根からそのノードまでの文字がキーに対応することを表し、その情報はTrieとは別のインデックスに格納されている。枝とノードを忠実に辿ることで検索を行い、検索に失敗した場合でも容易に訂正候補を示すことができる。

Trie の木構造をポインタベースで実装した場合、ノードへのアクセスは効率的に行えるが、作業領域量が大きくなる問題がある[4]。たとえば、Trie 中のノード数がn個の時、木のポインタ表現には$\log n$ビット必要なので全体でO(n log n)ビットが必要となり、非常に大きなキー集合を扱うことは難しい。そのため、性能を保ちつつ作業領域量を削減する簡潔データ構造LOUDS[5]を併用する。これにより、作業領域量をO(n) にまで削減することができる。

3.2. LOUDS

LOUDS（Level-Ordered Unary Degree Sequence:ラウズ）[5]とは、木構造を表現する簡潔データ構造（Succienct Data Structure:SDS）の１つであり、G. Jacobsonによって提唱された。SDSは、データを格納するのに情報理論的下界に近い領域量だけ使いつつ、かつ、索引を利用して高速な操作を実現したデータ構造である。ここでの情報理論的下界とは、位数Lの集合の中にある１つの要素を格納するときに必要なデータサイズ$\log L$ビットである。LOUDSはあくまで木の形のみを表現しており、Trieとして扱うためには帰途は別に文字の情報とキーの終端であるかの情報を保存する必要がある。図2のTrie木をLOUDS表現で表したものを配列B、登録したキーの終端の情報を格納したものを配列T、枝の文字を格納したものを配列Eとして表1に示す。

配列B、T、E は図2のTrie のノードを幅優先で探索した時の値が格納されている。配列B は始めにスーパールートと呼ぶ“10” を入れた後、注目したノードの子ノードの数だけ1を並べ、その後に0を置くという作業を繰り返してできた配列である。例えば表1の配列B で見ると、根であるノード０の持つ子ノード数は３であり、１ を３つ並べて0を置くので“1110”となる。このように「あるノードの子ノードの数だけ1を並べ、次に0を置く」という操作によって作られた配列をノードブロックと呼び、ノードブロック内の1と0の間には親子関係を持つ。

配列T は色付きのノードだった場合1を置き、違う場合は0を置くという作業を繰り返してできた配列である。色付きのノードは根からそのノードまでのパス上の枝に付随する文字をつないだ文字列がキーに対応していることを表す。

配列E は枝に付随する文字を格納している。配列B のノードブロックから、子ノードと親ノードは１対１の関係にあることがわかる。そのため、配列B 中の子ノードを表す1と配列E は1対１で対応している。

LOUDS を使うことによって、Trieの木としての役割を

表 1. Trie の LOUDS 表現の例.

i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
B	1	0	1	1	1	0	1	1	0	1	0	1	0	1	1	0	0	1	0	0	0	0	0
T	0	0	1	0	0	1	1	1	1	1	1												
E	t	i	w	e	o	n	e	a	n	n													

図 3. rank/select の基本操作.

配列B のように表すことができる。このように0と1によってコンパクトに木を表現したものを簡潔木と呼び、ノード数がn の時、高々$2n$+o(n) ビットで表現できる。[6)]

Trie の枝の長さの総長をl とすると、各枝に付随する文字をすべてつなげた配列は$E[0; l)$となる。Eに格納している文字の種類がσの時、配列Eは$l \log \sigma$ビットで表現できる。

LOUDSでの配列B上でのノードの移動、配列Bと配列E、Tの対応は節3.3で述べるrank/select操作によって求めることができる。これらを合わせると、TrieはLOUDSにより、$2l + l \log \sigma$+o($l \log \sigma$) ビットで表現できる。[6)]

3.3. rank/select操作

LOUDSを用いることによってTrieを簡潔な表現で表すことができるが、配列として表しただけでは本来の木と同様に扱うことができない。そのためrank/select操作が必要になる。rank/select操作はSDSで最も基本的、かつ重要な操作であり、これを利用することによって表1 をTrieとして扱えるようになる。rank/select操作の演算は以下のように定義する。

- $\text{rank}_b(B,i)$ ： $B[0,i)$中の$b \in \{0,1\}$の数を返す。
- $select_b(B,i)$ ： B中で先頭から見て$i+1$番目に出現した$b \in \{0,1\}$の位置を返す。

図3にrank/select操作の例を挙げる。$rank_1(B,6)$はB[0,6]中に1が4回出現しているので、$rank_1(B,6) = 4$となる。同様に、$select_0(B,3)$はBの先頭から3+1番目の0の出現位置は7であるので、$\text{select}_0(B,3) = 7$となる。このようにrank/select操作を使うことで添字bの出現数と先頭からの一を求めることができる。また、これらを組み合わせることでTrie木をLOUDS表現にした配列Bのノード間の移動や、枝を格納した配列E、キーの終端を表す情報を格納した配列Tの対応付けを行うことができる。

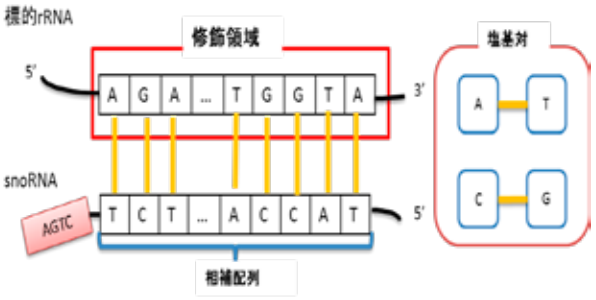


図 4. 塩基対の例.

4. 提案手法

提案手法では、boxC/D 型snoRNA上の相補配列を用いることで、標的rRNAの修飾領域候補を高速に抽出する。標的rRNAを修飾するsnoRNA の相補配列は、２つのbox間に塩基長9～20塩基の長さで存在し、図4のように修飾領域の配列と塩基対をなすことが知られている。塩基対はAとT、GとCの２つがあり、ある塩基配列中の塩基を、塩基対をなす塩基に置き換えた配列を相補鎖と呼ぶ。したがって、相補配列の相補鎖は標的rRNA の修飾領域の塩基配列に類似する。そこで、相補鎖とsnoRNA名を組として扱った完備辞書（LOUDSTrie）を作成し、標的rRNAの塩基配列を検索する。辞書と一致すれば当該rRNA 配列はsnoRNA に修飾されることを意味し、そのsnoRNA名が判明する。

4.1. 完備辞書の作成

標的rRNAの修飾領域を抽出する辞書はsnoRNA名と相補配列の相補鎖を格納する。相補鎖をキーとしてTrieを構築し、キーの終端を表すノードに対応して別にsnoRNA名を保存する。構築したTrieをLOUDSで表現することで、Trieとしての性能を保ったまま、作業領域量を削減する。

4.2. サンプルの切り出し

辞書を用いた標的rRNAの塩基配列に対する検索方法を説明する。修飾を受ける標的rRNAは種類によって約1,700から5,000塩基長を持つ塩基配列である。そのため、そのまま先ほどの辞書に当てはめて修飾領域を特定することはできず、先に検索しやすい形に切りそろえる必要がある。提案手法では標的rRNAの塩基配列を20塩基の長さで、先頭から１塩基ずつずらしながら塩基配列（Query）を切り出す。辞書を用いてQueryの検索を行い、辞書とマッチした結果を出力する。検索の様子を図5に示す。

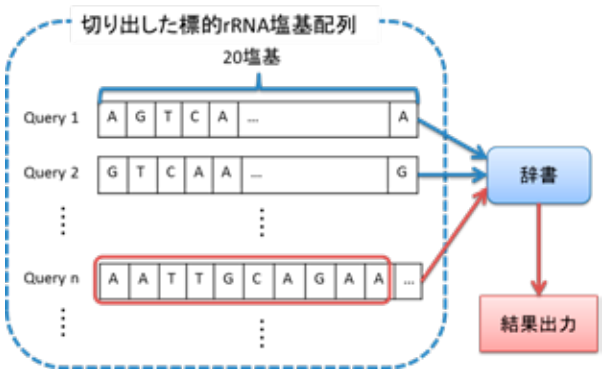
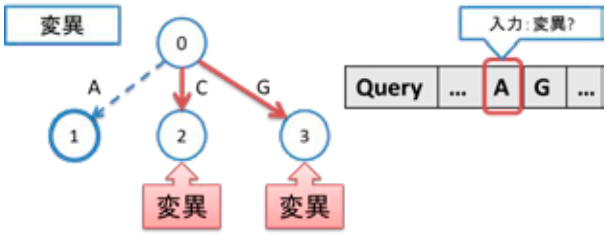
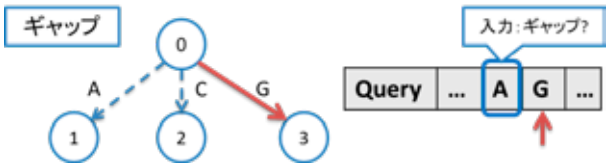


図 5. 辞書を用いた検索.



(a). 変異の対応.



(b). ギャップの対応.

図 6. 変異とギャップの対応.

4.3. 変異とギャップへの対応

ゲノム配列の検索においては以下の点に注意する必要がある。

- 変異：登録した相補鎖と比べ、一部の塩基が別の塩基と置き換わっている。
- ギャップ：登録した相補鎖と比べ、新たな塩基が挿入している。

変異とギャップの存在は、辞書を忠実に辿る検索において影響を与える。そのため、これらの可能性を考慮しつつ検索を行う必要がある。変異の可能性を考えて辞書を辿る例を図6(a)に示す。入力と枝が一致するノード1は変異なしと判断して検索を続ける。入力と枝が一致しないノード2、ノード3は入力を変異と判断して検索を続ける。

次にギャップの可能性を考えて辞書を辿る例を図6(b)に示す。入力の次の塩基を見て、その塩基と一致する枝がある場合、入力をギャップと判断する。

辞書を辿るたびに変異とギャップの可能性を許すことは網羅的な検索となり、計算量増加と検出精度の低下につながる。そのため、変異とギャップの出現は少ないほうが望ましい。

提案手法では、変異とギャップにそれぞれ1のペナルティを与えて累積し、ペナルティの上限値を定めることで塩基配列の相違を認めつつ検索範囲を狭める枝切りを行う。

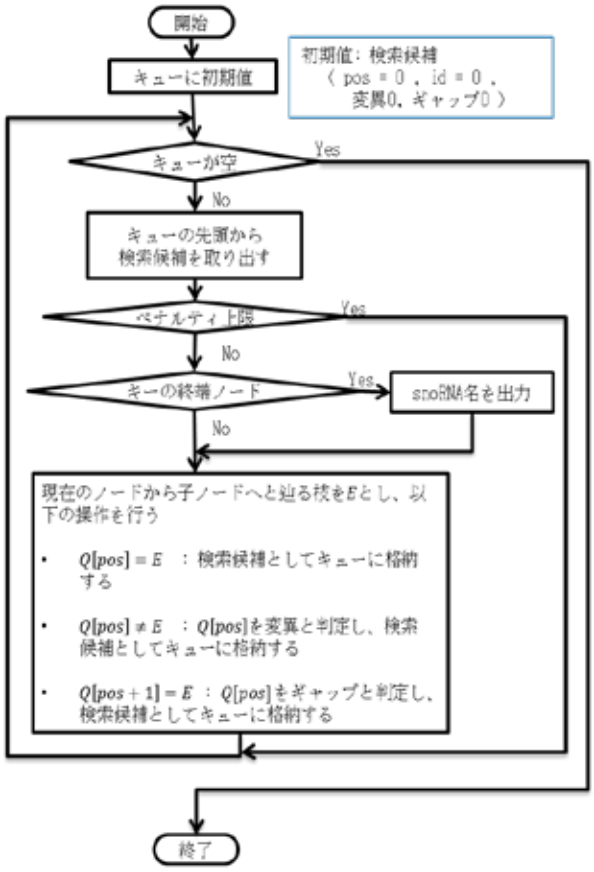


図 7. 検索アルゴリズム.

4.4. 検索過程

Queryを先頭から1 塩基ずつ取り出し、幅優先で辞書の検索を行う。この時、Query の注目塩基の位置を *pos*、注目塩基を*Q*[*pos*] とする。検索時には、変異とギャップを考慮するため分岐が生じる。分岐によりできた検索すべき候補は別に用意するキューに格納する。キューに格納する内容は、*pos*、ノード番号 (*id*) 、変異の出現回数、ギャップの出現回数を組としたもので、どの*Q*[*pos*] をどの*id* から検索再開するのかを求めることができる。この検索のアルゴリズムを図7 に示す。

5. 評価実験と考察

5.1. 実験データと実験条件

実験に使用した塩基配列データは、“snOPY”から取り出した。実験では５種の生物のsnoRNAの相補配列とsnoRNA 名を使用し、同生物の標的rRNA 塩基配列中の修飾領域を検出する。

評価実験に用いる正例データは、snoRNA による修飾が確認されている507個を使用する。負例データは、同塩基配列中の正例データ以外の配列とする。

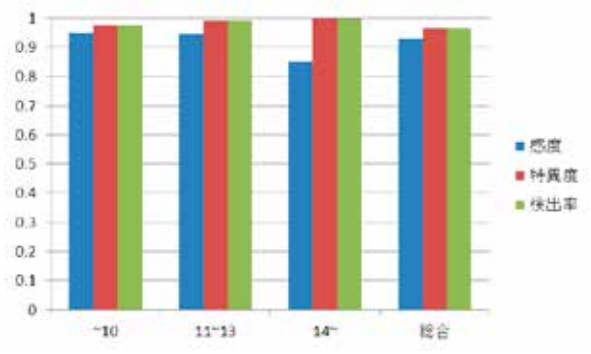


図 8. スクリーニング結果.

5.2. 評価方法

評価指標として、正例と負例をともに正しく判定できる確率を表す検出率、正例を正しく判定できる確率を表す感度、負例を正しく判定できる確率を表す特異度の３つを用いる。評価を行う際、修飾領域の塩基長を10塩基以下、11～13塩基、14塩基以上と、それらを合わせた総合で４つの各評価指標を求める。

5.3. 実験結果と考察

ペナルティの上限値を0～2と変化させて実験を行う。その中で感度、特異度、検出率ともに0.9以上となったペナルティ2 での実験結果を図8に示す。すべての修飾領域塩基長を合わせた総合の場合、図8 から感度0.927、特異度0.965、検出率0.964と高い精度を持つことがわかる。しかし、14 塩基を超える修飾領域を検出する場合は、感度0.849と、10塩基以下の感度0.950や11～13塩基の感度0.943と比べ、やや低くなっている。これは、14塩基以上の修飾領域検出に関してはペナルティ2では途中で検索が打ち切られているためと考えられる。ペナルティの許容量を引き上げることにより、14塩基以上の修飾領域の検出精度は向上すると予想されるが、それと同時に短い修飾領域の誤検出の増加が予想でき、感度が下がる可能性が高い。そのため、14塩基以上の修飾領域の検出に対応する別の手段が必要になると考えられる。

また、スクリーニングの処理時間は5種の生物の平均で約0.4秒となった。最も処理時間がかかった例は、Homo sapienceの辞書用データ136個を用いて約5,000塩基長の28SrRNAをスクリーニングした場合で約0.6秒以内であり、十分高速である。

6. おわりに

現在、snoRNAは約200種類もの分子種が同定されている。新たなsnoRNA の発見と機能の解析は、生命システムを理解するうえで今後ますます重要になると考えられる。しかし、snoRNAの機能解析はなかなか進まないのが現状である。その原因として、相同性の決定等の解析作業は専門家の手作業によるものが多く、時間がかかることが挙げられる。そのため、解析作業をより効率的に、かつ正確に行う手法が必要である。

本研究では、snoRNAが修飾する標的rRNA修飾領域候補を抽出する一次スクリーニング法を提案した。修飾領域の抽出は異種生物間の相同性を推測するにあたって必要な過程である。

実験の結果、提案手法における一次スクリーニングの精度は、許容ペナルティが2の時に感度0.927、特異度0.965、検出率0.964、平均処理時間は約0.4秒となることがわかった。これにより、本研究における一次スクリーニングは有用性を示すことができた。

今後の課題としては、検出精度の低かった14塩基以上の修飾領域に対しての精度の向上と、今回行わなかったboxH/ACA型snoRNAに対する実験が挙げられる。

参考文献

1) 塩見 春彦: noncoding RNAによる遺伝子発現制御機構と生命現象, 実験医学, Vol.28, pp.18-23, 2010.
2) 河合 剛太, 金井 昭夫: 機能性non-coding RNA, クバプロ, 2006.
3) 剣持 直哉: snoRNA Orthological Gene Database snOPY, http://snoopy.med.miyazaki-u.ac.jp/, 2008.
4) 岡野原 大輔: 最近の(trie)の話(xbwなど), http://research.preferred.jp/2011/05/trie_survey/, 2011.
5) G. Jacobson: Space-efficient static trees andgraphs, Proceedings of the 30th Annual Symposiumon Foundations of Computer Science, SFCS '89, pp. 549–554, 1989.
6) 岡野原 大輔: 高速文字列解析の世界, 岩波書店, 2012.