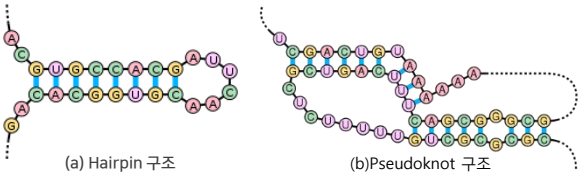
RNA 2차 구조 예측

AJOU SOFTCON 2021-2

이 름 권도현 지도교수 조다정 교수님

연구 배경

생물체의 유전 정보인 DNA, RNA의 염기 서열 분석 기술의 발전과 함께, 유전 정보와 질병 관계 분석이 중요해지고, 이를 이용하여 개인 유전체 분석, 질병의 조기 발견, 예방, 진단, 치료연구가 활발히 이루어 지고 있다. DNA 나 RNA를 이루는 분자는 외부의 개입 없이 간단한 구조의 분자들이 스스로 상호작용하여 복잡한 구조를 형성하는데 이는 질병과도 직접적인 연관이 있어, 이의 효율적인 탐색 기법 설계가 필요하다. 이 중에서도 본 연구는 RNA의 2차 구조의 탐색 알고리즘 설계를 목표로 하였다.



(a) Hairpin 구조 (b)Pseudoknot 구조

RNA는 DNA의 일부가 전사되어 형성되기 때문에 염기 4종류 (A, G, U, C) 가 이어져 있는 단일 가닥 형태를 지니고 있다. 그러나 RNA를 이루는 각 염기가 Watson-Crick pair (A-U, C-G) 나 wobble pair (G-U)와 같은 수소 결합에 의한 염기 쌍을 형성하면서 RNA는 고차원 구조를 가지게 된다. 염기쌍을 이룬 연속된 염기를 stem 구조라고 하고, 염기쌍을 이루지 못한 연속된 염기를 loop구조라고 한다.

RNA 2차 구조는 크게 2가지로 나뉘는데, 1개의 stem과 1개의 loop로 구성되어 있는 구조인 hairpin (stem-loop) 구조와, 최소 2개의 hairpin 구조 서로 맞물려 있는 pseudoknot 구조로 이루어져 있다

연구 진행 과정

해결하고자 하는 문제를 정확히 정의하고, pseudoknot이라는 다소 복잡한 구조의 이해를 위해 다음과 같은 문제들을 해결하였다.

1. 제한된 pseudoknot 결정 문제: Pseudoknot 구조는 hairpin 구조에 비해 다소 복잡해 효율적인 탐색이 까다롭기 때문에, 우선 주어진 RNA 염기 서열 w에서 pseudoknot 구조의 존재에 대한 결정 문제에 대해 탐구 하였는데, 이 문제는 greedy하게 pseudoknot을 이루는 모든 stem의 길이가 1이라고 가정하면 $O(N)$에 답을 구할 수 있다.

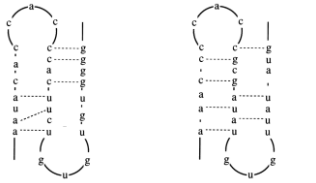
2. Stem의 길이가 최소 K인 pseudoknot 탐색: 여러 요소를 고려했을 때, 일반적인 상황에서 길이가 1인 Stem은 열역학적으로 안정하지 않을 가능성이 높아 형성되기 어렵다. 따라서 stem의 길이의 최솟값을 변수로 두어 앞의 문제를 확장하였는데, KMP (Knuth-Morris-Pratt) 알고리즘을 응용하면 범위가 제한된 입력에서는 $O(N)$, 전체 입력에 대해서는 $O(N^3)$에 pseudoknot 구조의 존재 유무를 확인할 수 있었다.

앞선 연구에서, 염기 서열에서 pseudoknot의 존재의 결정문제는 쉽게 해결됨을 알 수 있었다. 가능한 모든 2차 구조의 탐색 문제는 NP-hard 집합에 포함되어 있음이 밝혀져 있기 때문에, 대신 그 중에서도 열역학적으로 가장 안정된 구조를 찾는 문제인 MFE(Minimum free energy), 즉, 최대의 염기쌍을 지니는 구조 탐색) 가 RNA 2차 구조 예측의 주된 문제이다. 몇가지의 제한을 추가한 후 이 MFE를 가지는 2차 구조를 찾는 알고리즘의 설계를 본 연구의 최종 목표로 설정하였다.

MFE를 고려한 2차 구조의 탐색 문제의 경우 DP(Dynamic programming) 을 이용하여 최대 염기쌍을 계산하는 $O(N^4)$ 알고리즘이 존재한다. 하지만 해당 알고리즘에서 계산하는 2차 구조는 자연적으로 형성되기가 힘들다는 한계가 있는데, 본 연구에서는 보다 현실적인 탐색을 위해 다음과 같이 문제를 정의하였다.

RNA 염기서열 w가 주어졌을 때, Watson-Crick pair, wobble pair로 형성되는 RNA 2차 구조의 MFE를 계산하는 알고리즘을 설계한다. 이때,

1. Hairpin, pseudoknot 구조의 stem은 부분적인 loop를 포함할 수 있다.
2. Pseudoknot 구조를 이루는 2개의 hairpin 구조의 stem은 서로의 범위를 overlap 해서는 안된다.



(좌) 본 연구에서 고려하는 pseudoknot 구조
(우) 본 연구에서 고려하지 않는 pseudoknot 구조

3. Hairpin구조와 pseudoknot구조가 일자로 연결된 경우만 고려한다, 즉, 이미 연결된 여러 컴포넌트들을 둘러싸는 hairpin과 같은 구조는 고려하지 않는다.

다음과 같은 공간에 부분 해를 저장하여 전체 RNA 염기 서열에 대한 전체 최적해를 계산할 수 있다.

$Pseudoknot(l,\ r)$: 구간 (l,r)의 pseudoknot 구조의 최대 염기 쌍 수
$Hairpin(l,\ r)$: 구간 (l,r)의 hairpin 구조의 최대 염기 쌍 수
$Structure(l,r)$: 구간 (l,r)에 대한 부분 해

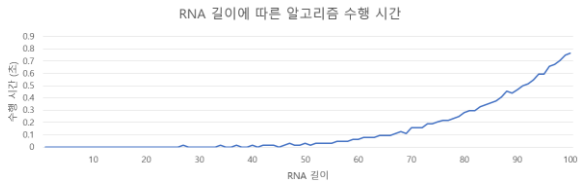
$Pseudoknot$과 $Hairpin$은 Longest Common Subsequence (LCS) 알고리즘을 응용하면 계산할 수 있다.
$Structure(1,|w|)$ 은 전체 문제의 해가 된다. 최적해를 가지는 염기 쌍 정보는 DP 역추적용 공간을 별도로 정의하여 복원할 수 있다.

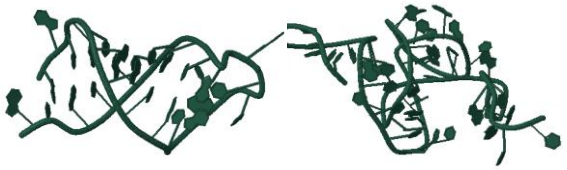
```
UCGACUGUAAAGCGGCGACUUUCAGUCGCUCUUUUUGUCGCGCGC
.()([])()(([[[[[[[[[]).]]]]]].]....].](()))
```

Dot-Bracket notation을 사용하여 복원된 RNA 염기 서열

결과 및 분석

시간 복잡도: $O(N^5)$ / 공간 복잡도: $O(N^4)$





복원된 RNA 염기 서열을 이용한 3차원상의 RNA 시각화
(좌) Hairpin 구조만 등장하는 RNA
(우) Pseudoknot 구조를 포함하는 RNA

본 연구의 한계점과 및 후속연구의 필요성에 대한 분석: 고안해낸 알고리즘의 시간적 공간적 비용이 높기 때문에 긴 RNA 데이터에 대해서는 해를 계산하는데 상당히 많은 시간과 저장공간이 필요할 것이다. 따라서 본 연구에서 제안하는 알고리즘의 최적화가 후속연구의 주제가 될 것이다. 또한 "열 역학적인 안정 " 에 대한 엄밀한 정의를 내리기 위한 에너지 함수에 대한 연구 및 Watson-Crick 과 wobble pair의 관계에 대한 연구가 필요성을 느꼈다.