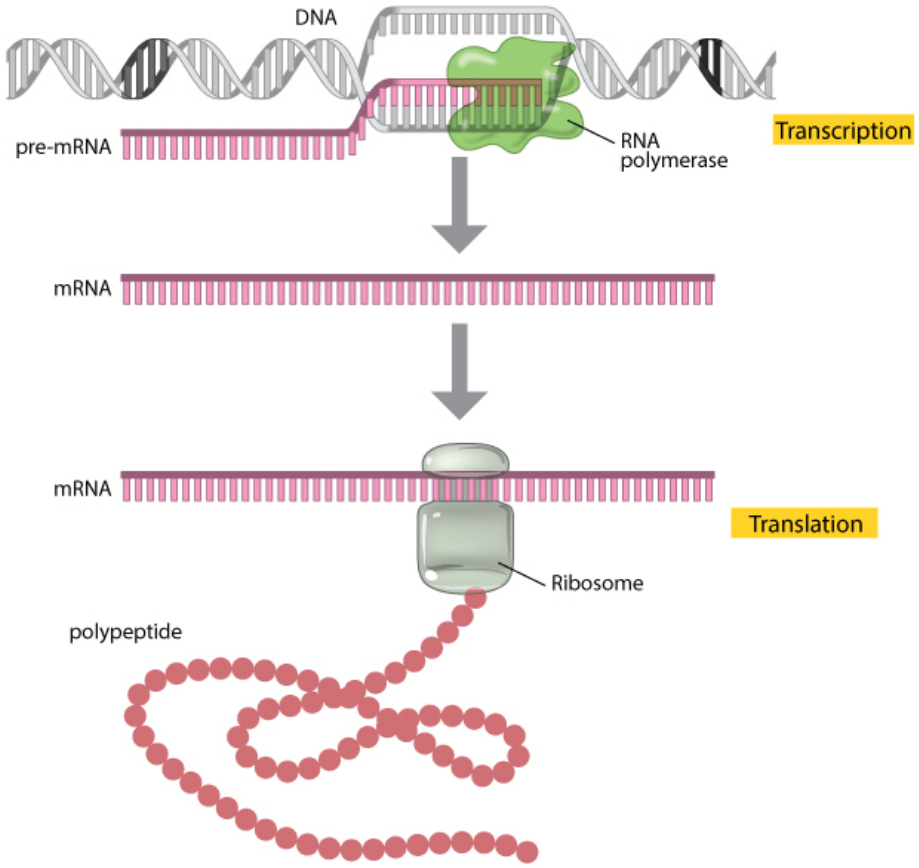
The RNA-folding Problem

Stefano Gualandi
Università di Pavia, Dipartimento di Matematica

email: stefano.gualandi@unipv.it
twitter: @famo2spaghi
blog: http://stegua.github.com

DNA, RNA, and proteins

https://www.nature.com/scitable/topicpage/translation-dna-to-mrna-to-protein-393



Predicting the secondary structure of an RNA molecule

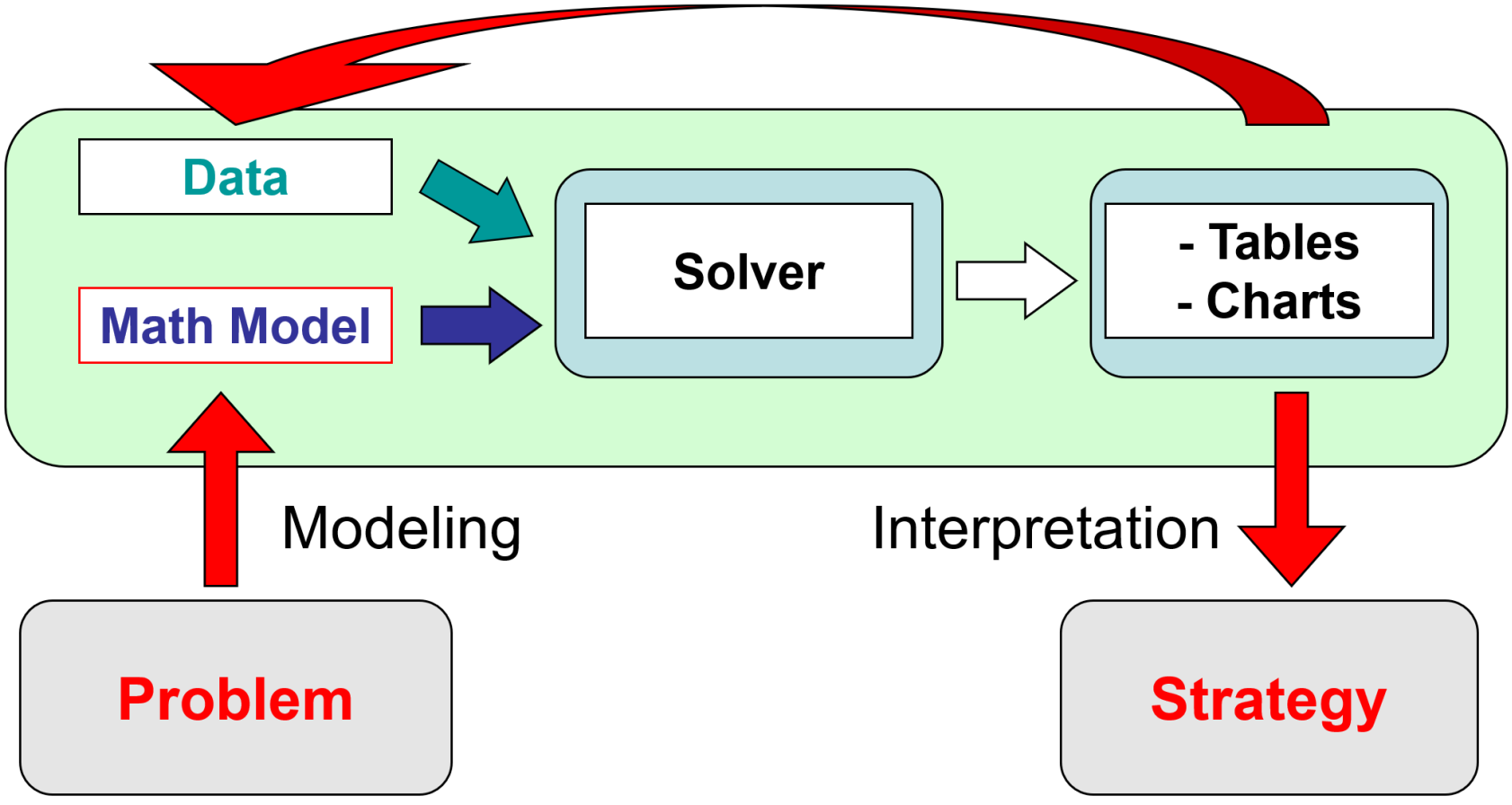
RNA Folding

RNA folding is the process by which a linear ribonucleic acid (RNA) molecule acquires secondary structure through intra-molecular interactions.

The folded domains of RNA molecules are often the sites of specific interactions with proteins in forming RNA-protein (ribonucleoprotein) complexes.

https://www.nature.com/subjects/rna-folding

Modeling as closed-loop activity



Problem Description: Basic approximation

Let s be a string of n characters of the RNA alphabet, that is:

$$s_i \in \{A, C, U, G\}, \quad \forall i = 1, \ldots, n$$

Example: $s = ACGUGCCCGAU$.

Problem Description: Basic approximation

Let s be a string of n characters of the RNA alphabet, that is:

$$s_i \in \{A, C, U, G\}, \quad \forall i = 1, \ldots, n$$

Example: $s = ACGUGCCCGAU$.

We define a pairing as a set of *disjoint* pairs of characters in s. Indeed, each character can be in at most a pair.

Problem Description: Basic approximation

Let s be a string of n characters of the RNA alphabet, that is:

$$s_i \in \{A, C, U, G\}, \quad \forall i = 1, \ldots, n$$

Example: $s = ACGUGCCCGAU$.

We define a pairing as a set of *disjoint* pairs of characters in s. Indeed, each character can be in at most a pair.

A **pair** is called complementary if the two characters in the pair are $\{A, U\}$ or $\{C, G\}$.

Problem Description: Basic approximation

Let s be a string of n characters of the RNA alphabet, that is:

$$s_i \in \{A, C, U, G\}, \quad \forall i = 1, \ldots, n$$

Example: $s = ACGUGCCCGAU$.

We define a pairing as a set of *disjoint* pairs of characters in s. Indeed, each character can be in at most a pair.

A **pair** is called complementary if the two characters in the pair are $\{A, U\}$ or $\{C, G\}$.

A **pairing** is called complementary if all of the matched pairs in the pairing are complementary.

Dan Gusfield. *Integer Linear Programming in Computational and System Biology*. Cambridge Press, 2019. (Chapter 6: RNA Folding)

The secondary structure

The secondary structure of an RNA molecule is specified by a pairing in its nucleotide sequence s.

The secondary structure

The secondary structure of an RNA molecule is specified by a pairing in its nucleotide sequence s.

In a first basic model of RNA folding, we require that any pairing be complementary.

The secondary structure

The secondary structure of an RNA molecule is specified by a pairing in its nucleotide sequence s.

In a first basic model of RNA folding, we require that any pairing be complementary.

This **constraint** appears because characters A and U represent complementary nucleotides that can **bond**; the same holds for C and G.

The secondary structure

The secondary structure of an RNA molecule is specified by a pairing in its nucleotide sequence s.

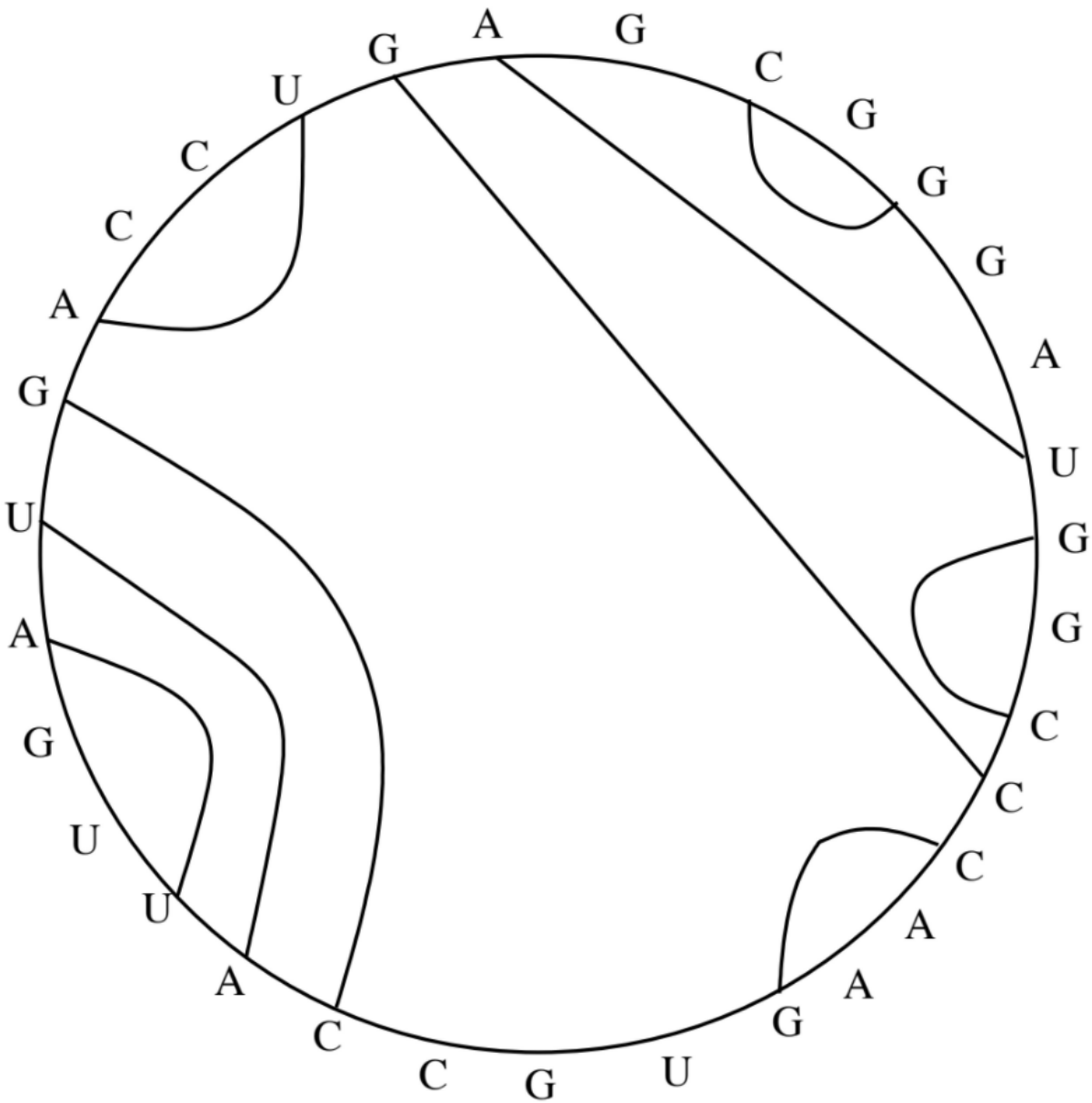
In a first basic model of RNA folding, we require that any pairing be complementary.

This **constraint** appears because characters A and U represent complementary nucleotides that can **bond**; the same holds for C and G.

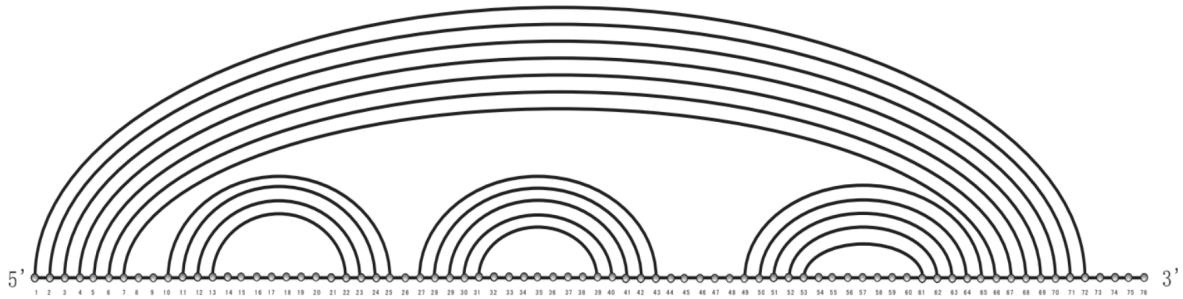
Definition 1 (Nested Pairing)

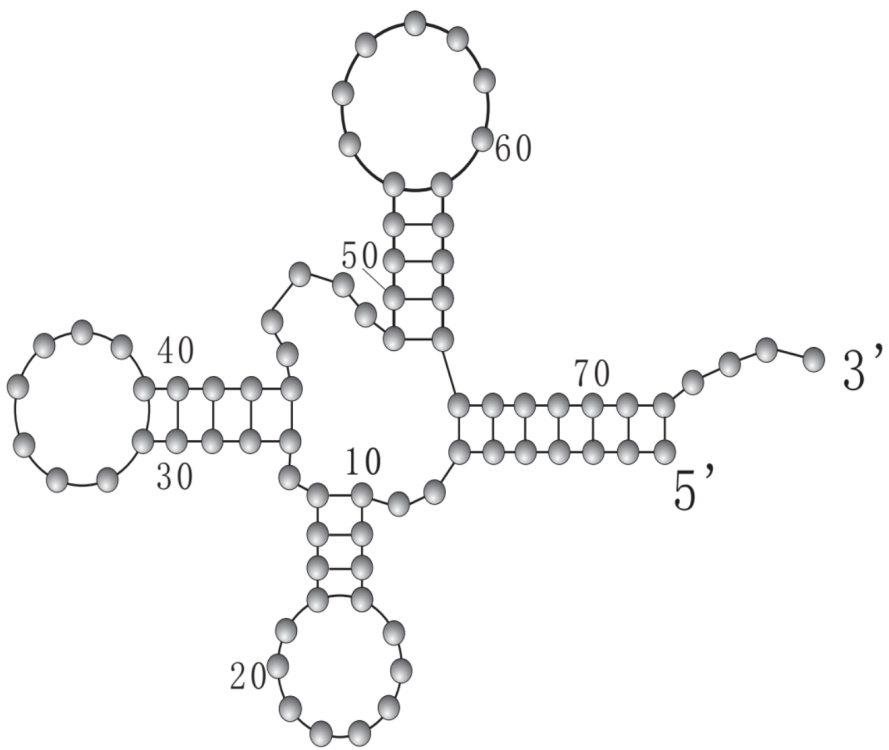
If we draw the RNA string s as a circular string, we define a nested pairing, also called a non-crossing pairing, a complementary pairing where each pair in the pairing is connected by a straight line inside the circle, and where none of the lines cross each other.

The secondary structure: Nested Pairing



The secondary structure: Nested Pairing





Fold stability

From biological and chemical considerations, it is generally asserted that *as a first approximation* the secondary structure (fold), of an RNA molecule is specified by a nested pairing that is the most stable, that is, the one with *minimum free energy*.

Fold stability

From biological and chemical considerations, it is generally asserted that *as a first approximation* the secondary structure (fold), of an RNA molecule is specified by a nested pairing that is the most stable, that is, the one with *minimum free energy*.

In our basic model, the stability is measured by the number of matched pairs. Hence, the most stable pairing will be the one with the largest number of matched pairs.

Fold stability

From biological and chemical considerations, it is generally asserted that *as a first approximation* the secondary structure (fold), of an RNA molecule is specified by a nested pairing that is the most stable, that is, the one with *minimum free energy*.

In our basic model, the stability is measured by the number of matched pairs. Hence, the most stable pairing will be the one with the largest number of matched pairs.

Problem: The simple RNA-folding problem

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.

Formulate this problem using Integer Linear Programming.

Data

$s = s_1, \ldots, s_n$	nucleotide sequence of a RNA molecule
$\{A, U\}, \{C, G\}$	complementary pairs
E	collections of ordered pairs $(i, j) = 1, \ldots, n$, such that $i < j$.

Data

$s = s_1, \ldots, s_n$	nucleotide sequence of a RNA molecule
$\{A, U\}, \{C, G\}$	complementary pairs
E	collections of ordered pairs $(i, j) = 1, \ldots, n$, such that $i < j$.

Decision Variables

$$x_{ij} = \begin{cases} 1 & \text{if } i, j \text{ are paired} \\ 0 & \text{otherwise.} \end{cases}$$

Data

$s = s_1, \ldots, s_n$	nucleotide sequence of a RNA molecule
$\{A, U\}, \{C, G\}$	complementary pairs
E	collections of ordered pairs $(i, j) = 1, \ldots, n$, such that $i < j$.

Decision Variables

$$x_{ij} = \begin{cases} 1 & \text{if } i, j \text{ are paired} \\ 0 & \text{otherwise.} \end{cases}$$

Objective: Max Pairs

$$\max \sum_{(i,j) \in E} x_{ij}$$

Data

$s = s_1, \ldots, s_n$ nucleotide sequence of a RNA molecule

$\{A, U\}, \{C, G\}$ complementary pairs

E collections of ordered pairs $(i, j) = 1, \ldots, n$, such that $i < j$.

Decision Variables

$$x_{ij} = \begin{cases} 1 & \text{if } i, j \text{ are paired} \\ 0 & \text{otherwise.} \end{cases}$$

Objective: Max Pairs

$$\max \sum_{(i,j) \in E} x_{ij}$$

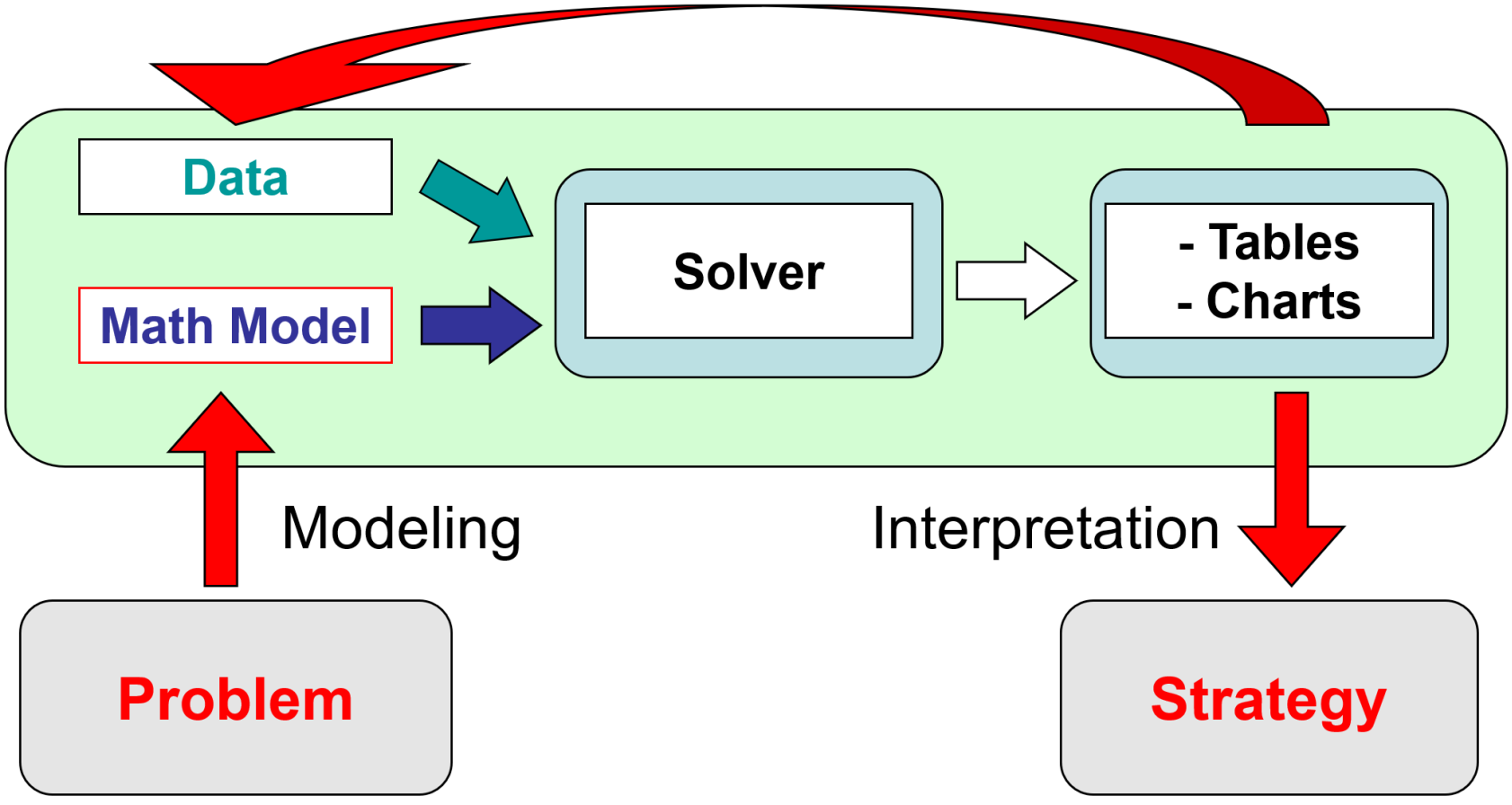
Constraints

$$x_{ij} = 0 \qquad \forall \{i, j\} \in E : \{s_i, s_j\} \neq \{A, U\} \text{ or } \{C, G\}$$

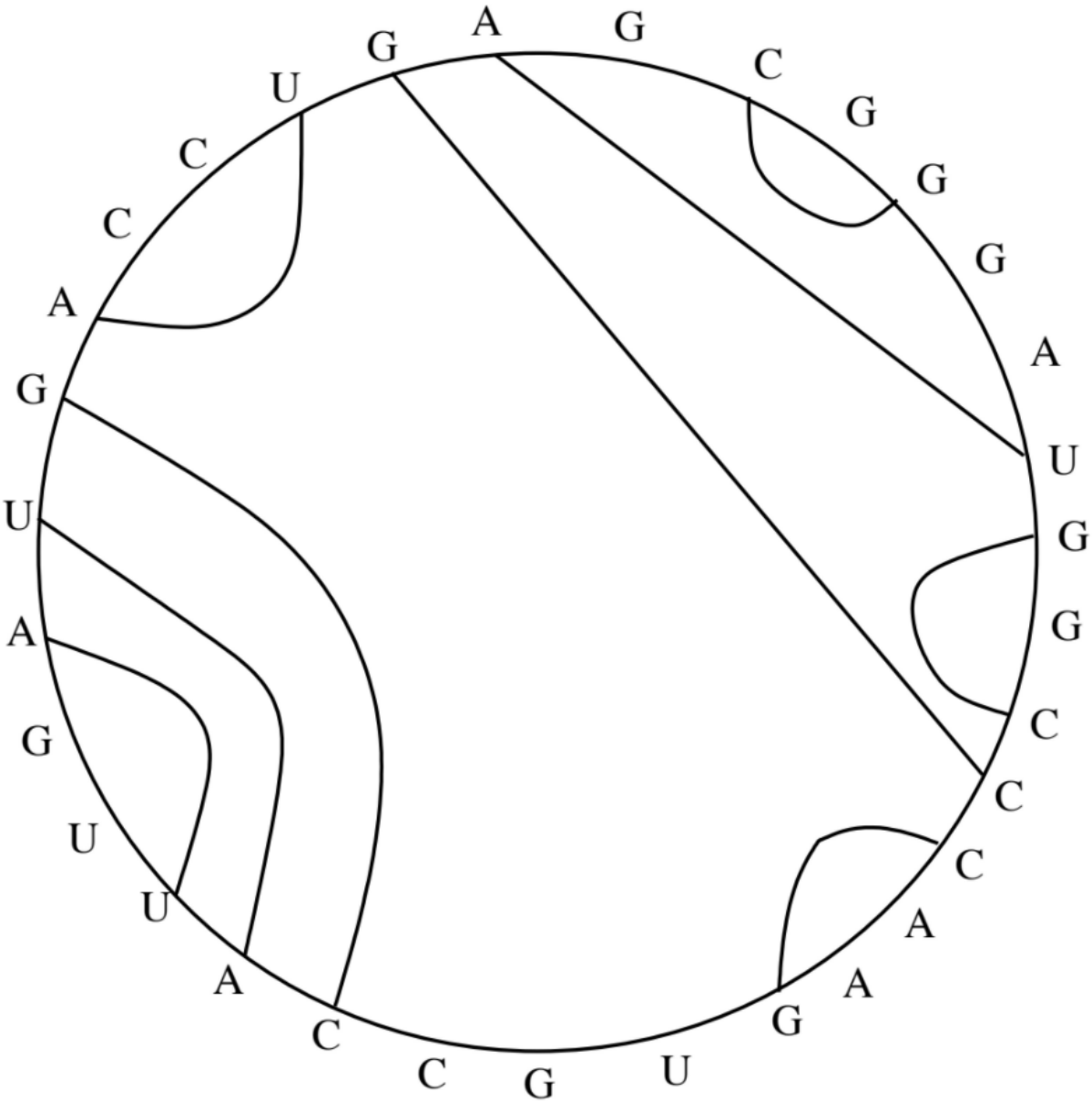
$$\sum_{(k,j) \in E} x_{kj} + \sum_{(j,k) \in E} x_{jk} \leq 1 \qquad \forall j = 1, \ldots, n \quad \text{(at most a pair)}$$

$$x_{ij} + x_{vw} \leq 1 \qquad \forall i, j, v, w : i < v < j < w.$$

Modeling as closed-loop activity



The secondary structure: Nested Pairing



Data

V	set of pairs in s for (s_i, s_j) are complementary
E	set of pairs of non-crosssing pair $\{(s_i, s_j), (s_v, s_w)\}$
$G = (V, E)$	auxiliary graph

Data

V	set of pairs in s for (s_i, s_j) are complementary
E	set of pairs of non-crosssing pair $\{(s_i, s_j), (s_v, s_w)\}$
$G = (V, E)$	auxiliary graph

Decision Variables

$$x_i = \begin{cases} 1 & \text{if pair } i \text{ is selected} \\ 0 & \text{otherwise.} \end{cases}$$

Data

V set of pairs in s for (s_i, s_j) are complementary

E set of pairs of non-crosssing pair $\{(s_i, s_j), (s_v, s_w)\}$

$G = (V, E)$ auxiliary graph

Decision Variables

$$x_i = \begin{cases} 1 & \text{if pair } i \text{ is selected} \\ 0 & \text{otherwise.} \end{cases}$$

Objective: Max Pairs

$$\max \sum_{i \in V} x_i$$

Data

V set of pairs in s for (s_i, s_j) are complementary
E set of pairs of non-crosssing pair $\{(s_i, s_j), (s_v, s_w)\}$
$G = (V, E)$ auxiliary graph

Decision Variables

$$x_i = \begin{cases} 1 & \text{if pair } i \text{ is selected} \\ 0 & \text{otherwise.} \end{cases}$$

Objective: Max Pairs

$$\max \sum_{i \in V} x_i$$

Constraints

$$x_i + x_j \leq 1 \qquad \forall \{i, j\} \in E$$

Maximum Clique

This problem defined over a generic graph $G = (V, E)$ is called the **Maximum Clique Problem** and is a fundamental problem of Combinatorial Optimization. It belongs to Karp's 21 NP-hard problems.

The Spectrum of Computational Complexity

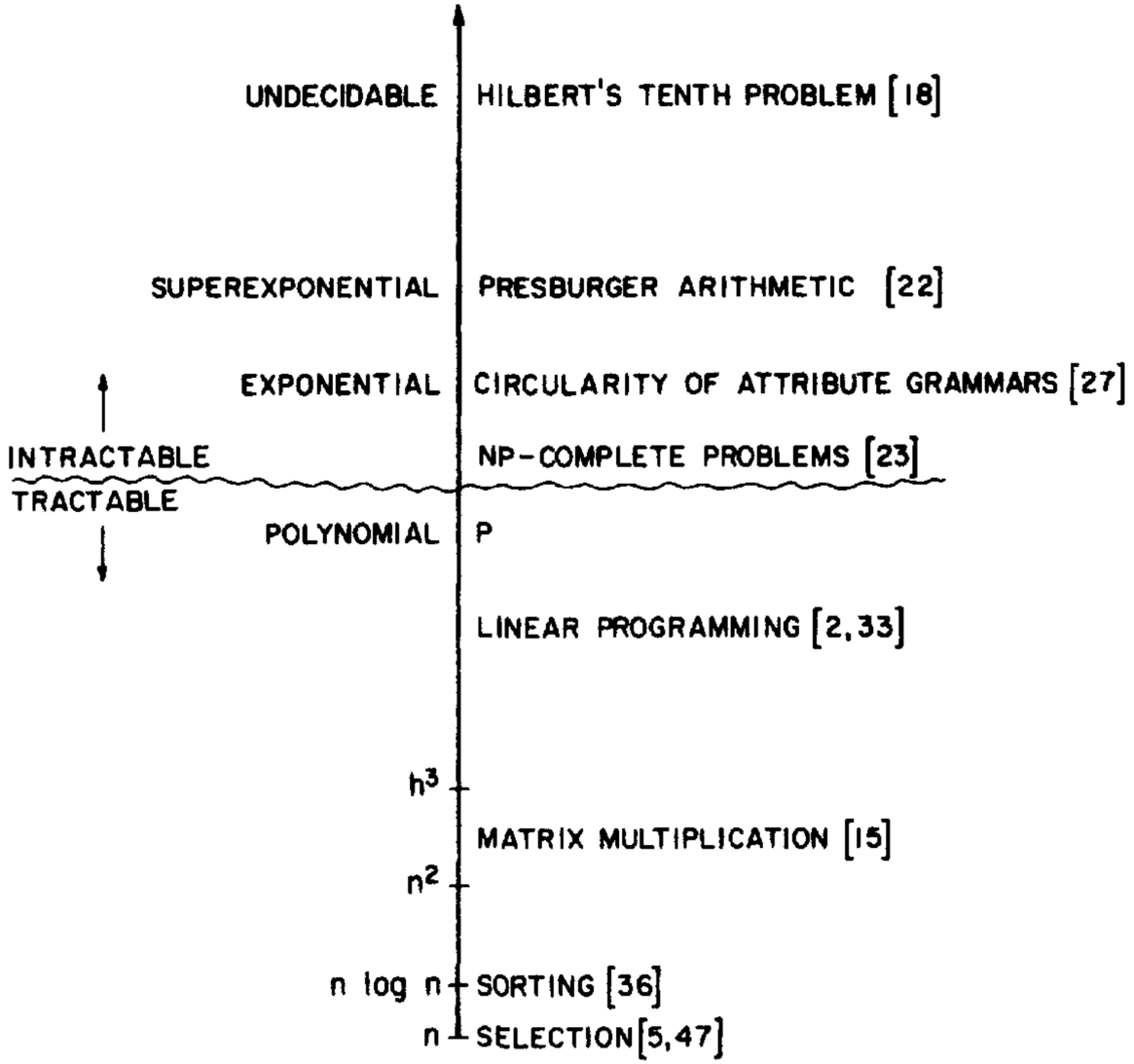


FIG. 1.3. *The spectrum of computational complexity.*

The Spectrum of Computational Complexity

COMPLEXITY \ SIZE	20	50	100	200	500	1000
$1000n$	.02 sec	.05 sec	.1 sec	.2 sec	.5 sec	1 sec
$1000n \lg n$	.09 sec	.3 sec	.6 sec	1.5 sec	4.5 sec	10 sec
$100n^2$	.04 sec	.25 sec	1 sec	4 sec	25 sec	2 min
$10n^3$	.02 sec	1 sec	10 sec	1 min	21 min	2.7 hr
$n^{\lg n}$	.4 sec	1.1 hr	220 DAYS	125 CENT	5×10^8 CENT	
$2^{n/3}$	.0001 sec	.1 sec	2.7 hr	3×10^4 CENT		
2^n	1 sec	35 YR	3×10^4 CENT			
3^n	58 min	2×10^9 CENT				

The Spectrum of Computational Complexity

TIME / COMPLEXITY	1 sec	10^2 sec (1.7 min)	10^4 sec (2.7 hr)	10^6 sec (12 DAYS)	10^8 sec (3 YEARS)	10^{10} sec (3 CENT.)
$1000n$	10^3	10^5	10^7	10^9	10^{11}	10^{13}
$1000n \lg n$	1.4×10^2	7.7×10^3	5.2×10^5	3.9×10^7	3.1×10^9	2.6×10^{11}
$100n^2$	10^2	10^3	10^4	10^5	10^6	10^7
$10n^3$	46	2.1×10^2	10^3	4.6×10^3	2.1×10^4	10^5
$n \lg n$	22	36	54	79	112	156
$2^{n/3}$	59	79	99	119	139	159
2^n	19	26	33	39	46	53
3^n	12	16	20	25	29	33

The advantage of ILP modeling: Flexibility

More involved models of the RNA-folding problem includes:

- **Minimum Distance Constraint.** More realistic models include the chemical constraints that a position (nucleotide) cannot pair with any position that is *too close*.

The advantage of ILP modeling: Flexibility

More involved models of the RNA-folding problem includes:

- **Minimum Distance Constraint.** More realistic models include the chemical constraints that a position (nucleotide) cannot pair with any position that is *too close*.

- **Differential Binding Strengths.** The binding strength of $\{C, G\}$ is largest than $\{A, U\}$, since it has three hydrogen bonds, instead of two. Hence, we can add weights in the objective function terms in order to formulate this *differential binding strength*.

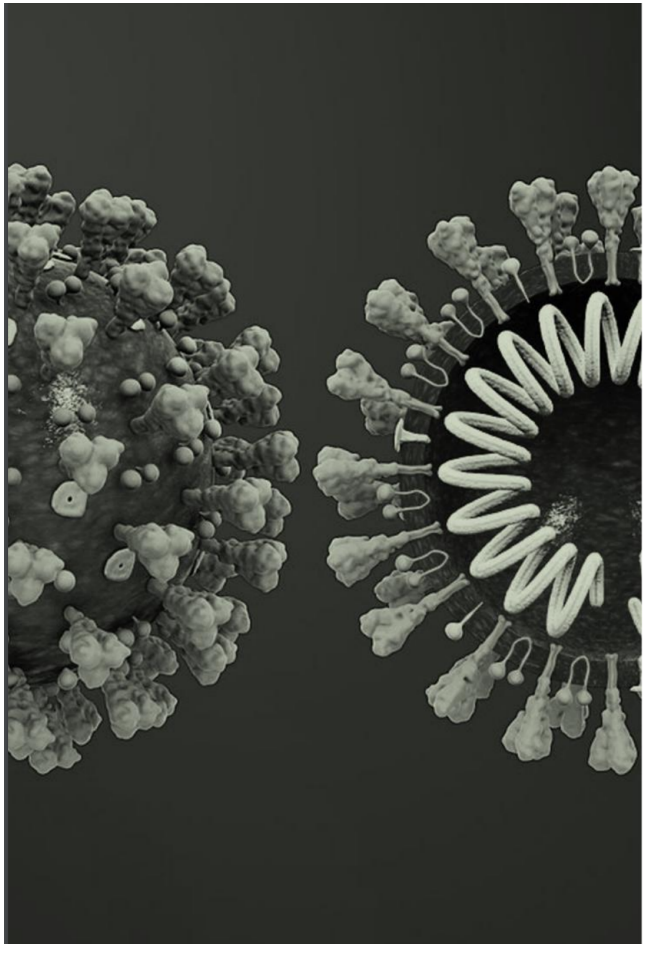
The advantage of ILP modeling: Flexibility

More involved models of the RNA-folding problem includes:

- **Minimum Distance Constraint.** More realistic models include the chemical constraints that a position (nucleotide) cannot pair with any position that is *too close*.
- **Differential Binding Strengths.** The binding strength of $\{C, G\}$ is largest than $\{A, U\}$, since it has three hydrogen bonds, instead of two. Hence, we can add weights in the objective function terms in order to formulate this *differential binding strength*.
- **Allowing Some Noncomplementary Matched Pairs.** In some models, also noncomplementary pairs can match, but they have a very low strength weights, and most of the time, only a given (small) percentage of noncomplementary pairs can be part of the finale solution (*EXERCISE: How can you model this extra condition?*)

Dan Gusfield. *Integer Linear Programming in Computational and System Biology*. Cambridge Press, 2019. (Chapter 6: RNA Folding)

RNA-folding, COVID-19, ... and final projects



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***

RNA-folding, COVID-19, ... and final projects

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