AJOU SOFTCON 2021-2

자기주도연구

RNA secondary structure prediction

팀 원 권도현

지도교수 조다정 교수님

아주대학교 | 소프트웨어학과 / 소프트웨어교육센터

소개

권도현 (201920771)

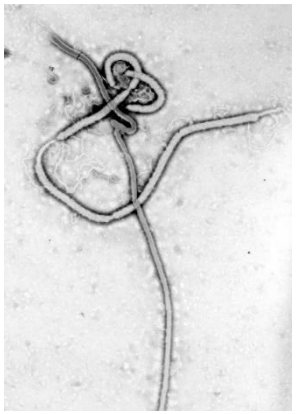
- 소프트웨어학과 3학년 재학
- 자기주도연구1 수강
- 지도교수: 조다정

1. 연구 배경 (개요)

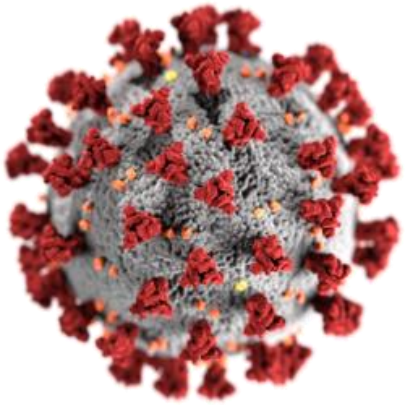
- RNA를 유전체로 가지는 RNA 바이러스들은 변종이 매우 쉽게 발생하고
- 백신 개발에도 RNA가 사용 됨

↓

더욱 효율적인 RNA 분석 연구의 필요

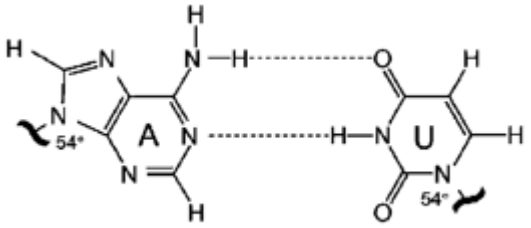


에볼라 바이러스

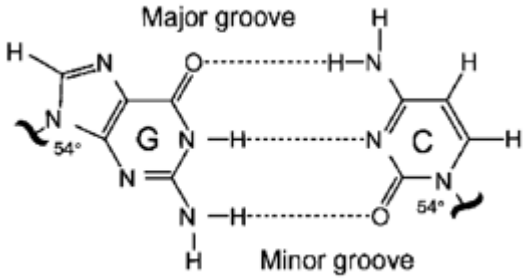


SARS-CoV-2

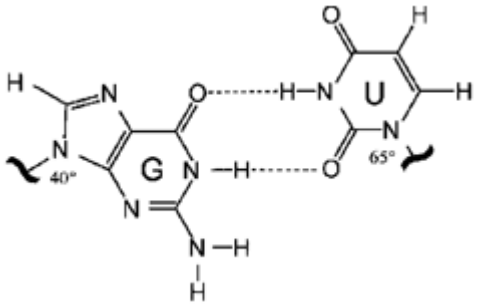
1. 연구 배경 (염기 쌍 형성 규칙)



Watson-Crick pairing (A-U)



Watson crick pairing (G-C)



Wobble pairing (G-U)

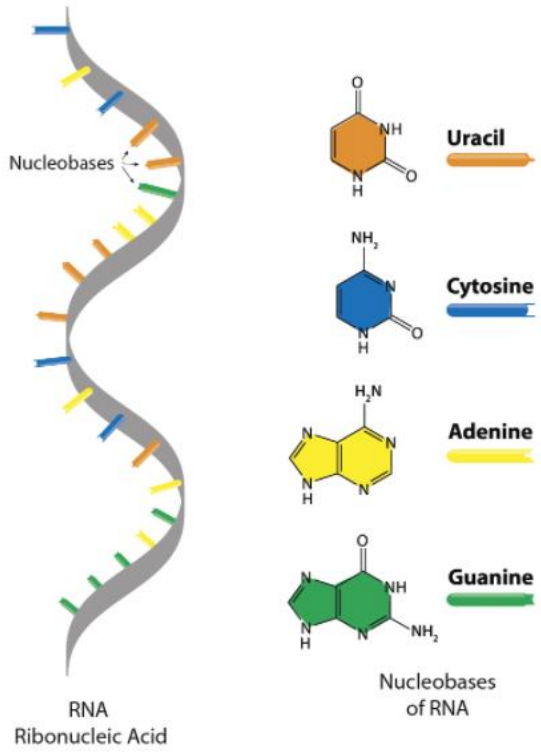
RNA의 염기 (Base)

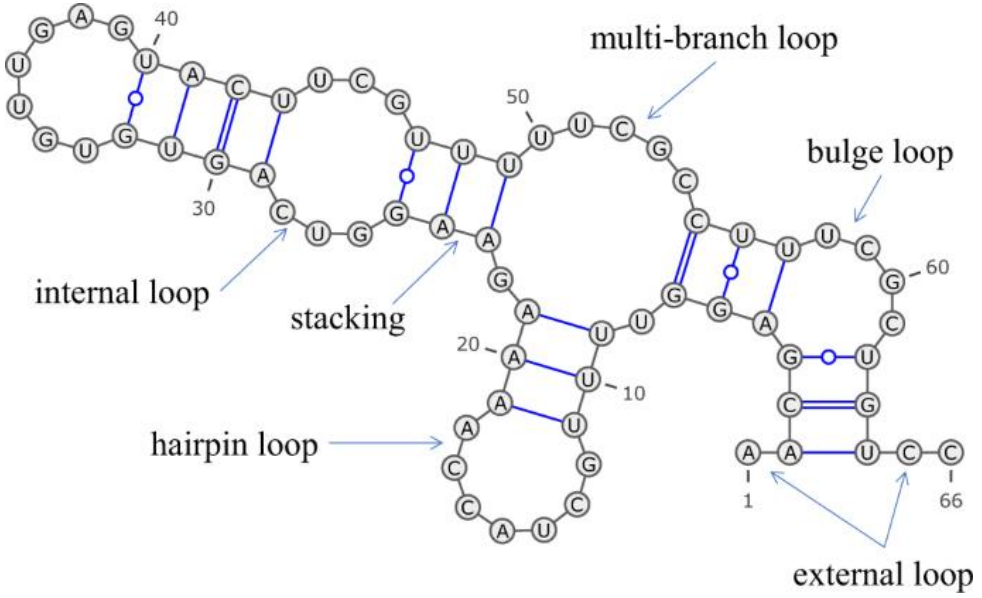
- 아데닌 (A, Adenine)
- 우라실 (U, Uracil)
- 구아닌 (G, Guanine)
- 사이토신 (C, Cytosine)

RNA의 염기 쌍 형성 규칙

- Watson-Crick pairing
- Wobble pairing

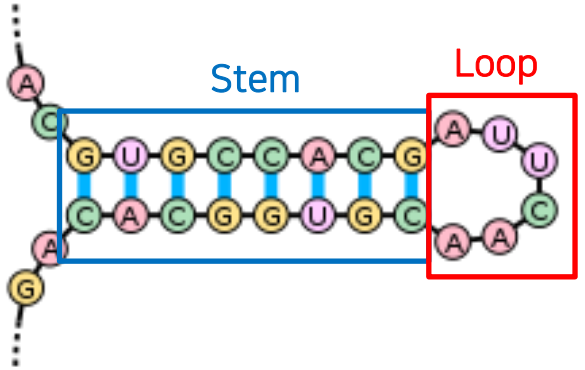
1. 연구 배경 (구조의 형성)


Nucleobases
RNA
Ribonucleic Acid
Uracil
Cytosine
Adenine
Guanine
Nucleobases
of RNA

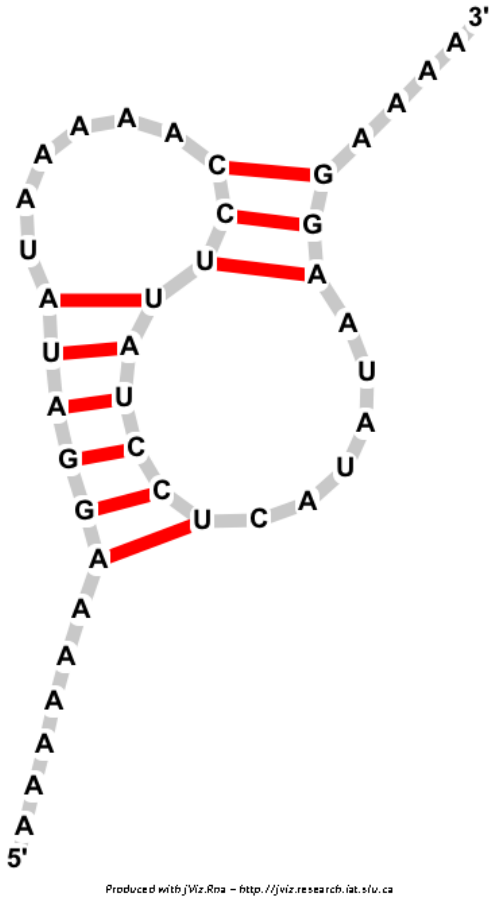

multi-branch loop
bulge loop
internal loop
stacking
hairpin loop
external loop
1
10
20
30
40
50
60
66

1. 연구 배경 (RNA 2차 구조)

- Stem: 염기쌍이 형성된 부분
- Loop: 염기쌍을 이루지 못한 부분
- Stem-Loop (Hairpin): Stem과 Loop 구조가 이어져 있는 구조
- Pseudoknot: 하나의 stem 구조의 절반이 다른 stem 구조에 삽입되어 있는 형태



Stem-Loop (Hairpin) 구조



Pseudoknot 구조

2. 연구 과정 (문제 정의)

Pseudoknot 구조 결정문제: 주어진 RNA sequence에 pseudoknot 구조가 존재할 수 있는지?

Simple pseudoknot 결정 문제: Greedy $O(N)$

Stem 길이 조건 추가

최소 K 길이의 Stem을 가지는 Pseudoknot 결정 문제: KMP $O(N^3)$

2. 연구 과정 (문제 정의)

- 단순 pseudoknot 존재의 결정문제는 상대적으로 빠른 시간 ($\sim O(N^3)$)에 해결 됨
- 가능한 모든 2차 구조 탐색은 NP-Hard

열역학적으로 가장 안정된 (Minimum free energy, MFE) 2차 구조 탐색 문제는 해결 가능

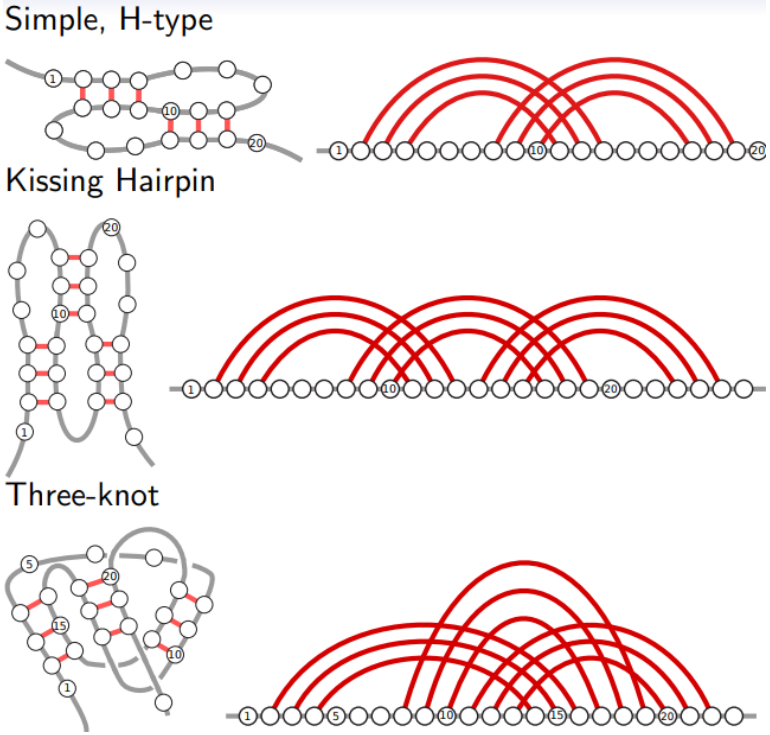
2. 연구 과정 (기수행 연구 분석)

class	R&G	A/U	L&P	D&P	CCJ	R&E
time	$O(n^4)$	$O(n^4)/O(n^5)$	$O(n^5)$	$O(n^5)$	$O(n^5)$	$O(n^6)$
space	$O(n^2)$	$O(n^3)/O(n^3)$	$O(n^3)$	$O(n^4)$	$O(n^4)$	$O(n^4)$

기수행 연구의 여러 알고리즘들

- 연구마다 허용하는 pseudoknot의 형태가 다름
- 범용적인 알고리즘일수록 시간/공간 복잡도 ↑
- 제한된 알고리즘일수록 시간/공간 복잡도 감소 ↓

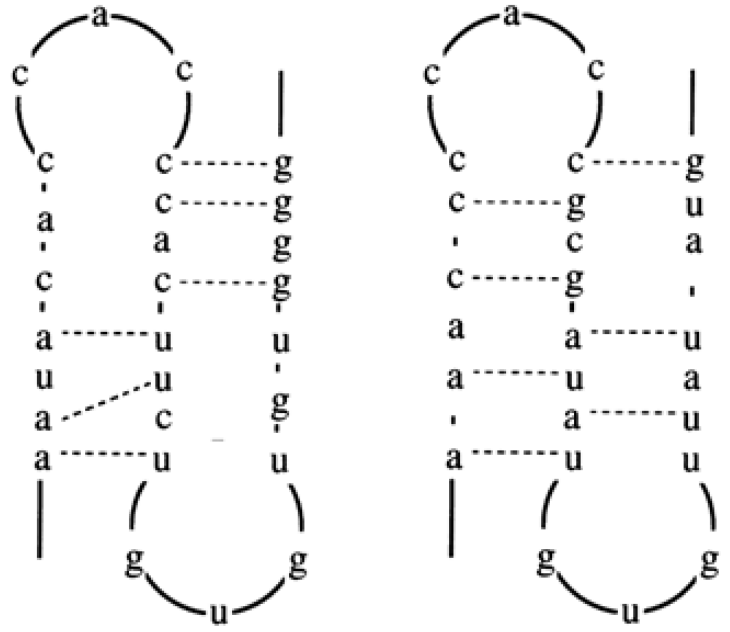
기수행 연구의 결과는 자연적으로 형성이 힘든 경우도 고려한다는 한계점이 있음



2. 연구 과정 (문제 정의)

현실적인 탐색을 위한 제한

1. Hairpin, pseudoknot 구조의 stem은 부분적인 loop가 허용된다 (mismatch)
2. Pseudoknot 구조를 이루는 2개의 hairpin 구조의 stem은 서로의 범위를 overlap 하는 경우는 고려하지 않음
3. Hairpin구조와 pseudoknot가 직렬로 연결된 경우만 고려
4. Wobble pair도 고려



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

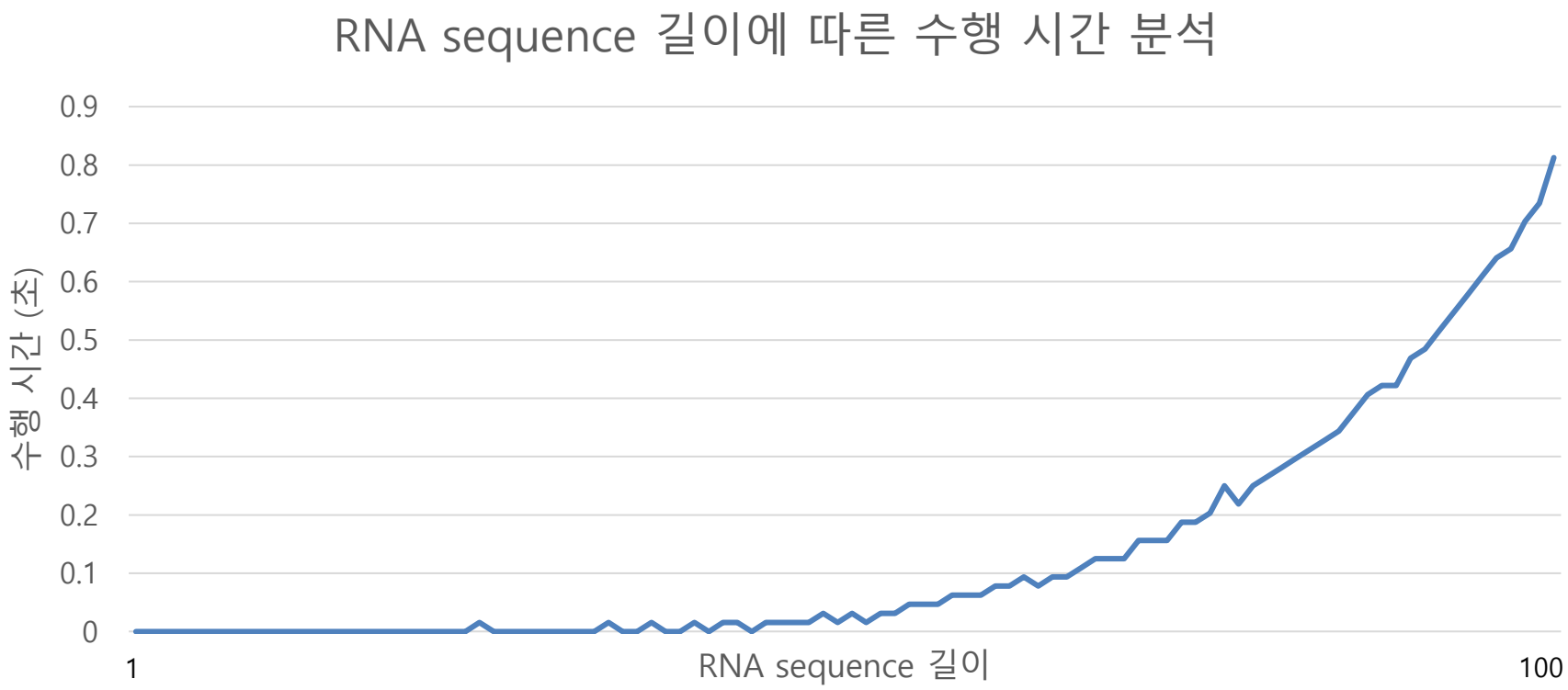
2. 연구 과정 (구현)

DP 공간 정의

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

전체 해: Structure(1, |RNA sequence|)

3. 연구 결과 (성능 분석)



$$\begin{cases} \boldsymbol{time\ complexity: O(N^5)} \\ \boldsymbol{space\ complexity: O(N^4)} \end{cases}$$

$$(\ \boldsymbol{N = |RNA\ sequence|}\ )$$

3. 연구 결과 (시각화)

RNACOMPOSER를 이용해 RNA를 시각화한 모습

```
ACGUGCCACGAUUCAACGUGGCACAG
.(((((((((.((.)))))))))).)
```

Hairpin 구조만 등장하는 RNA

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

Pseudoknot 구조를 포함하는 RNA

4. 연구 분석 및 추후 연구 계획

1. 여전히 시간 및 공간적으로 여전히 실용적이지 못함
   - 기존의 SOTA 알고리즘 ($O(N^4)$)에 비해 느린 이유는 본 연구만의 추가된 제한에 의함
   - 알고리즘의 최적화가 후속연구 주제

2. 더 현실적이고 범용성 있는 제한의 필요
   - 열 역학적인 관점에서의 Watson-Crick pair와 Wobble pair 차이에 대한 분석의 필요
   - 형성될 확률이 가장 높은 RNA 2차 구조?

AJOU SOFTCON 2021-2

감사합니다