RNA Pseudoknot 구조의 시각화를 위한 새로운 표현 기법과 시각화 알고리즘

김우택[0], 이유진*, 한경숙*
인하대학교 자동화공학과[0], 인하대학교 컴퓨터공학부*
{g1991228[0], g2021343}@inhavision.inha.ac.kr, khan@inha.ac.kr

New Representation and Algorithm for Visualizing RNA Pseudoknot Structures

Wootaek Kim[0], Yujin Lee*, and Kyungsook Han*
Department of Automation Engineering[0], School of Computer Science and Engineering*, Inha University

요 약

RNA pseudoknot은 RNA 삼차 구조를 형성하는 중요한 구조요소일 뿐만 아니라, RNA 분자에서 중요한 역할을 한다. 지금까지 RNA pseudoknot 구조를 시각화하는 도구는 개발되어 있지 않기 때문에 대부분의 pseudoknot 구조의 시각화 작업은 수작업으로 이루어지고 있다. 본 논문은 RNA pseudoknot을 시각화를 위한 새로운 pseudoknot 표현 기법과 시각화 알고리즘에 대해서 소개한다. 새로운 표현기법은 모든 H-type pseudoknot을 uniform planar graph로 나타내고 RNA sequence의 진행방향을 따라가기가 쉽게 되어있다. 알고리즘을 이용하여 PseudoViewer라는 프로그램을 개발하였으며 PseudoViewer는 어떠한 시스템에서도 작동할 수 있는 Java로 구현되었다. 그 결과는 pseudoknot을 명확히 구분되고 보기 쉽도록 시각화됨을 보여준다.

1. 서 론

복잡한 분자 구조의 시각화는 구조를 쉽게 이해하게 해주며 생물학의 많은 부분에 도움을 줄 수 있게 한다. 본 논문은 RNA pseudoknot 구조를 사용자의 수작업을 거치지 않고 명확하게 그려낼 수 있는 새로운 알고리즘에 대해서 설명하고 있다.

Pseudoknot은 loop상의 염기들이 loop외부의 염기들과 상보적으로 쌍을 이룰 때 형성되는 RNA의 3차 구조요소이다. Pseudoknot은 모든 종류의 viral RNA 분자에서 구조적 특징을 일으킬 뿐만 아니라 RNA의 몇몇 중요한 기능을 야기한다. 예로 coding 영역에서 형성되는 pseudoknot 구조는 frameshift나 read-through를 야기하며, noncoding 영역에서 형성되는 pseudoknot 구조는 5′ noncoding 영역에서의 internal ribosomal entry site (IRES)에 참여하거나 3′ noncoding 영역에서 *translational enhancer*를 형성함으로써 translation이 시작되게 하기도 한다 [1]. 비록 pseudoknot은 위상학적으로 14종류로 분류되지만 [2], 대부분의 pseudoknot은 H-type pseudoknot에 해당된다. H-type의 H는 hairpin loop의 염기들이 hairpin loop 외부의 염기들과 결합한다고 하여 hairpin의 H를 의미한다.

현재 RNA pseudoknot 구조를 그려내는 방법은 알려져 있지 않다. 몇몇 컴퓨터 프로그램은 RNA 구조를 시각화하도록 개발되어졌지만 ([3,4,5,6]), 그것들 모두가 RNA의 이차 구조만을 시각화하도록 개발되었다. 따라서 RNA pseudoknot을 표현하기 위해서는 전적으로 수작업에 달려있다. graph 이론의 관점에서 RNA 2차 구조는 tree인 반면 RNA pseudoknot 구조의 형태는 graph (nonplanar graph가 될 수도 있음)이다. 그렇기 때문에 RNA pseudoknot 구조의 시각화는 계산적으로 이차 구조의 시각화보다 훨씬 어렵다. RNA 이차 구조 시각화의 어려운 문제중 하나가 구조요소의 겹침 현상인데, 이 문제는 시각화된 결과물의 가독성을 떨어뜨린다. 대부분의 RNA 이차 구조 시각화 프로그램에서는 프로그램의 반복 수행이나 사용자의 수작업을 통한 구조요소의 겹침 현상 제거 과정으로 인해 계산량이 늘어난다.

본 연구진은 H-type pseudoknot의 시각화를 위해 새로운 표현방법을 제시하는 한편, H-type pseudoknot을 겹침 현상없이 시각화할 수 있는 알고리즘을 개발하였으며 이 알고리즘을 바탕으로 PseudoViewer라는 프로그램을 개발하였다. 새로운 표현 방법은 모든 H-type pseudoknot을 uniform planar graph로 나타내고 RNA sequence의 진행 방향을 따라가기가 좀 더 쉽게 되어 있다. 개발된 알고리즘은 RNA pseudoknot 구조를 시각화할 수 있는 최초의 알고리즘이다. 알고리즘을 설계에는 다음 두 가지의 기준이 적용된다: (1) 구조요소의 겹침 현상을 최소화하고 시각화된 결과물의 가독성을 높인다. (2) pseudoknot 뿐만 아니라 RNA 전체 구조를 빠르고 명확하게 인식되도록 한다. 알고리즘과 프로그램 구현결과에 대해서는 본 논문의 뒷부분에 설명되어진다.

2. H-type Pseudoknot의 새로운 표현 기법

그림 1은 H-type의 전형적인 표현을 나타낸 것이다 [7]. 모든 H-type이 edge crossing이 나타나는데, edge crossing은 그림의 가독성을 떨어뜨리고 5′ 끝에서 3′ 끝으로의 RNA sequence의 진행방향을 따라가기가 어려워진다. 그러나 두 stem을 동축에 놓기 위해서는 edge

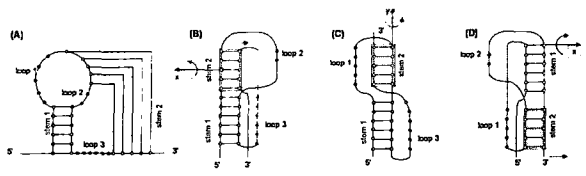

그림 1. H-type pseudoknot의 도식도 (a)일반적인 형태 (b)Loop 1이 없는 경우 (c) Loop 2가 없는 경우 (d) Loop 3이 없는 경우

crossing을 어쩔 수없이 그려야 한다. 두 stem을 동축에 놓는 것은 두 stem이 하나의 stem으로 보이게 하는 생물학적 의미를 가진다. 그러나 RNA 이차 구조를 포함한 pseudoknot의 그림은 기하적인 구조 (geometric structure)보다는 위상학적인 구조 (topological structure)로 표현된다. 즉 이러한 형태의 그림은 생물학적 의미의 희생을 감수하더라도 염기들 사이의 연결 관계를 명확히 보여 주는 것에 주안점을 두어야 한다.

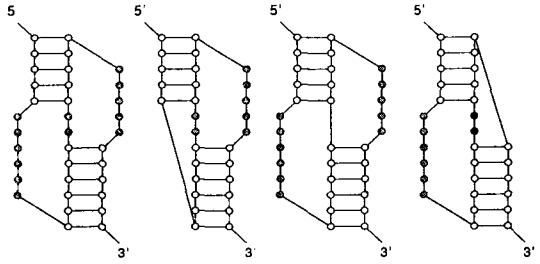

그림 2. H-type pseudoknot의 새로운 표현기법 (A) 일반적인 형태 (B) Loop 1이 없는 경우 (C) Loop 2가 없는 경우 (D) Loop 3이 없는 경우

본 논문은 모든 H-type pseudoknot을 일정하고 edge crossing 없이 표현하는 새로운 방법을 제시하고 있다. 그림 2는 그림 1과 동일한 pseudoknot을 새로운 방법으로 표현한 그림이다. 그림에서 edge crossing은 없으며 각 형태에 상관없이 두개의 내부 cycle을 포함한 동일한 형태를 나타내고 있다. 더욱이 5' 끝에서 3' 끝으로의 RNA sequence의 진행방향을 따라가기가 쉬워졌다. 새로운 표현 기법으로 그려진 그림에서는 두 stem이 동축에 놓여있지는 않지만 서로 평행하며 서로에게 인접해있다.

3. 시각화의 기본 전제

RNA 분자는 ribonucleotide가 연쇄적으로 연결된 사슬이다. 각 ribonucleotide는 다음 네 종류의 염기 adenine (A), cytosine (C), guanine (G), uracil (U) 중 하나를 포함한다. 한 가닥으로 구성된 RNA는 adenine과 uracil사이에, guanine과 cytosine사이에서 수소 결합을 함으로써 안정된 구조를 취하려 한다. A-U 염기쌍과 G-C 염기쌍을 canonical pair라고 하며 이보다 강력한 결합은 아니지만 G-U 염기쌍도 자주 발견되며 wobble pair라 부른다.

RNA 이차 구조에서는 다음의 두 가지 구조요소가 존재한다.

- stem(helix): 두 가닥으로 이루어진 부분으로서 염기쌍이 연속적으로 이루어진 영역
- regular loop: 한 가닥으로 이루어졌으며 hairpin loop과 internal loop, bulge loop, multiple loop, dangling end의 종류로 나뉜다.

개발된 알고리즘은 pseudoknot 구조를 시각화하기 때문에 다음 두 가지가 추가로 전제되어진다.

- pseudoknot: regular loop의 염기들이 loop 외부의 염기들과 결합함으로써 이루어진 구조 요소
- pseudoknot loop: pseudoknot뿐만 아니라 한 가닥으로 이루어진 부분들을 포함하고 있는 loop

4. Pseudoknot의 시각화 알고리즘

4.1 Pseudoknot

PseudoViewer는 pseudoknot을 표현하는데 널리 사용되는 pairing format의 ASCII 파일을 입력으로 받는다 [8]. pairing format은 pseudoknot 뿐만 아니라 이차 구조도 표현한다. 다음의 예는 G_5UGCAU_{10}과 $A_{17}UGCAU_{22}$가 쌍을 이룸으로써 pseudoknot이 형성되고 있다. 소괄호와 대괄호가 pseudoknot의 stem에 해당하며 ':'이 loop에 해당한다. 이차 구조에서는 소괄호와 ':'만이 사용되지만, pseudoknot의 경우 stem이 서로 중첩되기 때문에 대괄

```
# 1 2 3 4 5 6 7 8 9 | 1 2 3 4 5 6 7 8 9 | 1 2
$ C G U G G U G C A U A C G A U A A U G C A U
% ( ( ( : [ [ [ [ [ ) ) ) : : : ] ] ] ] ] ]
```

호가 추가되어 그 형태를 정의한다.

위와 같은 입력 데이터가 주어지면 PseudoViewer는 pseudoknot을 형성하는 stem을 분류하고 pseudoknot의 크기를 계산한다. 그 크기는 pseudoknot의 bounding box의 대각선 길이이며 그림 2에서 5' 끝의 염기와 3' 끝의 염기간의 길이와도 같다. pseudoknot의 크기는 pseudoknot loop의 크기를 결정하는 중요한 요소이다.

4.2 Regular Loop과 Pseudoknot Loop

RNA 이차 구조는 loop과 stem으로 형성되어 있다. 여기에 삼차 구조인 pseudoknot을 고려한다면 pseudoknot과 이차 구조 요소들과의 관계를 구분 지어야 한다. 개발된 알고리즘에서는 이러한 관계를 pseudoknot loop이라 명명된 새로운 loop을 적용하고 있다. pseudoknot loop이라는 것은 pseudoknot 자체를 하나의 이차 구조 요소로 구분지음으로써 pseudoknot이 포함된 single stranded 부분으로 정의한다. 그 형태는 원형이며 loop의 크기는 포함되어 있는 각각의 pseudoknot들의 크기와 염기들의 수, 연결된 stem의 수에 의해 크기가 결정된다. 그림 3에서 가운데 가장 큰 원 형태의 loop이 pseudoknot loop이다. 그리고 기존 이차 구조에서 loop은 본 논문에서 pseudoknot loop과 구분하기 위하여 regular loop이라 칭하고 있다.

5. 구현 결과

PseudoViewer는 Java로 구현되었으며 Java를 지원하는 모든 플랫폼에서 실행 가능하다. 그림 3은 ORSV (odontoglossum ringspot virus) RNA의 구조를 시각화한 결과이다. 염기들은 10개 단위로 번호를 매겼으며

pseudoknot의 경우에는 시작염기만 번호와 함께 녹색의 바탕색으로 나타내진다. pseudoknot은 전체가 노란색 바탕색으로 표현되어 다른 구조요소들과 쉽게 구분되도록 하였다. stem을 이루는 염기쌍의 관계는 canonical pair (A-U, G-C)인 경우는 filled circle, wobble pair (G-U)인 경우는 opened circle로 표현하였다. 구조요소들의 겹침 현상은 없으며, 한 개 이상의 pseudoknot으로 이루어진 pseudoknot loop은 regular loop과 같은 원형 형태를 취하고 있다. pseudoknot을 수작업으로 시각화한 것과 달리 PseudoViewer에 의해 시각화된 pseudoknot은 그 형태가 명확히 구분되어지고 육안으로도 보기 쉽도록 되어 있다.

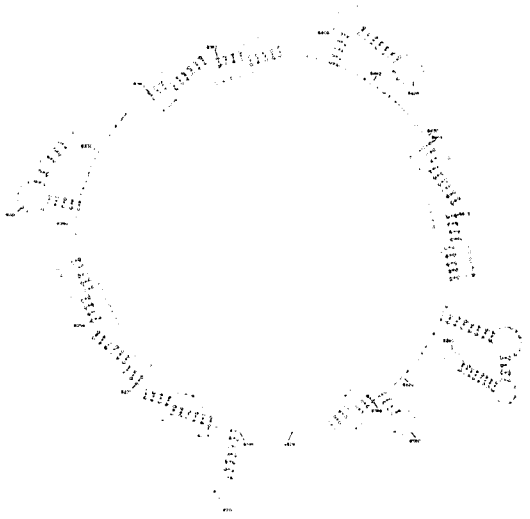

그림 3. 8개의 pseudoknot을 포함하는 ORSV RNA의 구조

6. 결 론

RNA pseudoknot의 시각화는 pseudoknot 내에 형성되는 내부 cycle뿐만 아니라 pseudoknot과 다른 구조 요소들로 형성되는 외부 cycle로 이루어진 graph (nonplanar일 수 있음)이다. 따라서 RNA pseudoknot을 명확하게 표현한다는 것은 RNA 이차 구조보다 그 계산이 더욱 어려워진다. 본 논문에서는 RNA pseudoknot을 이차원으로 시각화하는 알고리즘과 새로운 표현 방법을 제시하고 있으며 이런 알고리즘을 바탕으로 개발된 PseudoViewer라는 프로그램을 소개하고 있다. 새로운 표현기법은 H-type pseudoknot을 염기간의 연결 관계를 명확하고 일관되게 planar graph로 시각화한다.

RNA pseudoknot로 인하여 형성되는 내부 cycle과 외부 cycle을 tree의 node안에 숨김으로써 graph보다는 tree로서 전체 구조를 표현하였다. top level의 RNA 구조는 tree로 표현되고 tree의 loop은 그 depth 값에 따라 차례대로 그려진다. 구현 결과는 PseudoViewer가 H-type pseudoknot을 명확히 구분되고 보기 쉽도록 시각화됨을 보여준다. PseudoViewer의 알고리즘은 세계 최초로 개발된, pseudoknot을 포함하는 RNA 구조의 시각화하는 알고리즘이다.

PseudoViewer는 아직 완벽하게 완성되어진 것은 아니며, 여러 기능을 수행할 수 있도록 개선 작업에 있다. 첫 번째, PseudoViewer는 현재 H-type pseudoknot만을 시각화할 수 있으며 다른 형태의 pseudoknot을 시각화할 수 있는 알고리즘을 개발 중에 있다. 다른 형태의 pseudoknot에 대해 본 연구진은 시각화에 적합하도록 크게 다섯 종류로 분류를 하고 있다. 분류된 다섯 종류의 pseudoknot은 그림 2의 형태에서 조금씩 변형된 형태를 가지고 있으며 현재까지 알려진 pseudoknot중에서 극히 일부분을 제외하고는 대다수의 pseudoknot이 다섯 종류에 포함된다. 두 번째, Java를 지원하는 웹 브라우저를 사용하여 언제든지 실행할 수 있도록 웹 기반의 프로그램으로 가능토록 할 계획이다. 이것은 PseudoViewer의 개발 언어를 Java로 선택한 이유이기도 하다. 마지막으로 현재는 이차원으로 표현되는 것을 삼차원으로 표현할 수 있도록 개발하는 것이며, 이는 RNA pseudoknot 시각화 작업의 궁극적인 목표이다.

후기

본 연구는 한국과학재단의 지역대학우수과학자 지원연구 (과제번호 2001-1-30300-018-2)의 지원에 의하여 수행되었음.

6. 참고 문헌

1. Deiman, B.A.L.M., Pleij, C.W.A.: Pseudoknots: A Vital Feature in Viral RNA. Seminars in Virology 8. 166-175. 1997
2. Pleij, C.W.A.: Pseudoknots: a new motif in the RNA game. Trends in Biochemical Sciences 15. 143-147. 1990
3. Han, K., Kim, D., Kim, H-J.: A vector-based method for drawing RNA secondary structure. Bioinformatics 15. 286-197. 1999
4. Chetouani, F., Monestié, P., Thébault, P.,gaspin, C., Michot, B.: ESSA: an integrated and interactive computer tool for analyzing RNA secondary structure. Nucleic Acids Research 25. 3514-3522. 1997
5. Shapiro, B.A., Maizel, J., Lipkin, L.E., Currey, K., Whitney, C.: Generating non-overapping displays of nucleic acid secondary structure. Nucleic Acids Research 12. 75-88. 1984
6. Winnepenninckx, B., Van de Peer, Y., Backeljau, T., De Wachter, R.: CARD: A Drawing Tool for RNA Secondary Structure Models. BioTechniques 16. 1060-1063. 1995
7. Hilbers, C.W., Michiels, P.J.A., Heus, H.A.: New Developments in Struture Determination of Pseudoknots. Biopolymers 48. 137-153. 1998
8. van Batenburg, F.H.D., Gultyyaev, A.P., Pleij, C.W.A., Ng, J., Olihoek, J.: PseudoBase: a database with RNA pseudoknots. Nucleic Acids Res. 28. 201-204. 2000