The RNA Folding Problem

AN INTEGER LINEAR PROGRAMMING APPROACH

Problem Statement

The RNA Folding Problem is to predict the *secondary structure* of an RNA molecule, given its nucleotide sequence.
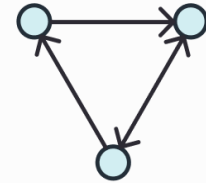
How?

$$\max c^\top x$$
$$\text{s.t. } Ax \le b$$
$$x \ge 0$$
$$x \in \mathbb{Z}^n$$

Integer Linear Programming

Dynamic programming

$$G = (V,E)$$

Graph theory
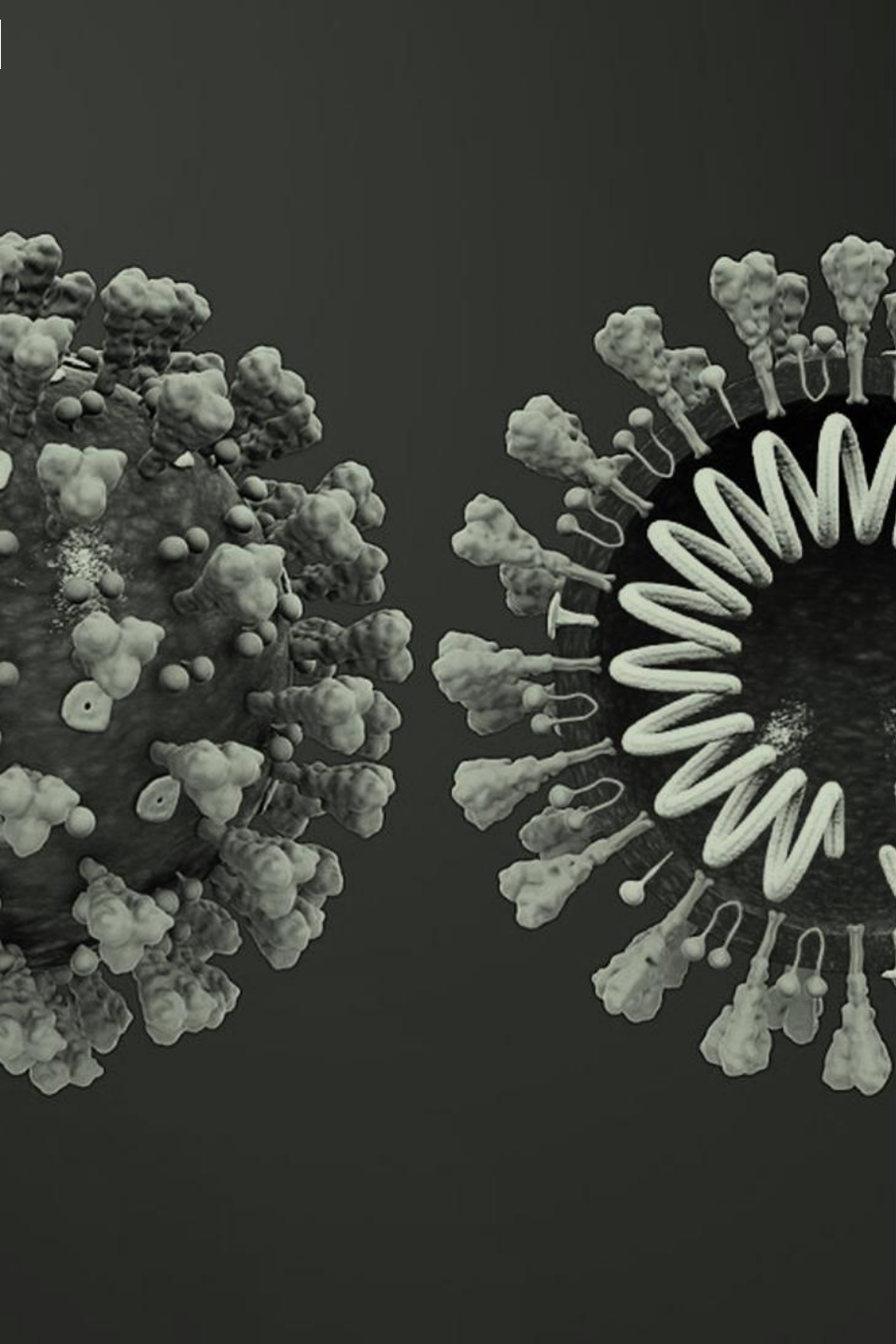
What about COVID-19?

- The Coronavirus genome is made up of a single strand of large positive polarity RNA

- RNA gives rise to 7 viral proteins and is associated with the N protein, which increases its stability

- To identify new antivirals it's necessary to know 3D structure of virus proteins that are responsible for cell infection and virus replication

- The enzyme forming the RNA chain of the virus is the ***RNA-dependent RNA polymerase***
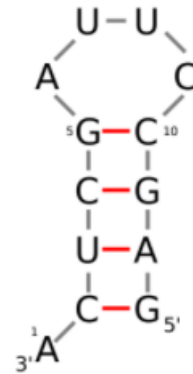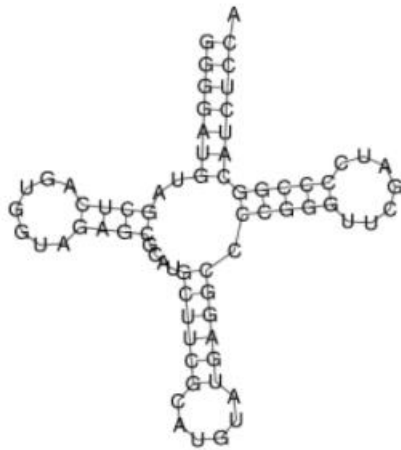
What about COVID-19?

The Linearfold algorithm:

- It predicts the secondary structure for the Covid-19 RNA sequence, reducing overall analysis time from 55 minutes to 27 seconds with respect the classical dynamic programming approach

- It achives *linear runtime*: in dynamic programming algorithms runtime scales cubically with RNA length

- It doesn't impose constraints on the output structure

- *Less use of memory* and *more scalability* with respect dynamic programming algorithms

Paper: https://academic.oup.com/bioinformatics/article/35/14/i295/5529205

Code: https://github.com/LinearFold/LinearFold

RNA secondary structure representations

A crude first model

Given the nucleotide sequence, s, of a RNA molecule, find a nested pairing that pairs the *maximum* number of nucleotides, compared to any other nested pairing.

A crude first model

$$Maximize \sum_{i<j} P(i,j)$$

→ *Maximize the number of paired nucleotides*

$$s.t. \sum_{j<k} P(j,k) + \sum_{k>j} P(k,j) \leq 1$$

→ *Each nucleotide can be paired to at most one other nucleotide*

$$P(i,j) + P(i',j') \leq 1$$

→ *Disallow crossing pairs*

$$P \in [0,1]$$

→ *Decision variable*
   0 : non complementary pair
   1 : complementary pair

A more complex model

$$Maximize \sum_{i<j} P(i,j) + Q(i,j)$$

→ *Maximize the number of paired nucleotides*

$$s.t. \sum_{j<k} P(j,k) + \sum_{k>j} P(k,j) \leq 1$$

→ *Each nucleotide can be paired to at most one other nucleotide*

$$P(i,j) + P(i',j') \leq 1$$

→ *Disallow crossing pairs*

$$P(i,j) + P(i+1,j-1) - Q(i,j) \leq 1$$

$$2Q(i,j) - P(i,j) - P(i+1,j-1) \leq 0$$

Stacked quartet detection

$$P \in [0,1]$$

→ *Decision variable*
   0 : non complementary pair
   1 : complementary pair

DEMO

https://github.com/badcortex/opt4ds