# 예측된 RNA 이차 서열 구조의 그룹핑 및 시각화

손현일[O] 정유진
한국외국어대학교
{hison[O], chungyj}@hufs.ac.kr

# Visualization and Grouping of RNA Secondary Structures

Hyunil Son[O] Yoojin Chung
Dept. of Computer Engineering, Hankuk University of Foreign Studies

**요 약**

RNA 분자의 이차구조는 그 분자의 생물학적 기능을 이해하는데 필요한 요소이다. 본 논문은 RNA 이차 구조의 loop과 helix를 그룹화 하여 하나의 객체 단위로 설계한 뒤 필요한 loop과 helix들의 연결구조를 부분 시각화 하는 알고리즘을 소개한다. 기존의 존재하는 시각화 알고리즘들은 정보손실에 해당하는 겹침에 의한 피해를 줄이기 위해 다양한 변형 방법을 이용하였으나 이번 알고리즘은 사용자에 의해 필요로 하는 정보만을 선택해 볼 수 있기에 정보 손실의 문제를 피할 수 있다.

## 1. 서론

RNA란 핵산 가운데 당 성분이 D-리보오스인 것을 말하며 리보핵산(ribonucleaicacid)의 약칭이다. 이는 DNA의 유전정보를 체내에서 단백질로 발현시키는 역할을 한다. DNA의 유전정보는 핵 내에서 mRNA에 전사되고, 세포질의 리보솜에서 tRNA의 작용을 통해 단백질로 발현된다. 즉 RNA는 DNA의 정보가 단백질로 발현되기까지의 중간역할을 담당하며 생물학적 기능을 이해하는데 중요한 요소이다.

RNA는 염기서열로 연결되어있는 사슬구조로 되어있는데, 이 염기서열은 네 개의 염기( (A) adenine, (U) uracil, (G) guanine, (C) cytosine )들을 포함하고 있다. 이들은 수소결합을 함으로써 이차구조로의 모양을 갖추게되고, 이차구조가 삼차원 형태를 갖는 것을 삼차구조라 한다. 이차구조로 형성되기위해 A와 U, G와 C 가 수소결합을 이루고 간혹 G와 U의 결합도 볼 수 있다.

제시하는 논문은 이렇게 서로 결합을 이루고있는 부분과 결합을 이루지않는 부분들을 그룹화하여 각각 번호를 정해놓고 이를 서로간의 연결정보를 가지고 있는 연결리스트로 만들어놓은 후 사용자가 원하는 부분만을 연결정보를 통하여 시각화하게 된다.

기존의 시각화 프로그램들이 가지고 있는 문제점인 겹침(정보손실)문제를 해결하기위해 이미 그려놓은 구조가 겹침이 발생하면 재구성하여 다시 그리는 과정을 통해 문제해결을 하거나[1][2], 이미 그려진 구조를 사용자의 손에 의한 작업을 통해 문제를 해결하거나[3][4], 그리는 과정에서 이러한 수작업이나 재구성을 하지않으면서 최소한의 왜곡만을 허용하는 알고리즘들[5][6]이 이미 연구되어졌다. 본 논문에서는 전체 구조의 시각화에 나타나는 이러한 문제점에서 벗어나 사용자가 필요로 하는 부분만을 그 부분들의 연결정보에 의해 부분 시각화하여 겹침에 대한 문제를 피할 수 있고, 보다 간결한 검색과 구조확인을 할 수 있음을 보여준다.

## 2. 용어 정의

입력받게되는 input서열은 [그림1]과같이 염기서열과 짝 정보를 가지고 있는 기호로 되어있다.

아래 [그림 1]에서 염기서열 아래에 있는 ' ( ' 과 ' ) ' 그리고 ' - ' 기호는 서열에 대하여 서로 쌍을 이루는지 이루지 않는지에 대한 RNA 이차구조의 예측된 정보이다.

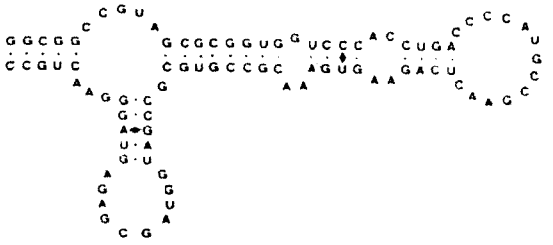
' ( ' 이렇게 열린기호에는 ' ) ' 이렇게 닫힌기호가 쌍을 이루게된다. 이렇게 쌍을 이루는 것을 helix 또는 stem이라 부른다. 본 논문에서는 helix 하나의 용어로 통일하겠다.

' - ' 기호는 쌍을 이루지 못한 염기서열을 말하고 이를 loop이라 부른다.

```
•  $   1 ggcggccguagcgcgguggucccac
•  %   1 (((((-----(((((((-((((--

•  $  26 cugaccccaugccgaacucagaagu
•  %  26 ((((-------------))))--))

•  $  51 gaaacgccgugcgccgaugguagcg
•  %  51 ))--)))))))-(((((-------

•  $  76 agaguagggaacugcc
•  %  76 ---)))))---)))))
```

[그림 1] 입력서열

helix는 두개의 가닥으로 이루어져 서로 쌍을 이루는 직선모양이고, loop은 서로 쌍을 이루지못한채 원모양을 형성하고 있다. [그림 2]



[그림 2] helix 와 loop

### 3. 알고리즘

loop과 helix들을 각각의 단위객체로 그룹화 시켜야한다. 그러기위해선 처음 입력받게 되는 서열에 대해 적절한 전처리과정이 필요하게 된다. 이렇게 전처리과정을 모두 마친 새로운 서열은 순서대로 입력받아 loop과 helix들을 구분지어 각각의 그룹으로 묶는작업이 필요로 하고, 이렇게 만들어진 그룹들은 하나의 연결리스트에 저장하게 된다.

#### 3.1 전처리

예측되어진 염기서열이 입력될 때 서열의 양끝이 Helix로 닫혀있어야 모든 helix를 그룹핑 할 수 있기 때문에 입력되어지는 서열 첫 부분 혹은 서열 끝부분이 닫혀있지 않다면 앞쪽과 뒤쪽을 ' ( '과 ' ) '으로 닫아주어야 한다.

bulge loop은 원의 모양을 하고 있어야 할 loop이 한쪽만 볼록 튀어나와 반원의 모습을 하고 있는 것이다. 이와 같이 ((((((----)))--))) 이렇게 된 bulge loop은 (((-(((----)))--))) 이렇게 양쪽으로 볼록한 원모양의 loop으로 그룹핑을 위해서 변형시켜야 한다.

연이어 나오는 helix에 대해서는 그룹을 나누기 위해 )))((( 모양은 )))-((( 이렇게 helix를 나누어 변형시켜야 한다.

#### 3.2 그룹화

그룹화는 다음과 같은 과정으로 이루어진다.

1) 전처리과정을 통해 수정되어진 서열들을 지역그룹(local group)으로 그룹핑한다.

ex. (((---))) -> (((, ---, )))

2) 지역그룹들을 하나의 완전한 helix와 loop을 이루기 위해 재구성한다.

ex. helix1: ((())), loop1 : ---

```
struct localGroupItem
{
        // 시작 index값
        int startNum;
        // 같은 형태의 연속된 길이
        int itemSize;
        // helix를 스택에 넣을때 loop스택의 현재 크기
        int loopStackSize;
};
```

[그림 3] item의 구조체

그룹화를 하기위해서 2가지의 스택을 사용한다. 하나는 helix의 지역그룹들을위한 스택이고, 다른 하나는 loop의 지역그룹들을위한 스택이다. [그림 3]은 그룹화를 하기위해 스택에 저장할 item의 구조체 모양이다.

**(1) loop에 대한 구조체 및 그룹화**

loop에 대한 구조체가 가지고 있어야 할 정보는 자신의 시작과 끝, 이루고 있는 염기번호, helix와 연결되어진 염기 등이다. 전처리 과정을 통해 다시 들어오는 RNA 서열을 하나씩 입력받으면서 연결 리스트에 저장할 때 하나의 단위그룹이 필요하게 된다. 이와 같은 하나의 loop그룹마다의 번호를 매기도록 구현 하였다.

이 과정에서의 문제점은 하나의 loop이기는 하나 서열순서상 하나의 loop에 하나이상의 helix가 존재하는 loop의 경우(이를 multiple loop 이라한다)는 서열이 차례대로 입력되어지게 되니 먼저 나온 완성되지 않은 loop의 지역그룹이 나온 후 인접한 helix를 만나고 그 뒤로 나머지 loop의 지역그룹이 나왔을 때야 비로소 하나의 완전한 loop이 생성된다는 것이다.

이와 같이 완성되지 않은 지역그룹들을 하나의 완전한 loop그룹을 이루기 위해 재구성해야 한다.

**(2) helix에 대한 구조체 및 그룹화**

helix에 대한 구조체는 그 자신의 한쪽쌍인 열린 helix 지역그룹이 나온 후 그에 대한 쌍인 닫힌 helix 지역그룹이 나와야 하나의 helix 그룹으로 묶을 수 있게 된다. 이 그룹이 가지고있는 정보는 자신과 연결되어있는 loop그룹의 연결정보, 자신들의 짝정보, 자신의 시작과 끝이다. 이러한 하나의 helix 그룹에 대해 번호를 매기게 설계,구현 하였다.

이렇게 구현된 helix와 loop그룹들을 순서에 맞게 하나의 연결리스트에 저장하게 된다.

(3) 스택을 이용한 그룹핑 단계

1) 지역그룹을 스택에 push

- 완전한 그룹화를 위해 helix 정보와 loop 스택의 크기정보를 저장한다.

2) 닫힌 helix ' ) '를 만나면, 스택에서 pop

- 열린 helix의 쌍인 닫힌 helix를 만나게 되면, 현재 loop을 먼저 저장된 loop 스택의 크기와 같이 되게 pop 시킨다.

3) 완전한 helix와 loop으로 재구성

- helix 와 loop의 그룹화를 한다.

### 3.3 시각화

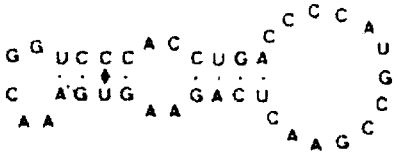
전처리과정을 거친 예측된 RNA 염기서열을 염기순서에 맞게 입력받아 앞서 말한 그룹화방법으로 번호가 정해진 helix와 loop 그룹으로 이루어진 연결리스트에서 사용자에 의해 선택되어진 loop이 먼저 그려지게 된다.

1) 사용자가 원하는 loop을 결정하게 된다.

2) 사용자가 선택한 loop이 결정되면 그에 인접한 loop과 그 사이 존재하게 될 helix들이 자동 선택되어진다.

3) 자동 결정된 연결 loop과 helix들은 연결리스트에 저장된 순서대로 차례대로 그려지게 된다.

4) 이미 그려져있는 loop중 더 많은 정보를 원하게 되는 loop을 선택하게 된다면 2)와 3)의 과정을 다시 반복하게 된다.



[ 그림 4 ] 기준 loop에 연결된 두 개의 loop

[그림 4]는 중간 loop을 사용자가 기준 loop으로 정했을 때의 경우이다. 선택된 가운데 loop에 인접한 두 개의 loop 중 연결리스트에 먼저 저장되어져있는 순서로 loop이 그려지게 된다. 처음 그려지는 기준 loop은 사용자에 의해 임의로 정해지는 것이다. 이렇게하여 사용자가 연이어 이어져 있는 loop을 선택하게 된다면 더 길게 연결되어지는 이차구조를 확인할 수 있다. 이때 그리는 순서는 역시 기준 loop에 연결되어있는 loop중에 연결리스트의 저장된 순서대로 그려지게 된다.

결국, 사용자는 RNA 이차구조에서 부분적구조의 연결정보를 알고 싶은 loop에 기준을 두어 원하는 구조정보를 확인할 수 있게된다.

### 4. 결론 및 향후과제

본 논문에서는 복잡한 전체이차구조를 표현하기보단 확인이 필요없을지 모르는 구조를 제외한 사용자가 원하는 필요한 부분만을 선택하여 그에 인접한 연결구조들만을 부분적 시각화하는 시스템을 설계하였다.

기존의 알고리즘들은 모든 helix와 loop에 대해 한번 이상의 검색시간을 필요로 하게되고 겹침(정보손실)이 발생하게되면 재구성하여 다시 겹침이없는 방향으로 그리기위해 검색을 반복하는 과정에 의하여 시간의 효율성이 낮다. 또한 이러한 반복과정을 거치지 않고 겹침을 피하기위한 추가적인 탐색 알고리즘을 통해 전체구조를 시각화하는 반면, 이 시스템에서는 실제로 유용하게 사용될 수 있는 부분적인 염기서열의 이차구조를 동적으로 표현하기 때문에 원하는 부분만의 검색시간만을 필요로하게 되고, 겹침을 피하기위한 새로운 탐색 알고리즘을 필요로하지 않게되므로 시간적 효율성과 알고리즘의 복잡도면을 개선할 수 있으라고 본다.

현재 모든 helix와 loop에 대한 그룹핑과 연결리스트까지 구현을 마쳤으며 그룹들이 가지고 있는 연결정보에 의해 부분 시각화하는 구현연구를 진행 중에 있다.

### 5. 참고문헌

[1].Bruccoleri, R., E., and Heinrich, G. (1988) An improved algorithm for nucleic acid secondary structure display. *CABIOS*, 4, 167-173.

[2].Muller, G., Gaspin, Ch., Etienne, A., and Westhof, E. (1993) Automatic display of RNA secondary structures. *CABIOS*, 9, 551-561.

[3].Devereux, J., Haeberli, P., and Smithies, O. (1984) A comprehensive set of sequence analysis programs for the Vax. *Nucleic Acids Res.*, 12, 387-395.

[4].Shapiro, B. A., Maizel, J., Lipkin, L.E., Currey, K., and Whitney, C. (1984) Generating non-overlapping displays of nucleic acid secondary structure. *Nucleic Acids Res.*, 12, 75-88.

[5].김도형, 한경숙 (1998) 벡터에 기반한 휴리스틱을 이용한 RNA 이차구조의 시각화. 제 25회 한국정보과학회 추계학술발표회 논문집, 25권, 2호, 633-635.

[6].김도형, 한경숙 (1999) RNA 이차구조의 시각화와 편집 한국정보과학회 논문지, 26권, 5호, 539-548.