# RNA Bioinformatics and Combinatorial Dynamic Programming

## ...through enumerative combinatorics

Yann Ponty

LIX, CNRS/Ecole Polytechnique, France

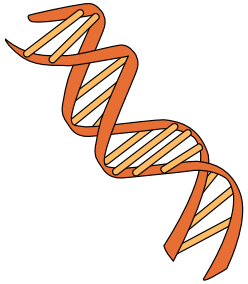
AMIB Team-Project, Inria, Saclay, France

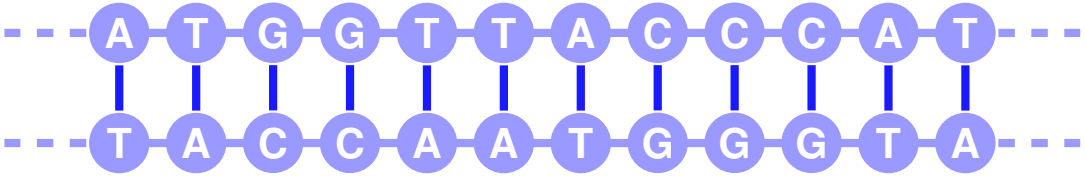
PIMS-Simon Fraser University, Burnaby, Canada

# Fundamental *dogma* of molecular biology

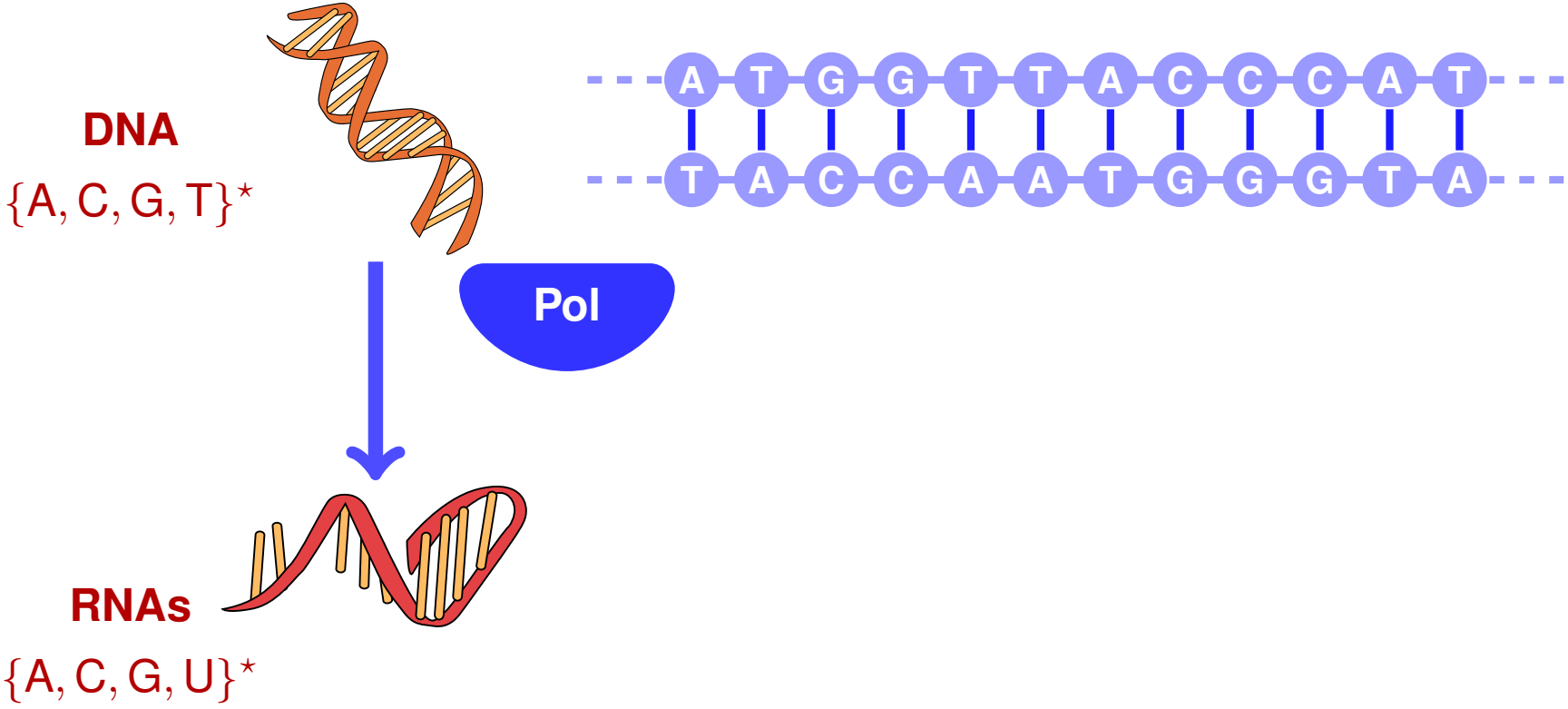
**DNA**

$\{A, C, G, T\}^\star$

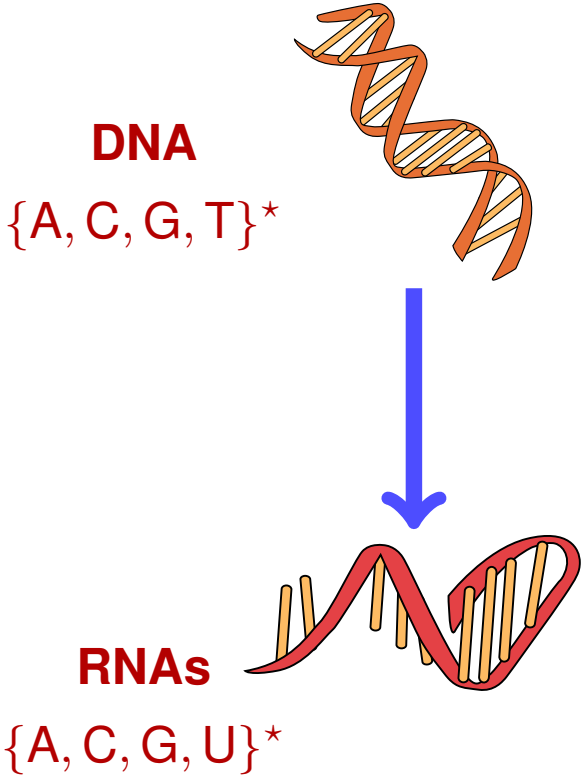


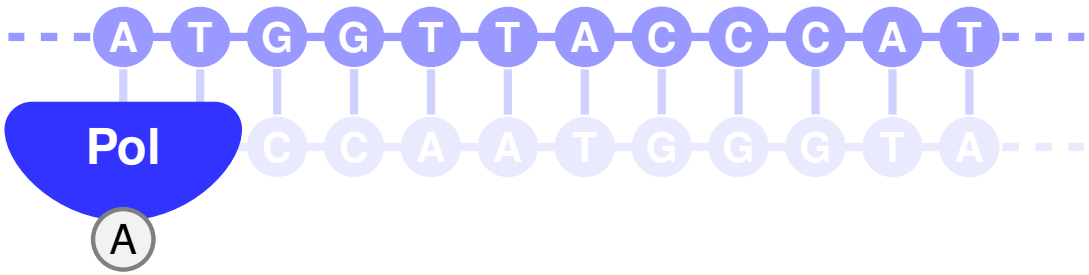


# Fundamental *dogma* of molecular biology

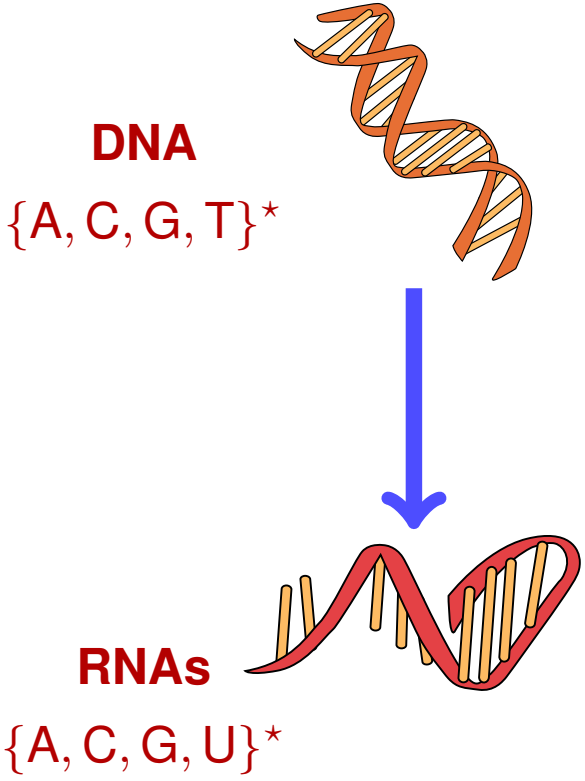


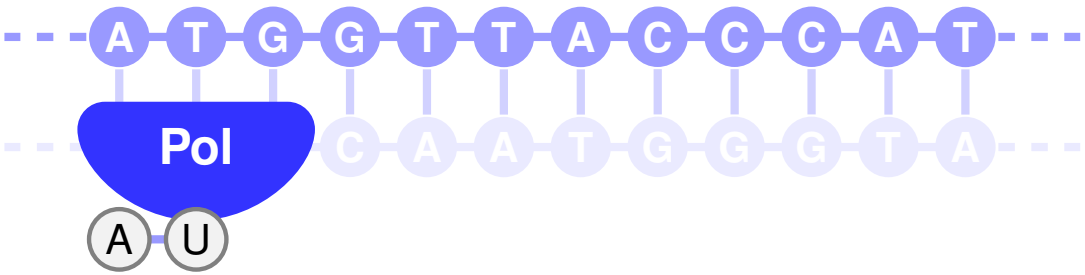
# Fundamental *dogma* of molecular biology



# Fundamental *dogma* of molecular biology

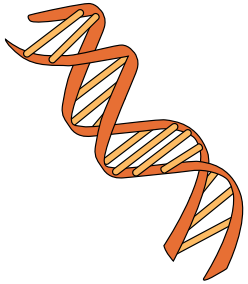


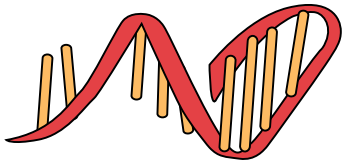



# Fundamental *dogma* of molecular biology

**DNA**

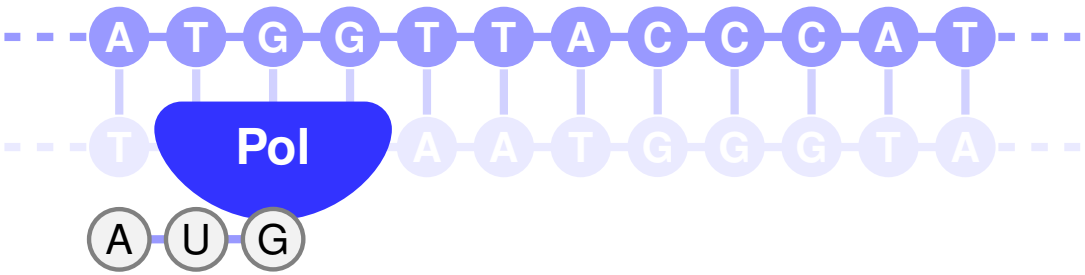
$\{A, C, G, T\}^\star$





**RNAs**

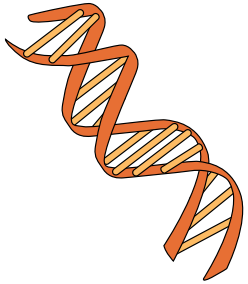
$\{A, C, G, U\}^\star$

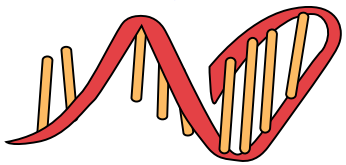


# Fundamental *dogma* of molecular biology

**DNA**

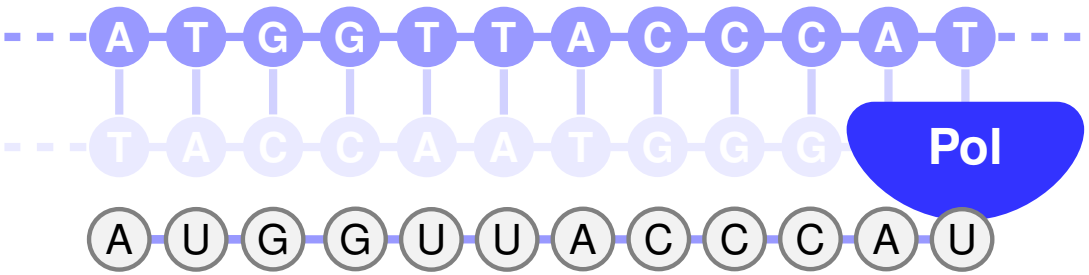
$\{A, C, G, T\}^\star$





**RNAs**

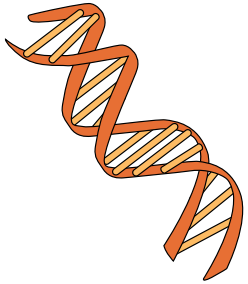
$\{A, C, G, U\}^\star$

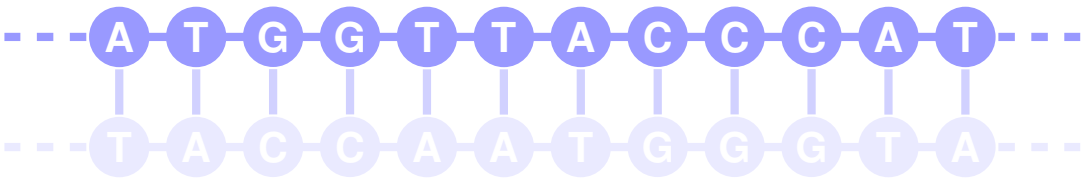


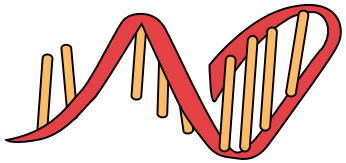
# Fundamental *dogma* of molecular biology

**DNA**

$\{A, C, G, T\}^\star$



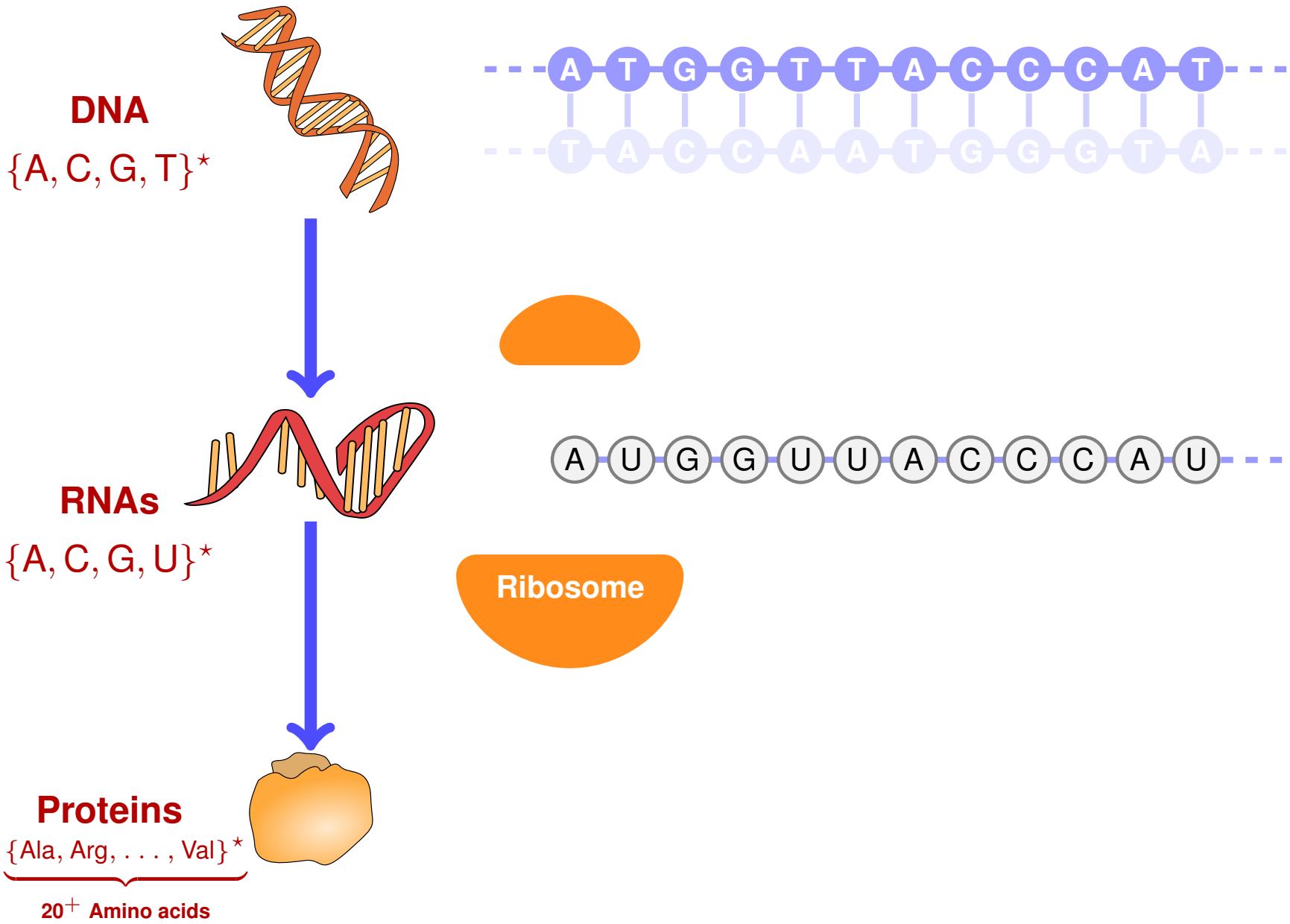


**RNAs**

$\{A, C, G, U\}^\star$

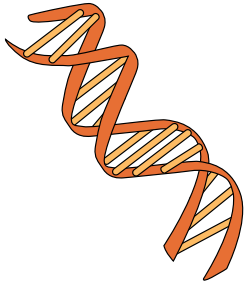
# Fundamental *dogma* of molecular biology

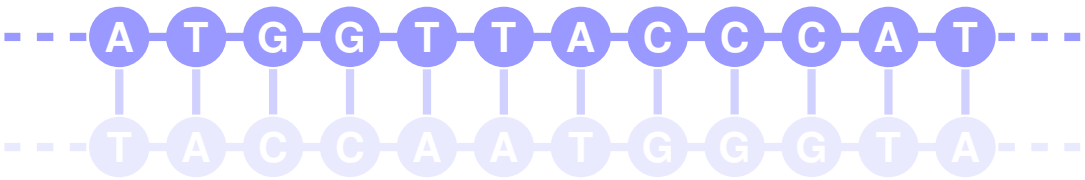


# Fundamental *dogma* of molecular biology

**DNA**

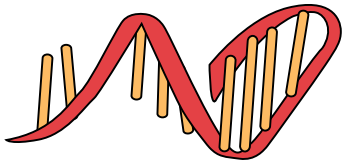
$\{A, C, G, T\}^\star$

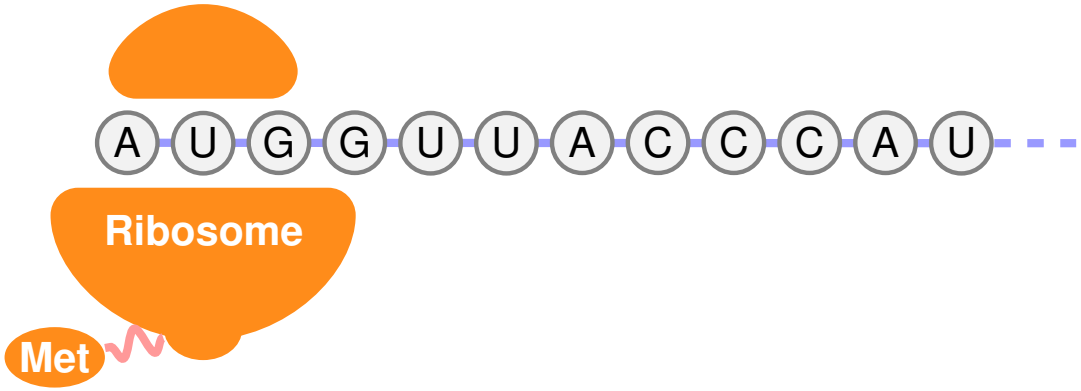





**RNAs**

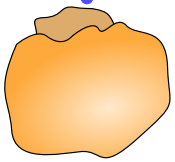
$\{A, C, G, U\}^\star$






**Proteins**

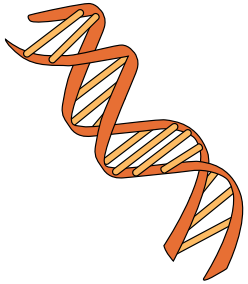
$\underbrace{\{\text{Ala}, \text{Arg}, \ldots, \text{Val}\}^\star}_{20^+ \text{ Amino acids}}$

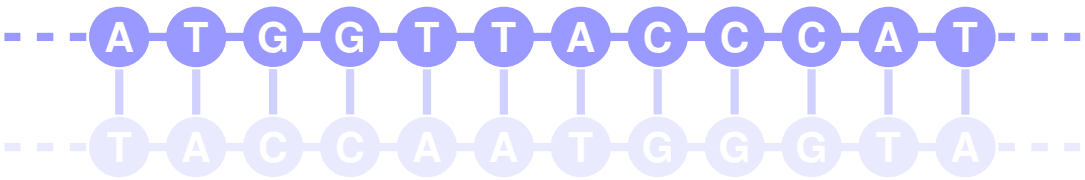


# Fundamental *dogma* of molecular biology

**DNA**

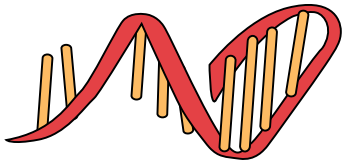
$\{A, C, G, T\}^\star$

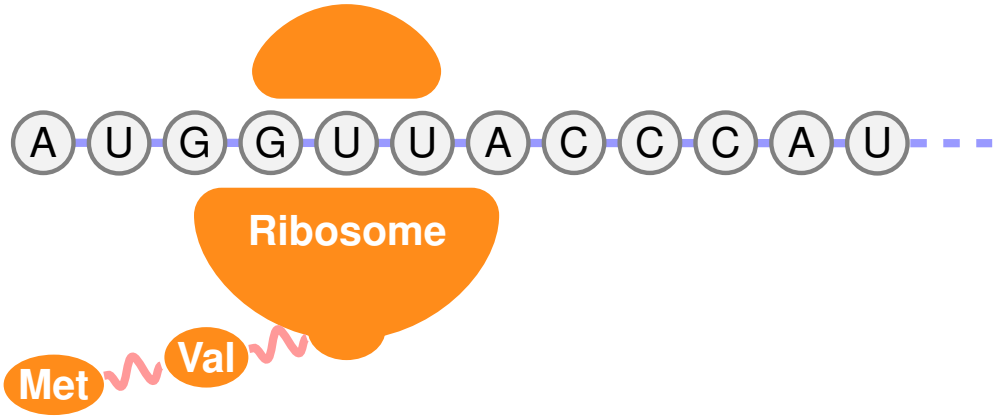





**RNAs**

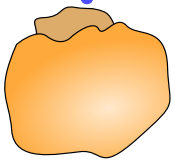
$\{A, C, G, U\}^\star$






**Proteins**

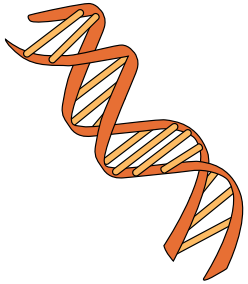
$\underbrace{\{\text{Ala}, \text{Arg}, \ldots, \text{Val}\}^\star}_{20^+ \text{ Amino acids}}$

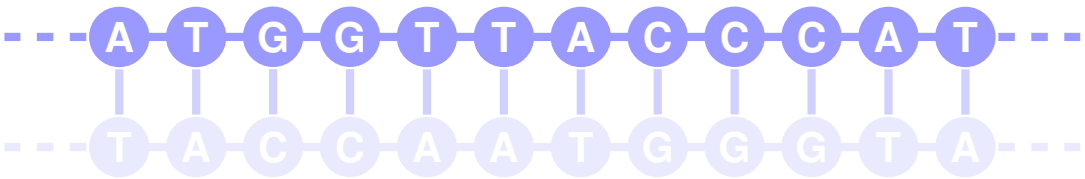


# Fundamental *dogma* of molecular biology

**DNA**

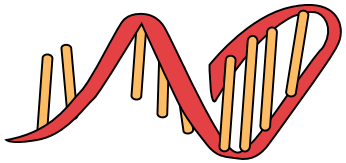
$\{A, C, G, T\}^\star$

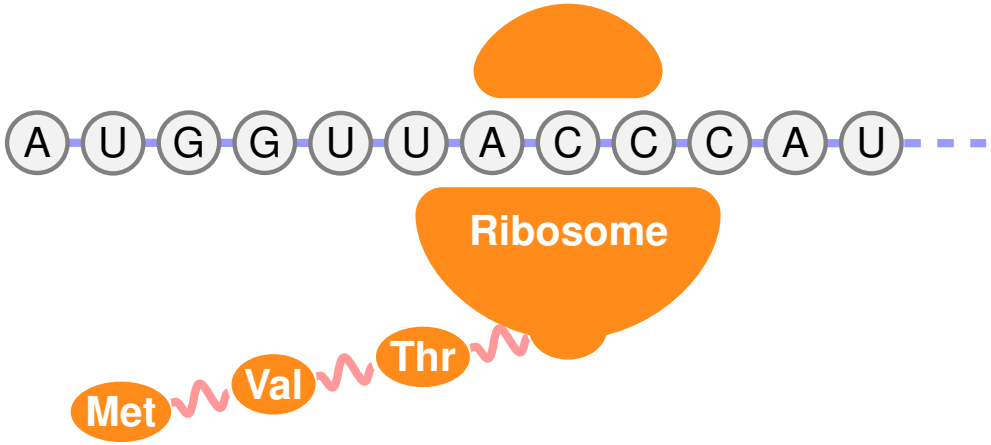





**RNAs**

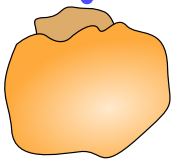
$\{A, C, G, U\}^\star$






**Proteins**

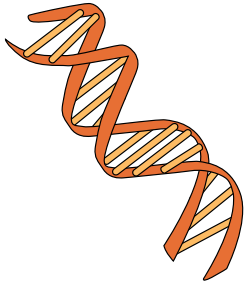
$\underbrace{\{\text{Ala}, \text{Arg}, \ldots, \text{Val}\}^\star}_{20^+ \text{ Amino acids}}$

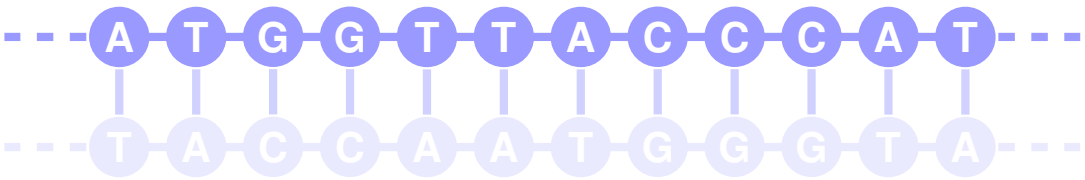


# Fundamental *dogma* of molecular biology

**DNA**

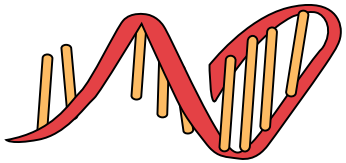
$\{A, C, G, T\}^\star$

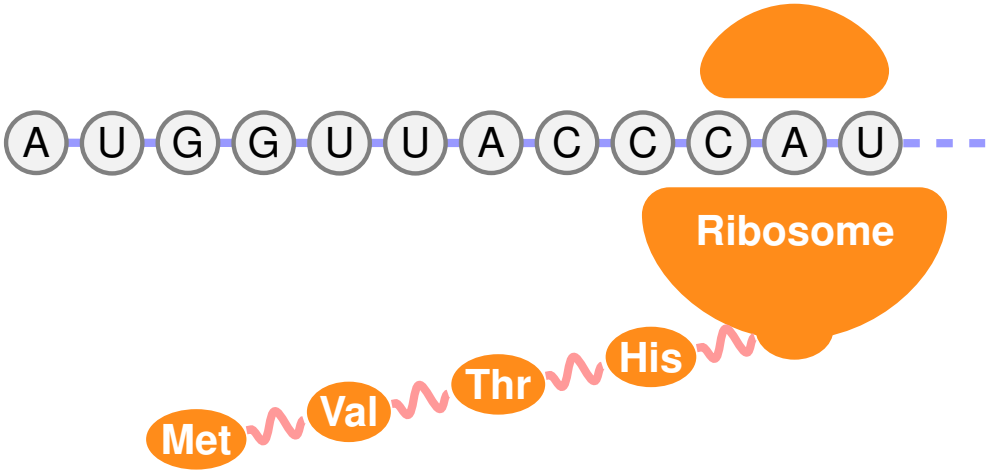





**RNAs**

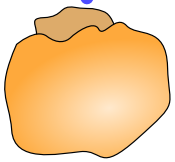
$\{A, C, G, U\}^\star$






**Proteins**

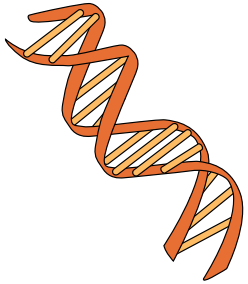
$\underbrace{\{\text{Ala}, \text{Arg}, \ldots, \text{Val}\}^\star}_{20^+ \text{ Amino acids}}$

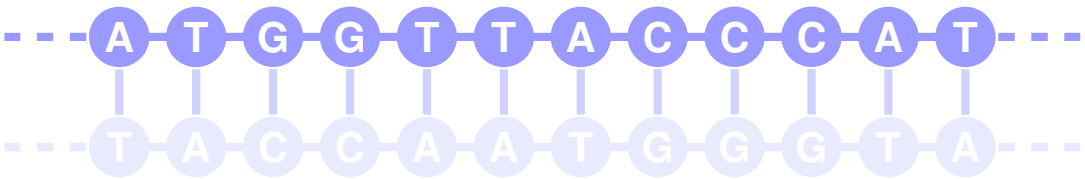


# Fundamental *dogma* of molecular biology

**DNA**

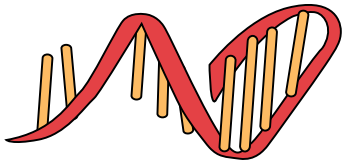
$\{A, C, G, T\}^\star$





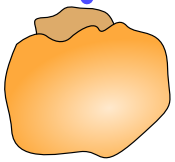

**RNAs**

$\{A, C, G, U\}^\star$

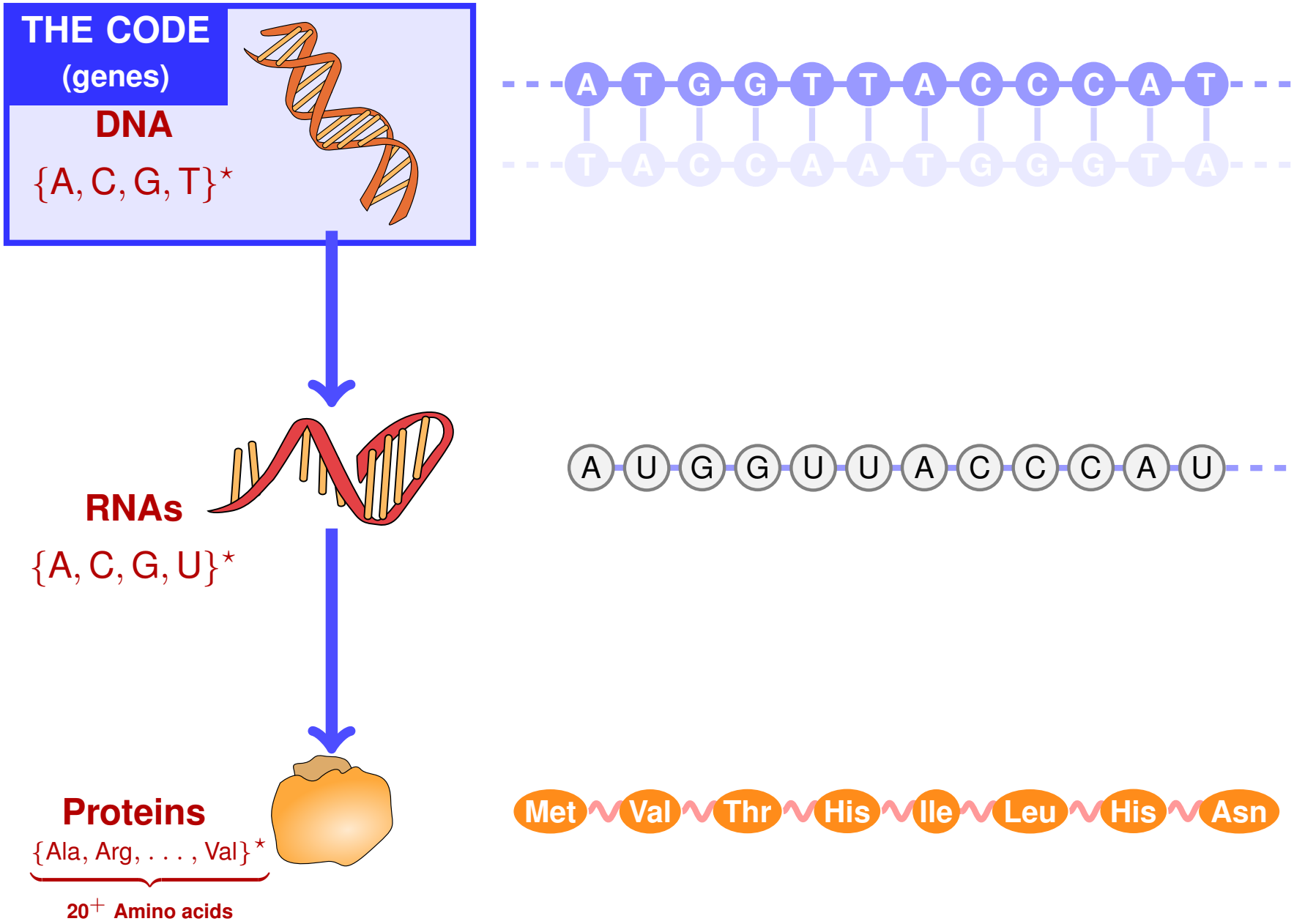



**Proteins**

$\underbrace{\{\text{Ala}, \text{Arg}, \ldots, \text{Val}\}^\star}_{20^+ \text{ Amino acids}}$



# Fundamental *dogma* of molecular biology

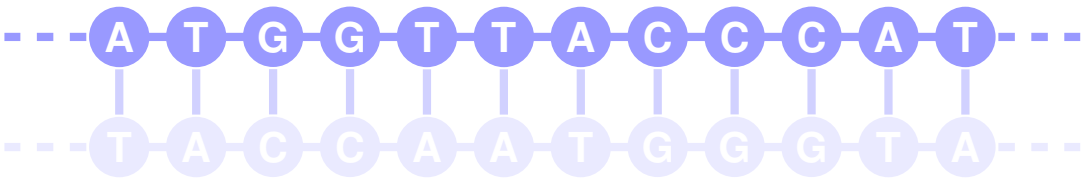


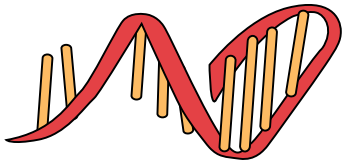
# Fundamental *dogma* of molecular biology

**THE CODE**
**(genes)**

**DNA**

$\{A, C, G, T\}^\star$

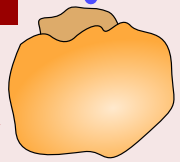


**RNAs**

$\{A, C, G, U\}^\star$


**THE MACHINE**
**(enzymes)**



**Proteins**

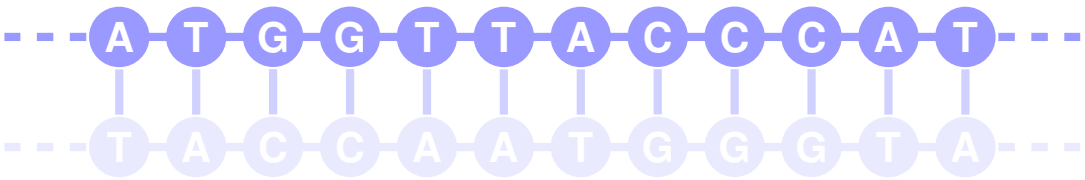
$\underbrace{\{\text{Ala}, \text{Arg}, \ldots, \text{Val}\}^\star}_{20^+ \text{ Amino acids}}$

# Fundamental *dogma* of molecular biology

**THE CODE**
**(genes)**

**DNA**

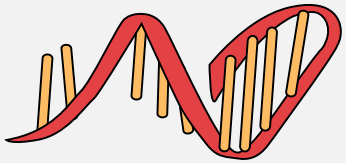
$\{A, C, G, T\}^\star$




**MEH...**

**RNAs**

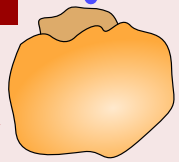
$\{A, C, G, U\}^\star$



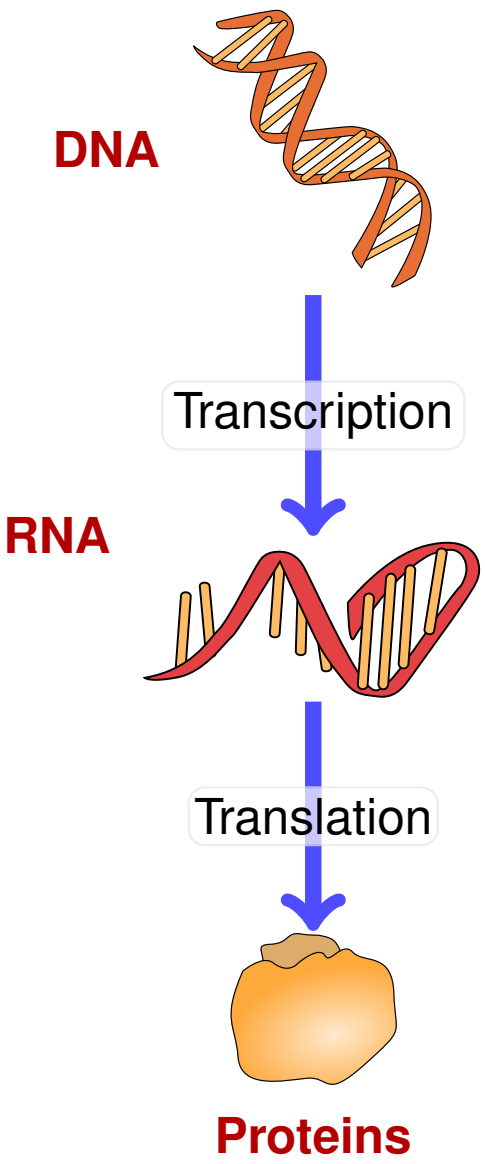

**THE MACHINE**
**(enzymes)**

**Proteins**

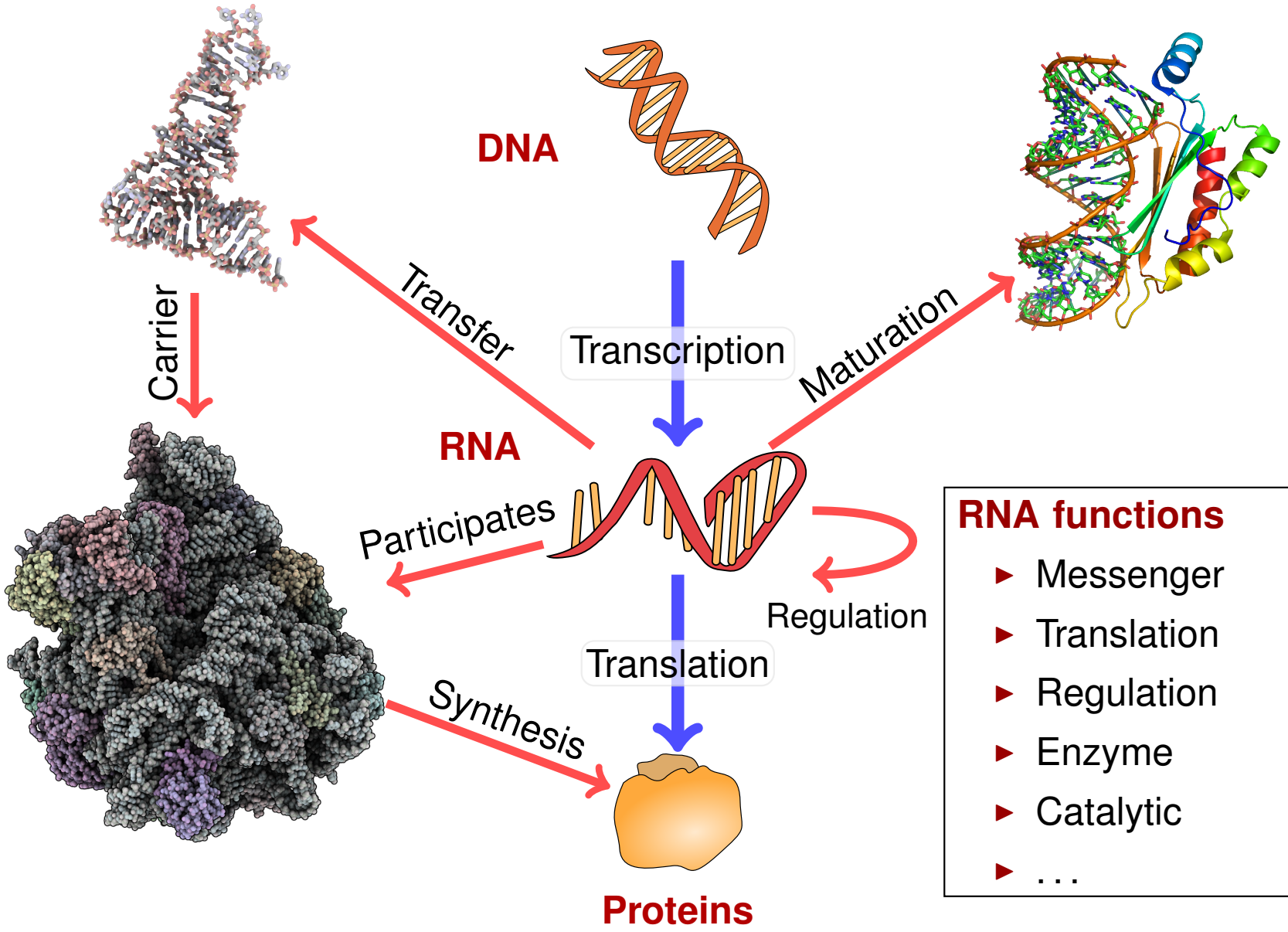
$\underbrace{\{\text{Ala}, \text{Arg}, \ldots, \text{Val}\}^\star}_{20^+ \text{ Amino acids}}$



# Fundamental *dogma* of molecular biology



# Fundamental *dogma* of molecular biology

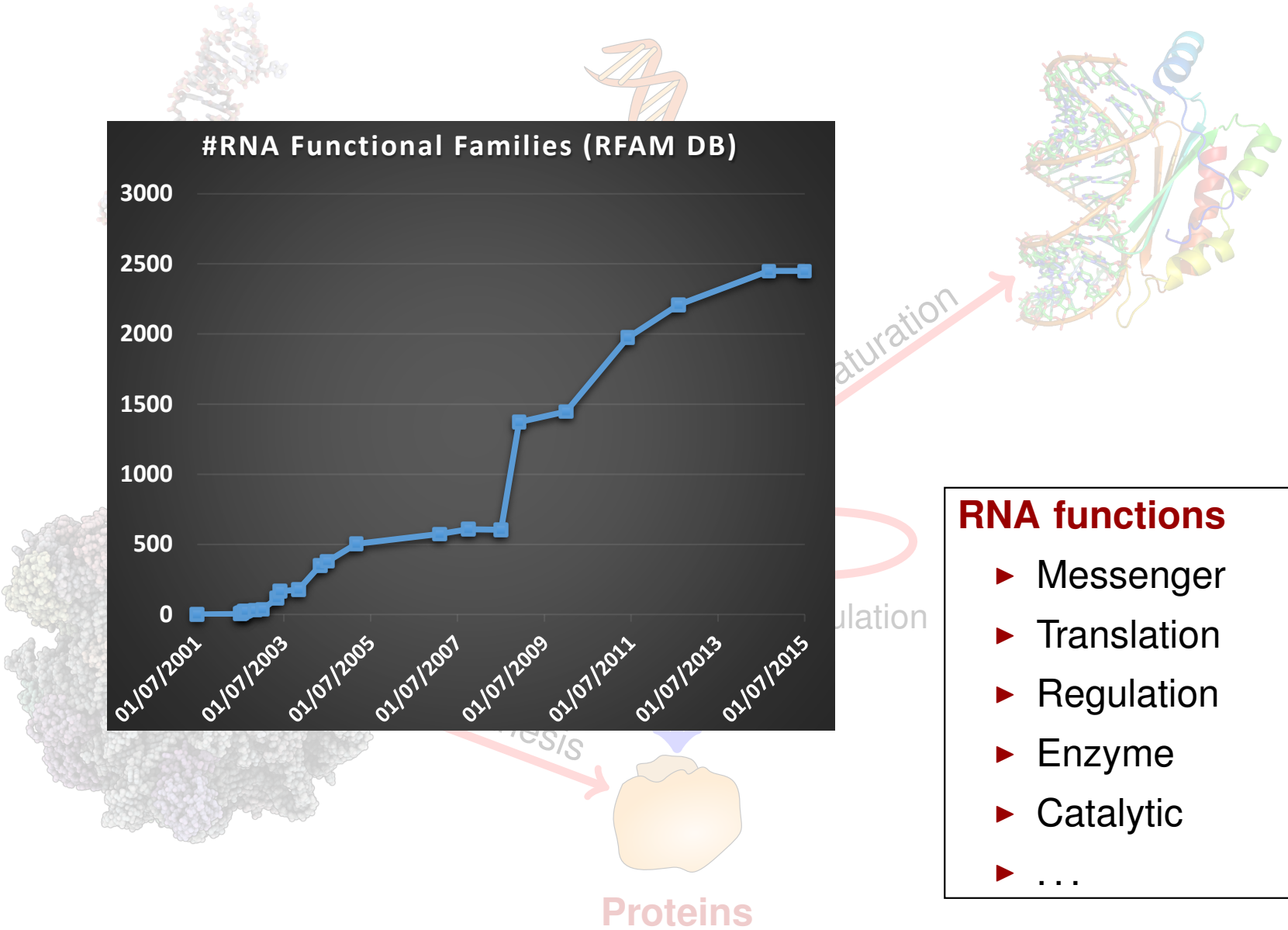



**RNA functions**

- Messenger
- Translation
- Regulation
- Enzyme
- Catalytic
- ...

# Fundamental *dogma* of molecular biology

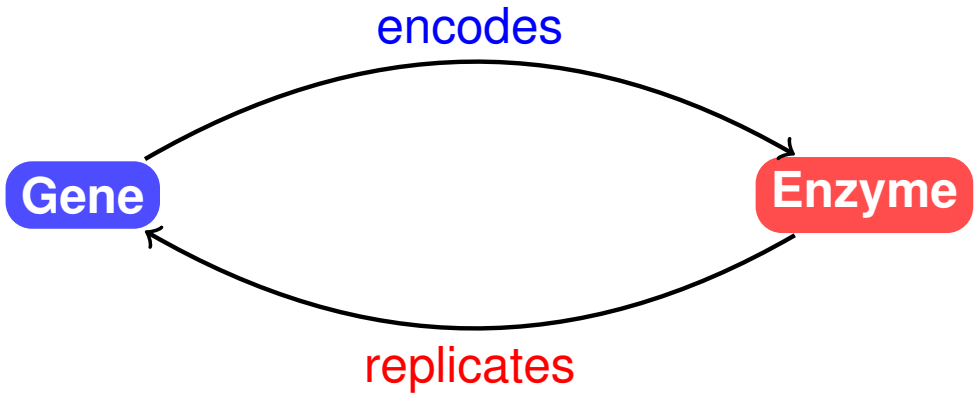



**RNA functions**

- Messenger
- Translation
- Regulation
- Enzyme
- Catalytic
- ...

# RNA world: Resolving the *chicken vs egg* paradox at the origin of life...



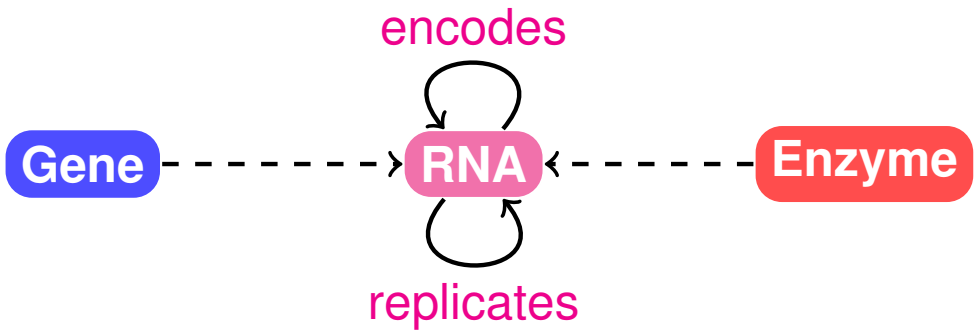


A gene big enough to specify an enzyme would be too big to replicate accurately without the aid of an enzyme of the very kind that it is trying to specify. So the system *apparently cannot get started*.

[...] This is the RNA World. To see how plausible it is, we need to look at why proteins are good at being enzymes but bad at being replicators; at why DNA is good at replicating but bad at being an enzyme; and finally why *RNA might just be good enough at both roles to break out of the Catch-22.*

**R. Dawkins**. *The Ancestor's Tale: A Pilgrimage to the Dawn of Evolution*

# RNA world: Resolving the *chicken vs egg* paradox at the origin of life...





A gene big enough to specify an enzyme would be too big to replicate accurately without the aid of an enzyme of the very kind that it is trying to specify. So the system *apparently cannot get started*.

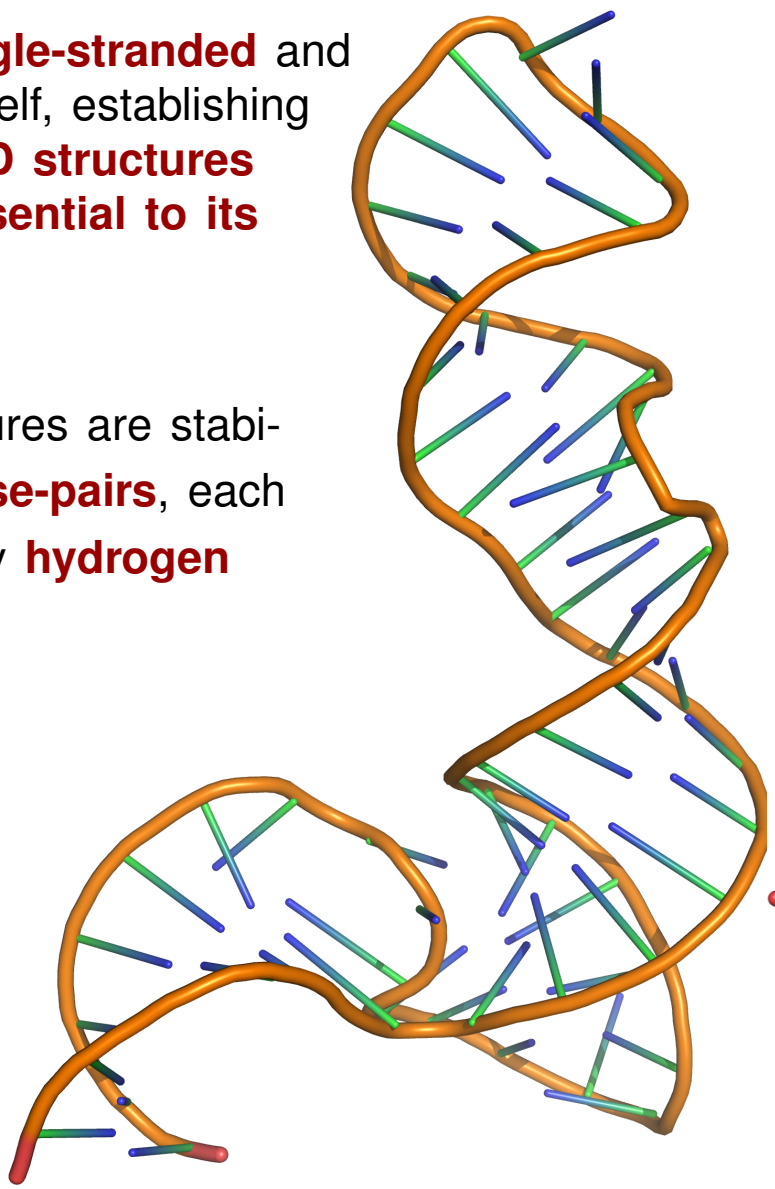
[...] This is the RNA World. To see how plausible it is, we need to look at why proteins are good at being enzymes but bad at being replicators; at why DNA is good at replicating but bad at being an enzyme; and finally why *RNA might just be good enough at both roles to break out of the Catch-22.*

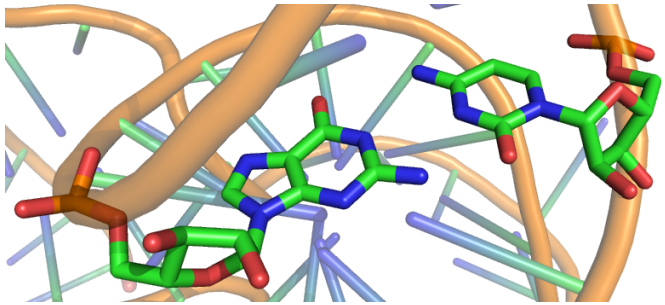
**R. Dawkins**. *The Ancestor's Tale: A Pilgrimage to the Dawn of Evolution*

# RNA folding

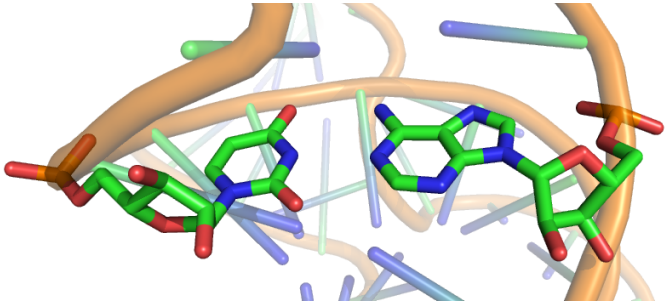
RNA is **single-stranded** and **folds** on itself, establishing **complex 3D structures** that are **essential to its function(s)**.

RNA structures are stabilized by **base-pairs**, each mediated by **hydrogen bonds**.



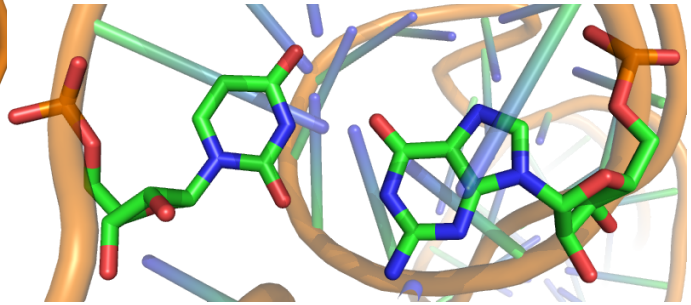


G/C



U/A

Watson/Crick base-pairs



U/G

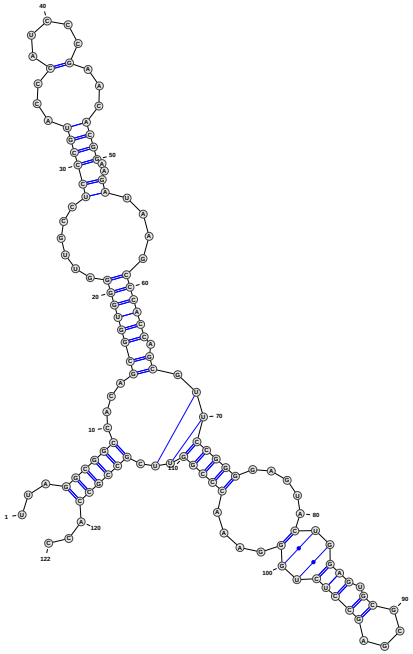
Wobble base-pair

**Canonical base-pairs**

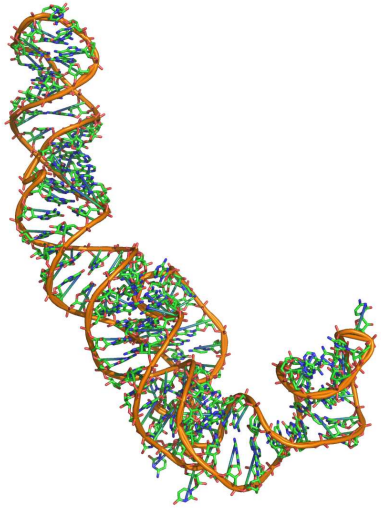
# RNA structure(s)

```
UUAGGCGGCCACAGC
GGUGGGGUUGCCUCC
CGUACCCAUCCCGAA
CACGGAAGAUAAGCC
CACCAGCGUUCCGGG
GAGUACUGGAGUGCG
CGAGCCUCUGGGAAA
CCCGGUUCGCCGCCA
CC
```

Primary structure



Secondary structure



Tertiary structure

Source: 5s rRNA (PDBID: 1K73:B)

**Definition (Secondary Structure)**

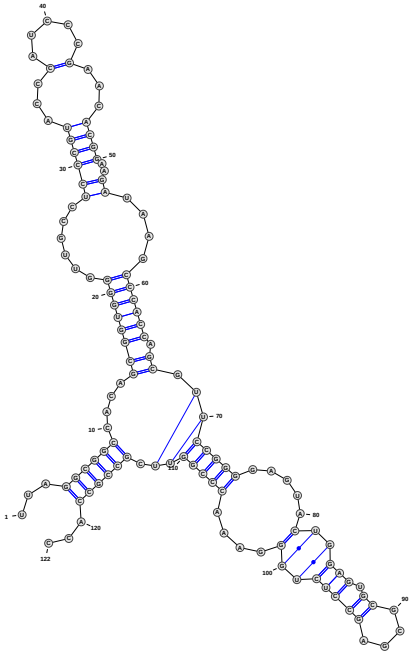
A **secondary structure** $S$ for an RNA $w$ is a set of **base-pairs** $(i, j) \in [1, n]^2$ such that:

- **Monogamy:** Each position $x \in [1, n]$ involved in **at most** one base-pair;
- **Non-crossing base-pairs:** $\nexists (i, j), (k, l) \in S$ such that $i < k < j < l$;
- **Steric constraints:** $\forall (i, j)$, one has $i < j$ and $j - i > \theta$ (where $\theta := 1$ typically).

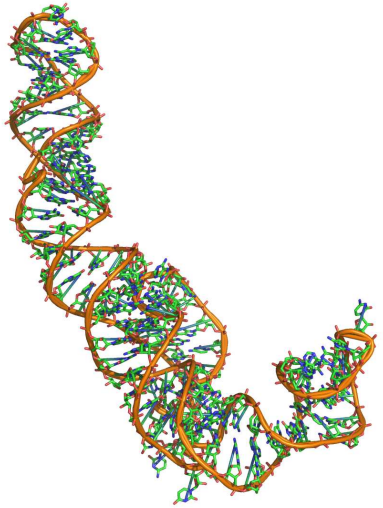
# RNA structure(s)

```
UUAGGCGGCCACAGC
GGUGGGGUUGCCUCC
CGUACCCAUCCCGAA
CACGGAAGAUAAGCC
CACCAGCGUUCCGGG
GAGUACUGGAGUGCG
CGAGCCUCUGGGAAA
CCCGGUUCGCCGCCA
CC
```

Primary structure



Secondary structure



Tertiary structure

Source: 5s rRNA (PDBID: 1K73:B)

**Definition (Secondary Structure)**

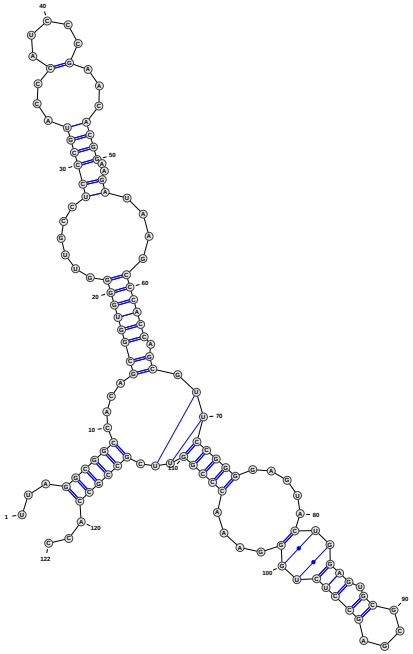
A **secondary structure** $S$ for an RNA $w$ is a set of **base-pairs** $(i, j) \in [1, n]^2$ such that:

- **Monogamy:** Each position $x \in [1, n]$ involved in **at most** one base-pair;
- **Non-crossing base-pairs:** $\nexists (i, j), (k, l) \in S$ such that $i < k < j < l$;
- **Steric constraints:** $\forall (i, j)$, one has $i < j$ and $j - i > \theta$ (where $\theta := 1$ typically).

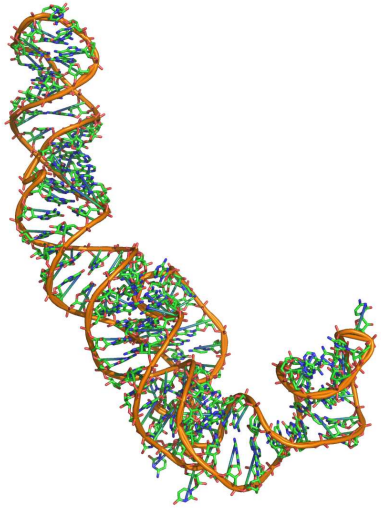
# RNA structure(s)

```
UUAGGCGGCCACAGC
GGUGGGGUUGCCUCC
CGUACCCAUCCCGAA
CACGGAAGAUAAGCC
CACCAGCGUUCCGGG
GAGUACUGGAGUGCG
CGAGCCUCUGGGAAA
CCCGGUUCGCCGCCA
CC
```

Primary structure



Secondary structure



Tertiary structure

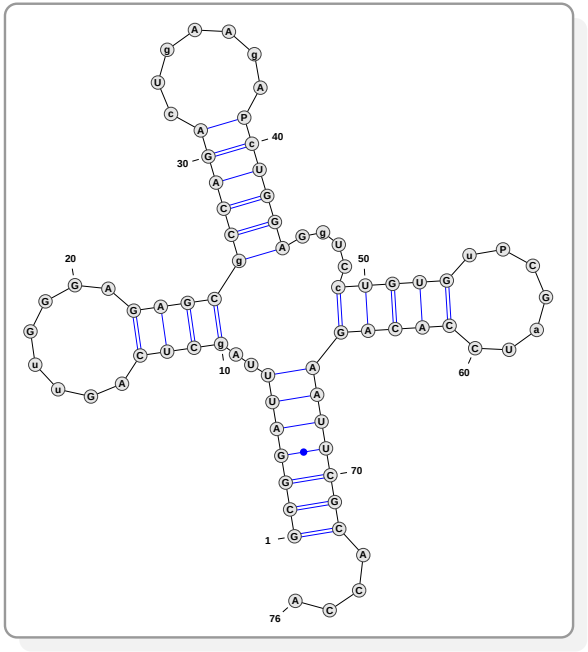
Source: 5s rRNA (PDBID: 1K73:B)

**Definition (Secondary Structure)**

A **secondary structure** $S$ for an RNA $w$ is a set of **base-pairs** $(i, j) \in [1, n]^2$ such that:

- **Monogamy:** Each position $x \in [1, n]$ involved in **at most** one base-pair;
- **Non-crossing base-pairs:** $\nexists (i, j), (k, l) \in S$ such that $i < k < j < l$;
- **Steric constraints:** $\forall (i, j)$, one has $i < j$ and $j - i > \theta$ (where $\theta := 1$ typically).

# Various representations for a versatile biomolecule



Outer-planar graphs

Hamiltonian-path, $\Delta(G) \leq 3$, 2-connected*

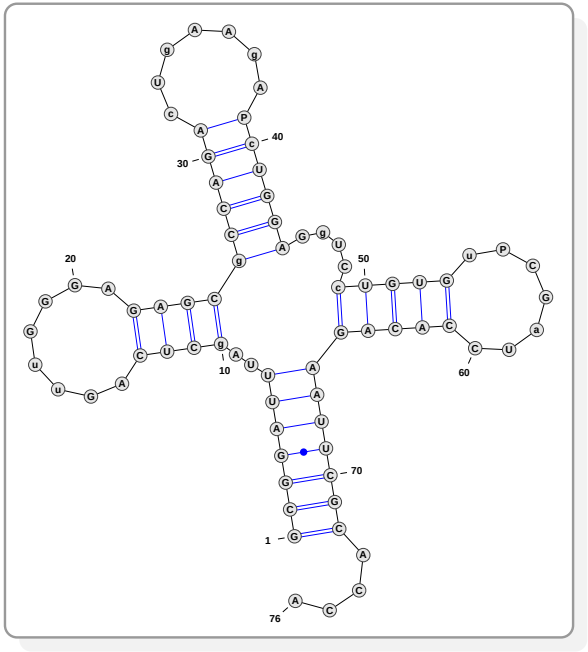
**Supporting intuitions**

Different representations

Common combinatorial structure

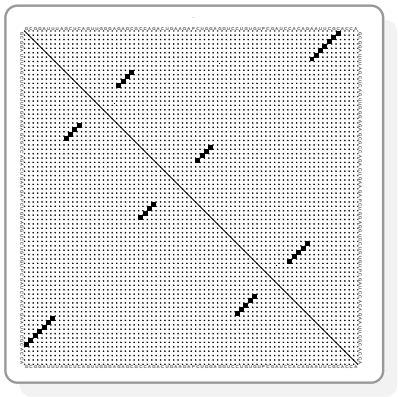
* Additional steric constraints

# Various representations for a versatile biomolecule



Outer-planar graphs

Hamiltonian-path, $\Delta(G) \leq 3$, 2-connected$^\star$



Dot plots

Adjacency matrices$^\star$

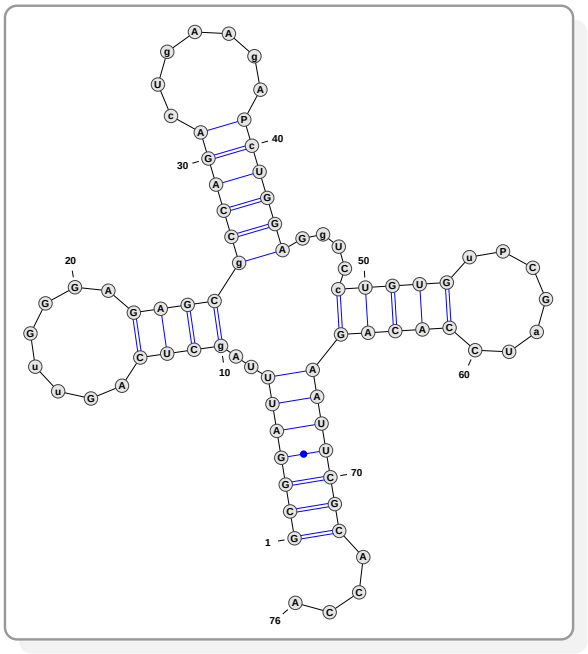
**Supporting intuitions**

Different representations

Common combinatorial structure

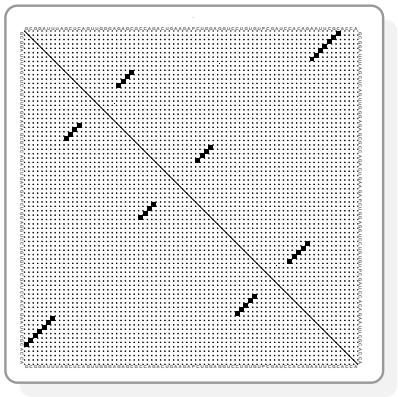
$^\star$Additional steric constraints

# Various representations for a versatile biomolecule



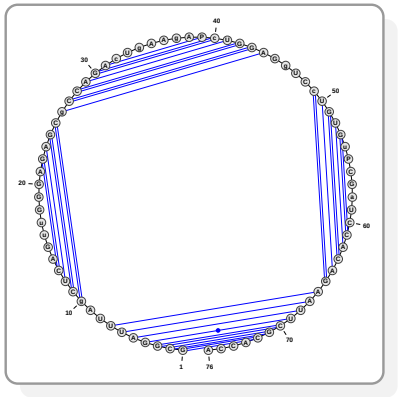
Outer-planar graphs

Hamiltonian-path, $\Delta(G) \leq 3$, 2-connected$^\star$



Dot plots

Adjacency matrices$^\star$



Non-crossing arc diagrams$^\star$

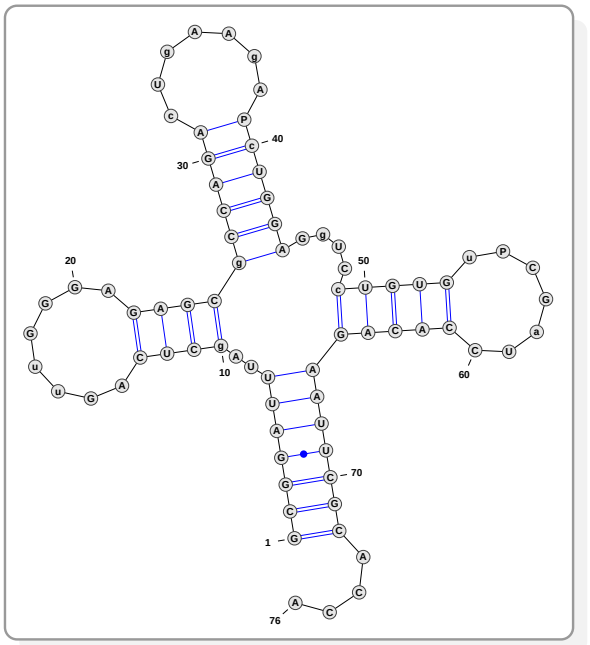
**Supporting intuitions**

Different representations

Common combinatorial structure

$^\star$Additional steric constraints

# Various representations for a versatile biomolecule

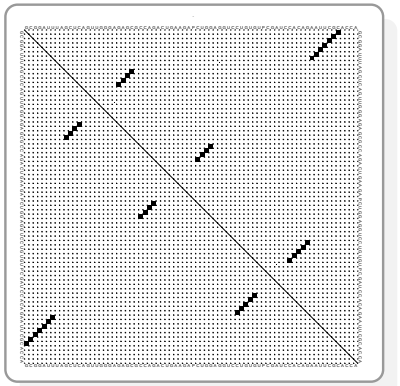


Outer-planar graphs

Hamiltonian-path, $\Delta(G) \leq 3$, 2-connected$^\star$

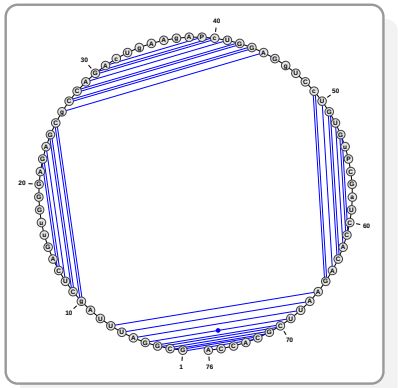
```
(((((((..((((........))))((((((.......))))))....(((((.......)))))))))))).....
```

Motzkin words$^\star$



Dot plots

Adjacency matrices$^\star$



Non-crossing arc diagrams$^\star$

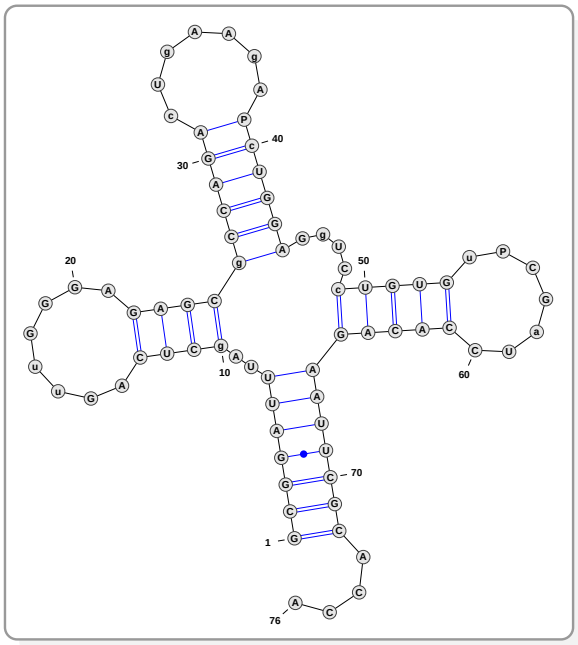
**Supporting intuitions**

Different representations

Common combinatorial structure

$^\star$Additional steric constraints

# Various representations for a versatile biomolecule

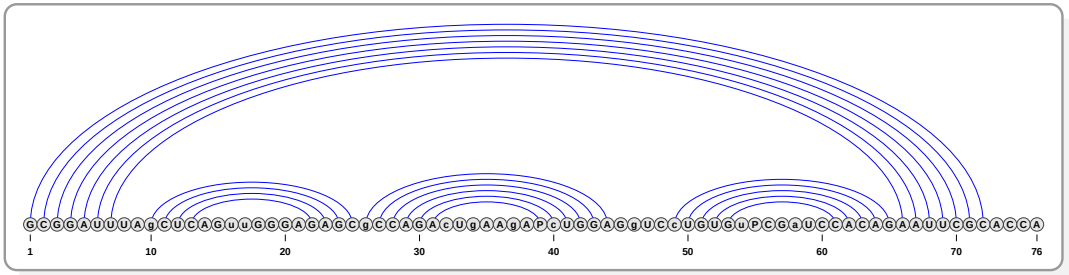


Outer-planar graphs

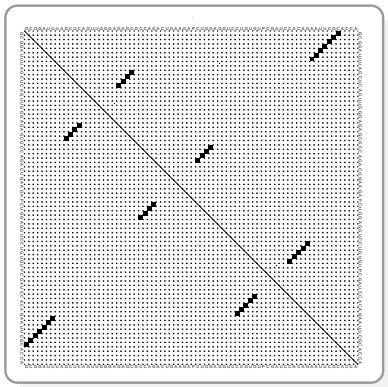
Hamiltonian-path, $\Delta(G) \leq 3$, 2-connected$^\star$

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

Motzkin words$^\star$

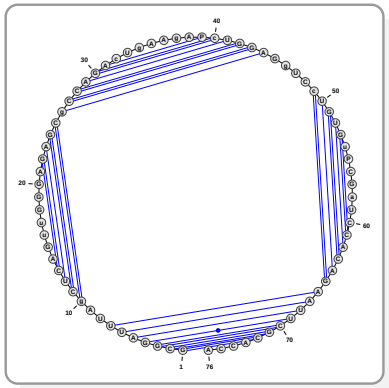


Non-crossing arc-annotated sequences$^\star$



Dot plots

Adjacency matrices$^\star$



Non-crossing arc diagrams$^\star$

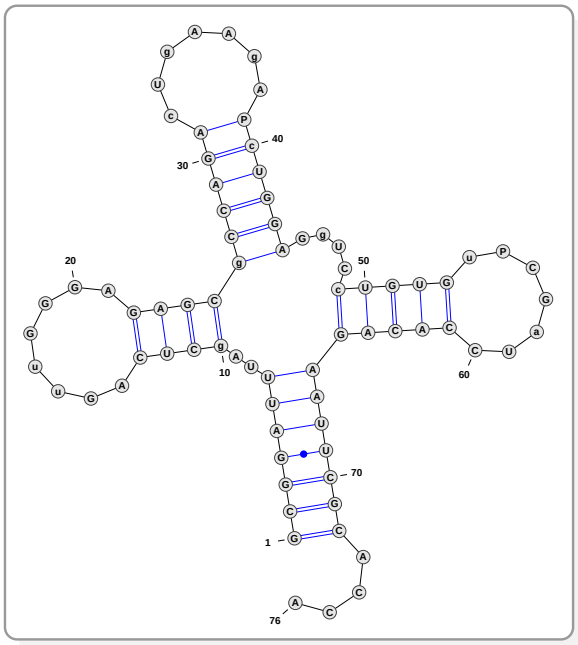
**Supporting intuitions**

Different representations

Common combinatorial structure

$^\star$Additional steric constraints

# Various representations for a versatile biomolecule

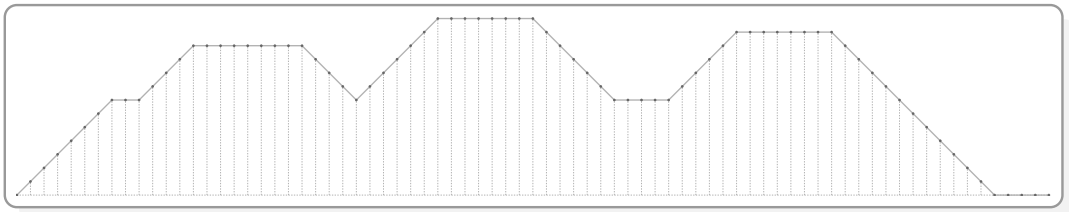


Outer-planar graphs

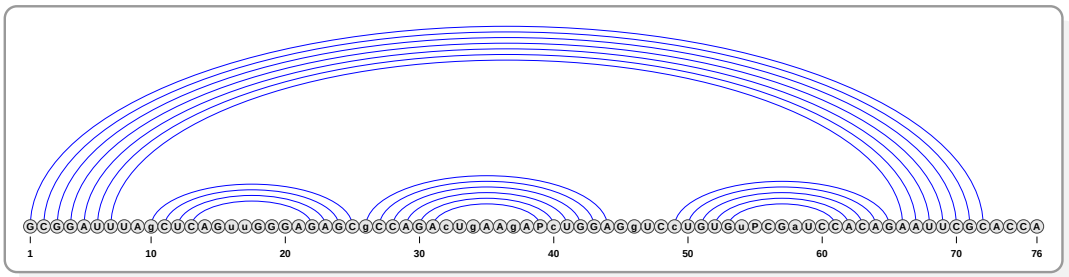
Hamiltonian-path, $\Delta(G) \leq 3$, 2-connected$^\star$

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

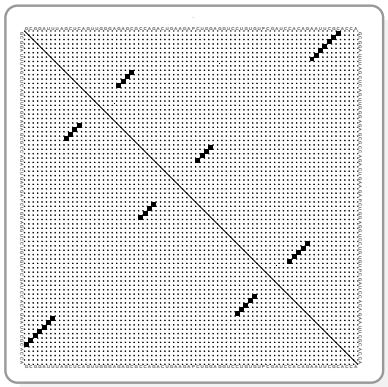
Motzkin words$^\star$



Positive 1D meanders$^\star$ over $\mathcal{S} = \{+1, -1, 0\}$

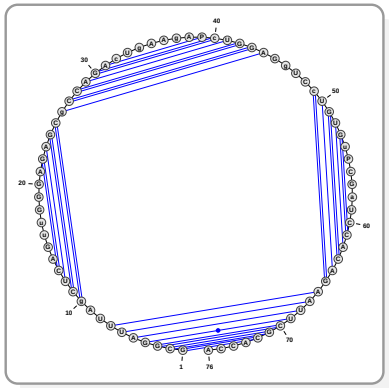


Non-crossing arc-annotated sequences$^\star$



Dot plots

Adjacency matrices$^\star$



Non-crossing arc diagrams$^\star$

**Supporting intuitions**

Different representations

Common combinatorial structure

$^\star$Additional steric constraints

# Part 1. Enumerative aspects

# Life through the lens of enumerative combinatorics

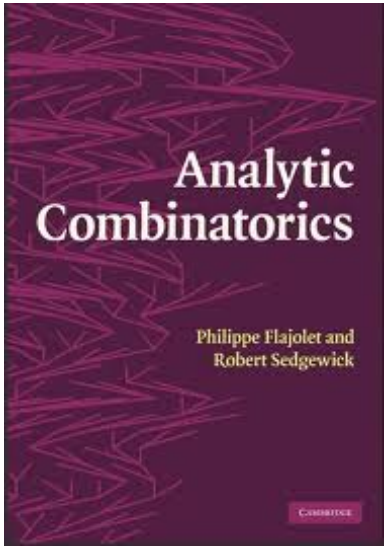
**Enumerative combinatorics** can be used to characterize the **precise (asymptotic) behavior** of **abstract models** for RNA sequence/structure.

**Typical problems**

- How many secondary structures on $n$ nucleotides? [Waterman, 1978]
- Expected #structures compatible with random RNA? [Zuker and Sankoff, 1984]
- Average distance between extremities? [Clote, Ponty, and Steyaert, 2012b]

The **symbolic method**, a **generic framework** for enumeration:

1. Find a suitable decomposition
2. Rephrase into grammar/specification
3. Translate equations & solve for generating function(s)
4. Singularity analysis yields asymptotics



[Flajolet and Sedgewick, 2009]

# Life through the lens of enumerative combinatorics

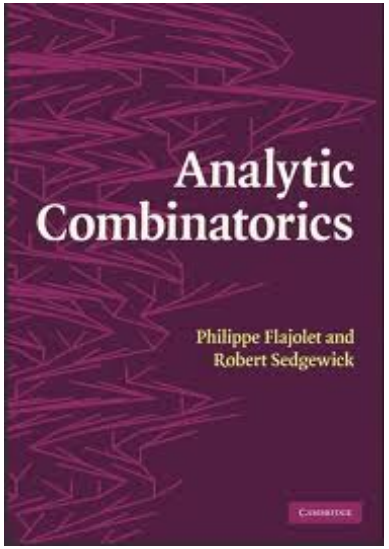
**Enumerative combinatorics** can be used to characterize the **precise (asymptotic) behavior** of **abstract models** for RNA sequence/structure.

**Typical problems**

- How many secondary structures on $n$ nucleotides? [Waterman, 1978]
- Expected #structures compatible with random RNA? [Zuker and Sankoff, 1984]
- Average distance between extremities? [Clote, Ponty, and Steyaert, 2012b]

The **symbolic method**, a **generic framework** for enumeration:

1. Find a suitable decomposition
2. Rephrase into grammar/specification
3. Translate equations & solve for generating function(s)
4. Singularity analysis yields asymptotics



[Flajolet and Sedgewick, 2009]

# Life through the lens of enumerative combinatorics

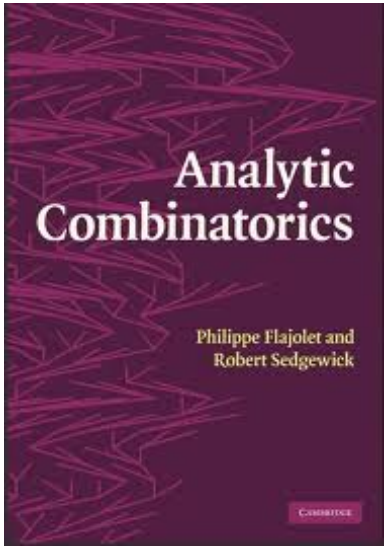
**Enumerative combinatorics** can be used to characterize the **precise (asymptotic) behavior** of **abstract models** for RNA sequence/structure.

**Typical problems**

- How many secondary structures on $n$ nucleotides? [Waterman, 1978]
- Expected #structures compatible with random RNA? [Zuker and Sankoff, 1984]
- Average distance between extremities? [Clote, Ponty, and Steyaert, 2012b]

The **symbolic method**, a **generic framework** for enumeration:

1. Find a suitable decomposition
2. Rephrase into grammar/specification
3. Translate equations & solve for generating function(s)
4. Singularity analysis yields asymptotics



[Flajolet and Sedgewick, 2009]

# Life through the lens of enumerative combinatorics

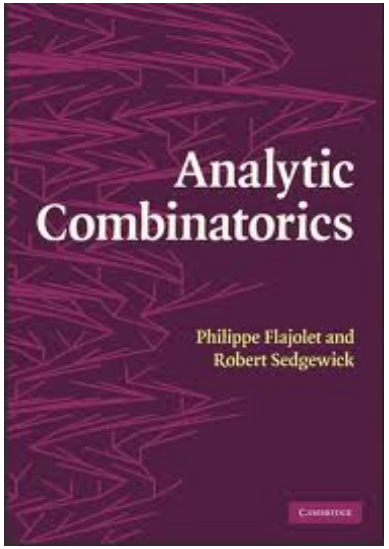
**Enumerative combinatorics** can be used to characterize the **precise (asymptotic) behavior** of **abstract models** for RNA sequence/structure.

**Typical problems**

- How many secondary structures on $n$ nucleotides? [Waterman, 1978]
- Expected #structures compatible with random RNA? [Zuker and Sankoff, 1984]
- Average distance between extremities? [Clote, Ponty, and Steyaert, 2012b]

The **symbolic method**, a **generic framework** for enumeration:

1. Find a suitable decomposition
2. Rephrase into grammar/specification
3. Translate equations & solve for generating function(s)
4. Singularity analysis yields asymptotics



[Flajolet and Sedgewick, 2009]

# Life through the lens of enumerative combinatorics

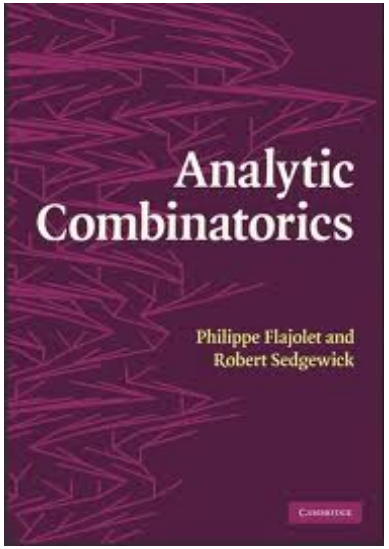
**Enumerative combinatorics** can be used to characterize the **precise (asymptotic) behavior** of **abstract models** for RNA sequence/structure.

**Typical problems**

- How many secondary structures on $n$ nucleotides? [Waterman, 1978]
- Expected #structures compatible with random RNA? [Zuker and Sankoff, 1984]
- Average distance between extremities? [Clote, Ponty, and Steyaert, 2012b]

The **symbolic method**, a **generic framework** for enumeration:

1. Find a suitable decomposition
2. Rephrase into grammar/specification
3. Translate equations & solve for generating function(s)
4. Singularity analysis yields asymptotics

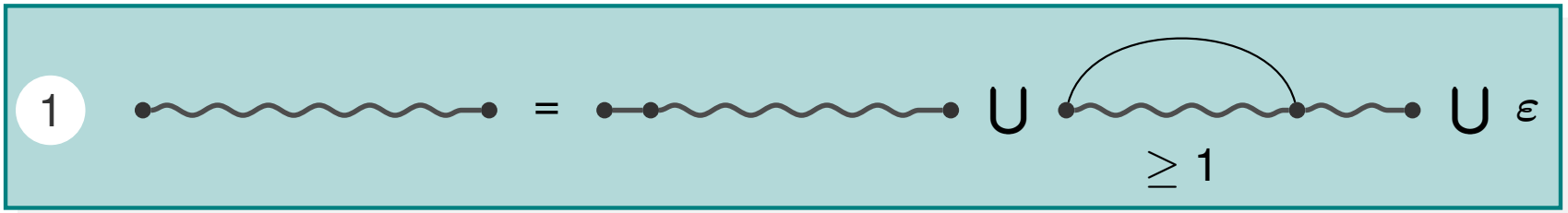


[Flajolet and Sedgewick, 2009]

# RNA secondary structures

**Goal: Generating function $S(z) = \sum_{n \geq 0} s_n z^n$**

where $s_n$ = #Secondary structures of length $n$



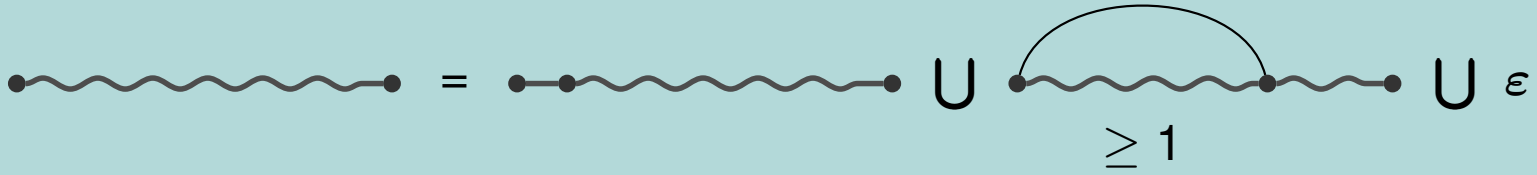

[Waterman, 1978] & [Vauchaussade de Chaumont and Viennot, 1985]

# RNA secondary structures

**Goal: Generating function $S(z) = \sum_{n \geq 0} s_n z^n$**

where $s_n$ = #Secondary structures of length $n$

1




2 $S \rightarrow \bullet\, S \mid (\, S_{>0} \,)\, S \mid \varepsilon$

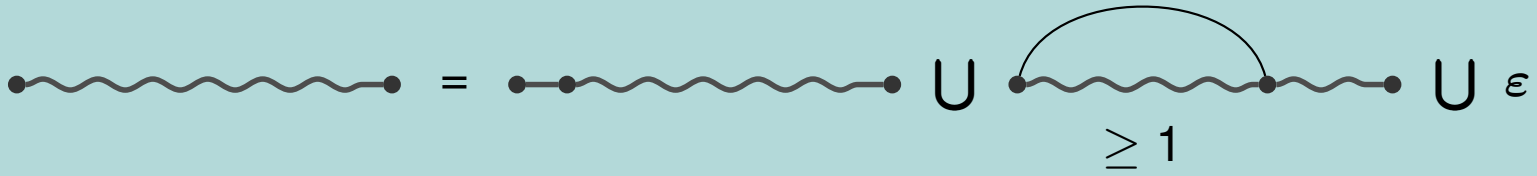
[Waterman, 1978] & [Vauchaussade de Chaumont and Viennot, 1985]

# RNA secondary structures

**Goal: Generating function $S(z) = \sum_{n \geq 0} s_n z^n$**

where $s_n$ = #Secondary structures of length $n$

1




2

$$S \rightarrow \bullet S \mid (\mathbf{T}) S \mid \varepsilon$$

$$\mathbf{T} \rightarrow \bullet S \mid (\mathbf{T}) S$$

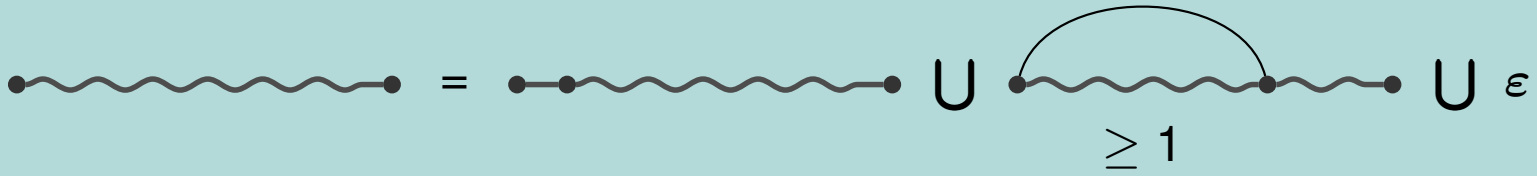
[Waterman, 1978] & [Vauchaussade de Chaumont and Viennot, 1985]

# RNA secondary structures

**Goal: Generating function $S(z) = \sum_{n \geq 0} s_n z^n$**

where $s_n$ = #Secondary structures of length $n$

1




2

$$S \rightarrow \bullet\, S \mid (\,\mathbf{T}\,)\, S \mid \varepsilon$$

$$\mathbf{T} \rightarrow \bullet\, S \mid (\,\mathbf{T}\,)\, S$$

3

$$S(z) = \frac{1 - z + z^2 - \sqrt{1 - 2z - z^2 - 2z^3 + z^4}}{2z^2}$$

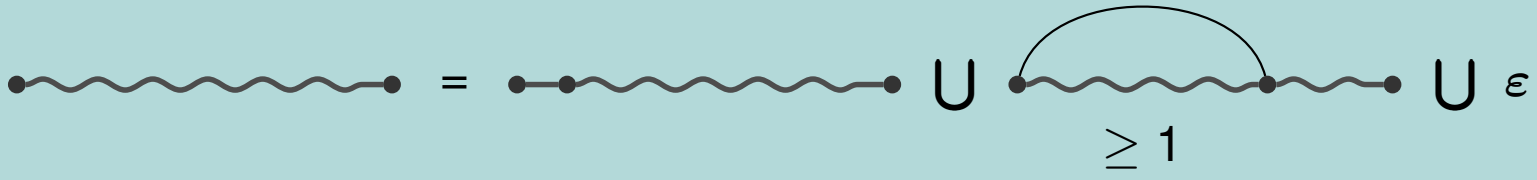
[Waterman, 1978] & [Vauchaussade de Chaumont and Viennot, 1985]

# RNA secondary structures

**Goal: Generating function $S(z) = \sum_{n \geq 0} s_n z^n$**

where $s_n$ = #Secondary structures of length $n$

1




2

$$S \rightarrow \bullet\, S \mid (\, \mathbf{T}\, )\, S \mid \varepsilon$$

$$\mathbf{T} \rightarrow \bullet\, S \mid (\, \mathbf{T}\, )\, S$$

3

$$S(z) = \frac{1 - z + z^2 - \sqrt{1 - 2z - z^2 - 2z^3 + z^4}}{2z^2}$$

4

$$\rho = \frac{3 - \sqrt{5}}{2} = 1 - \phi$$

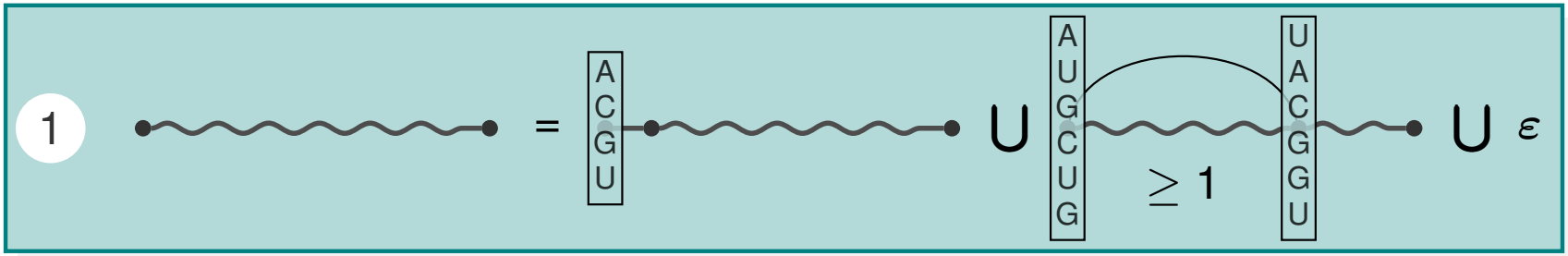
$$s_n = \sqrt{\frac{15 + 7\sqrt{5}}{8\pi}} \cdot \frac{\left(\frac{3+\sqrt{5}}{2}\right)^n}{n\sqrt{n}} (1 + \mathcal{O}(1/n)) \sim 1.1 \cdot \frac{2.6^n}{n\sqrt{n}}$$

[Waterman, 1978] & [Vauchaussade de Chaumont and Viennot, 1985]

# Expected #secondary structures compatible with RNA

**Goal: Generating function $S(z) = \sum_{n \geq 0} s_n z^n$**

where $s_n$ = #Compatible (Sequence/Sec. struct.) pairs of length $n$



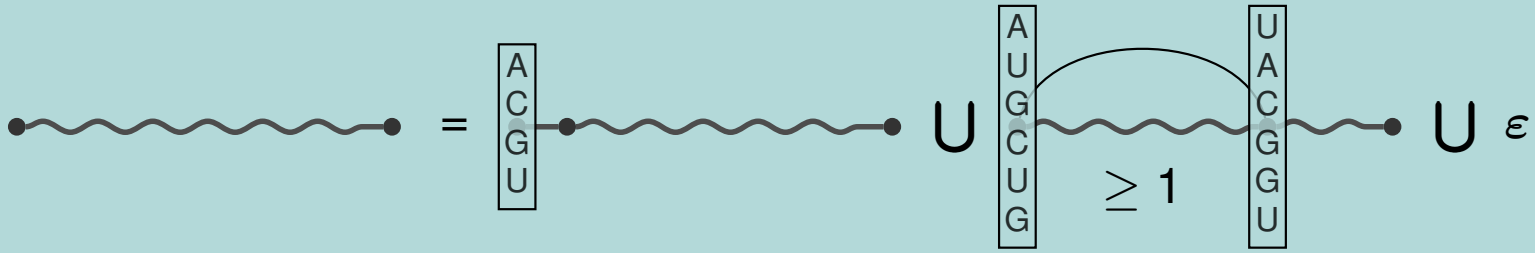

[Zuker and Sankoff, 1984]

# Expected #secondary structures compatible with RNA

**Goal: Generating function $S(z) = \sum_{n \geq 0} s_n z^n$**

where $s_n$ = #Compatible (Sequence/Sec. struct.) pairs of length $n$

1




2

$$S \to \begin{array}{l} ({}_{\mathrm{A}}\, T\,)_{\mathrm{U}}\, S \\ ({}_{\mathrm{U}}\, T\,)_{\mathrm{A}}\, S \\ ({}_{\mathrm{G}}\, T\,)_{\mathrm{C}}\, S \\ ({}_{\mathrm{C}}\, T\,)_{\mathrm{G}}\, S \\ ({}_{\mathrm{U}}\, T\,)_{\mathrm{G}}\, S \\ ({}_{\mathrm{G}}\, T\,)_{\mathrm{U}}\, S \end{array} \Bigg| \begin{array}{l} \bullet_{\mathrm{A}}\, S \\ \bullet_{\mathrm{U}}\, S \\ \bullet_{\mathrm{G}}\, S \\ \bullet_{\mathrm{C}}\, S \end{array} \Bigg| \, \varepsilon \qquad T \to \begin{array}{l} ({}_{\mathrm{A}}\, T\,)_{\mathrm{U}}\, S \\ ({}_{\mathrm{U}}\, T\,)_{\mathrm{A}}\, S \\ ({}_{\mathrm{G}}\, T\,)_{\mathrm{C}}\, S \\ ({}_{\mathrm{C}}\, T\,)_{\mathrm{G}}\, S \\ ({}_{\mathrm{U}}\, T\,)_{\mathrm{G}}\, S \\ ({}_{\mathrm{G}}\, T\,)_{\mathrm{U}}\, S \end{array} \Bigg| \begin{array}{l} \bullet_{\mathrm{A}}\, S \\ \bullet_{\mathrm{U}}\, S \\ \bullet_{\mathrm{G}}\, S \\ \bullet_{\mathrm{C}}\, S \end{array} .$$

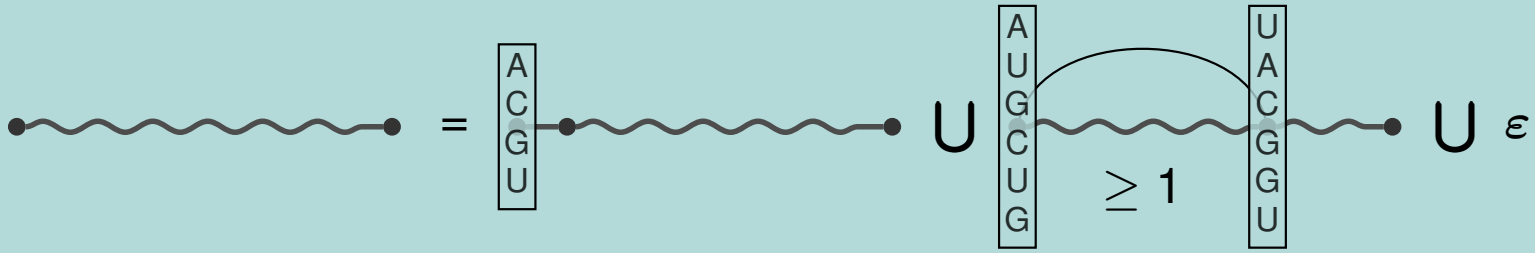
[Zuker and Sankoff, 1984]

# Expected #secondary structures compatible with RNA

**Goal: Generating function $S(z) = \sum_{n \geq 0} s_n z^n$**

where $s_n$ = #Compatible (Sequence/Sec. struct.) pairs of length $n$

1




2

$$S \to {}_{A}T_{U}\,S \;\; {}_{U}T_{A}\,S \;\; {}_{G}T_{C}\,S \;\; {}_{C}T_{G}\,S \;\; {}_{U}T_{G}\,S \;\; {}_{G}T_{U}\,S \;\Big|\; \bullet_{A}S \;\; \bullet_{U}S \;\; \bullet_{G}S \;\; \bullet_{C}S \;\Big|\; \varepsilon$$

$$T \to {}_{A}T_{U}\,S \;\; {}_{U}T_{A}\,S \;\; {}_{G}T_{C}\,S \;\; {}_{C}T_{G}\,S \;\; {}_{U}T_{G}\,S \;\; {}_{G}T_{U}\,S \;\Big|\; \bullet_{A}S \;\; \bullet_{U}S \;\; \bullet_{G}S \;\; \bullet_{C}S.$$

3

$$S(z) = \frac{1-4z+6z^2-\sqrt{1-8z-4z^2-48z^3+36z^4}}{12z^2}$$

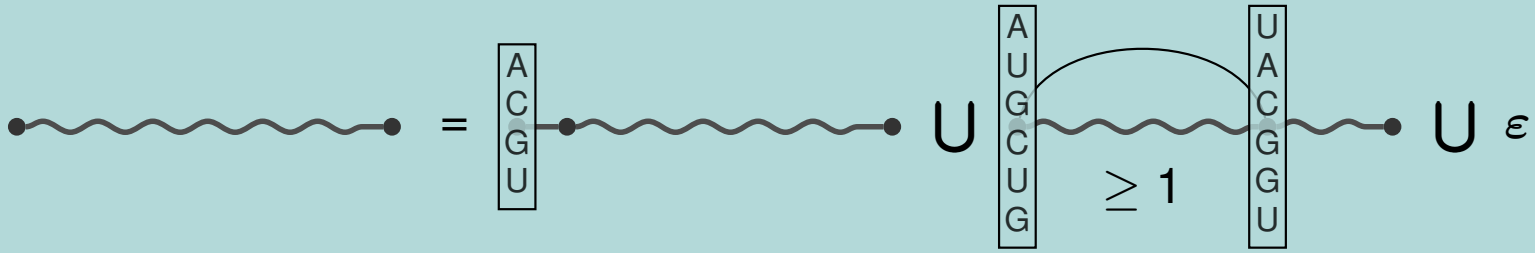
[Zuker and Sankoff, 1984]

# Expected #secondary structures compatible with RNA

**Goal: Generating function $S(z) = \sum_{n \geq 0} s_n z^n$**

where $s_n$ = #Compatible (Sequence/Sec. struct.) pairs of length $n$

1




2

$$S \to {}_{(A}\,T\,{}_{)U}\,S \;\; {}_{(U}\,T\,{}_{)A}\,S \;\; {}_{(G}\,T\,{}_{)C}\,S \;\; {}_{(C}\,T\,{}_{)G}\,S \;\; {}_{(U}\,T\,{}_{)G}\,S \;\; {}_{(G}\,T\,{}_{)U}\,S \;\Big|\; \bullet_A S \;\; \bullet_U S \;\; \bullet_G S \;\; \bullet_C S \;\Big|\; \varepsilon$$

$$T \to {}_{(A}\,T\,{}_{)U}\,S \;\; {}_{(U}\,T\,{}_{)A}\,S \;\; {}_{(G}\,T\,{}_{)C}\,S \;\; {}_{(C}\,T\,{}_{)G}\,S \;\; {}_{(U}\,T\,{}_{)G}\,S \;\; {}_{(G}\,T\,{}_{)U}\,S \;\Big|\; \bullet_A S \;\; \bullet_U S \;\; \bullet_G S \;\; \bullet_C S.$$

3

$$S(z) = \frac{1-4z+6z^2-\sqrt{1-8z-4z^2-48z^3+36z^4}}{12z^2}$$

4

$$\rho = \text{InfSing}(1-8z-4z^2-48z^3+36z^4) \quad 1/\rho \approx 8.164$$

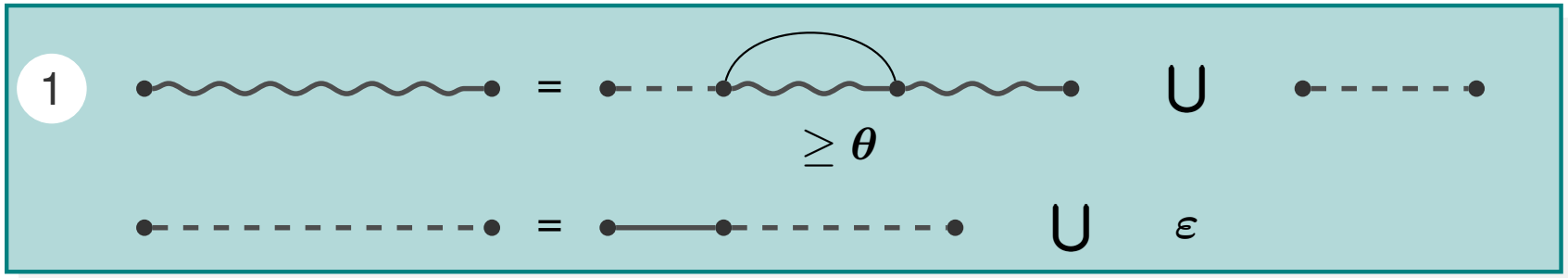
$$s_n \in \Theta\left(\frac{\rho^{-n}}{n\sqrt{n}}\right) \quad \to \quad \text{Expected\#Sec.Str.}=s_n/4^n \in \Theta(2.04^n/n\sqrt{n})$$

[Zuker and Sankoff, 1984]

# RNA secondary structures ($\theta$ constraint)

**Goal: Generating function $S_\theta(z) = \sum_{n\geq 0} s_{\theta,n}\, z^n$**

where $s_{\theta,n}$ = #Secondary structures of length $n$
having minimal base-pair distance=$\theta$

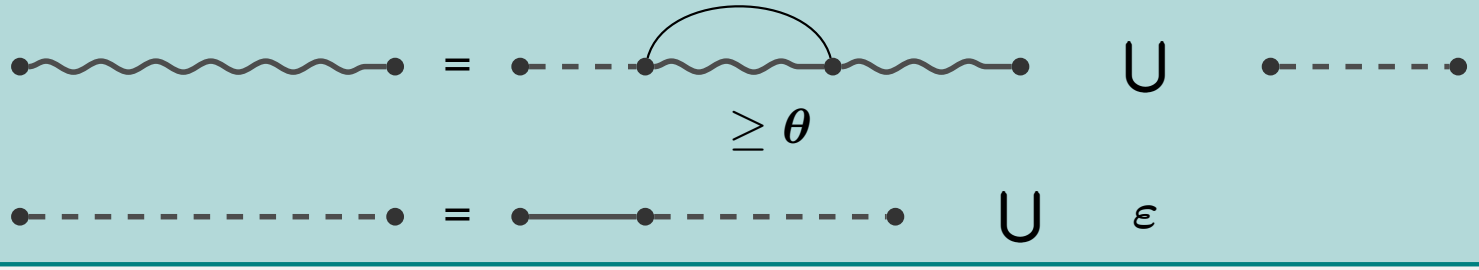


# RNA secondary structures ($\theta$ constraint)

**Goal: Generating function $S_\theta(z) = \sum_{n\geq 0} s_{\theta,n}\, z^n$**

where $s_{\theta,n}$ = #Secondary structures of length $n$ having minimal base-pair distance=$\theta$

1



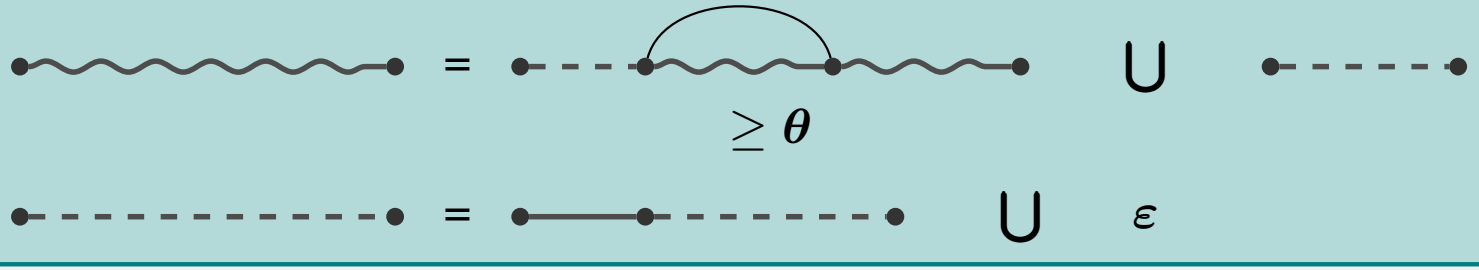

2 $\quad S \rightarrow U\,(\,S_{\geq\theta}\,)\,S \mid U \qquad U \rightarrow \bullet\, U \mid \varepsilon$

# RNA secondary structures ($\theta$ constraint)

**Goal: Generating function $S_\theta(z) = \sum_{n\geq 0} s_{\theta,n}\, z^n$**

where $s_{\theta,n}$ = #Secondary structures of length $n$ having minimal base-pair distance=$\theta$

1




2

$$S \rightarrow U\,(\,\mathbf{T}\,)\,S \mid U \qquad U \rightarrow \bullet\, U \mid \varepsilon$$

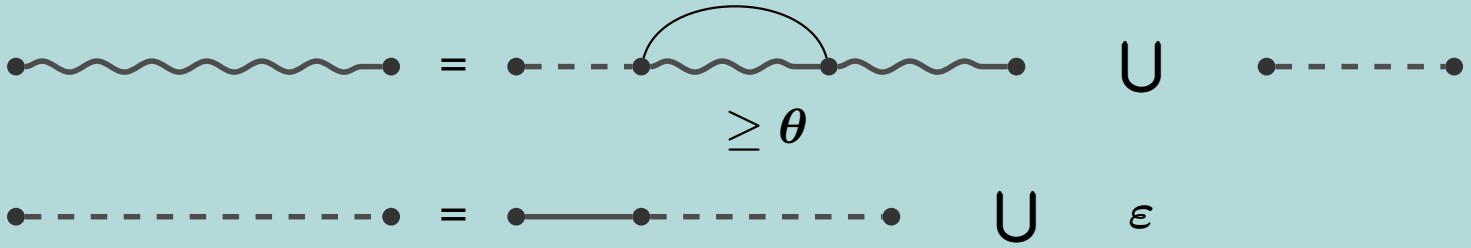
$$\mathbf{T} \rightarrow U\,(\,\mathbf{T}\,)\,S \mid \bullet^{\theta}\, U$$

# RNA secondary structures ($\theta$ constraint)

**Goal: Generating function $S_\theta(z) = \sum_{n\geq 0} s_{\theta,n}\, z^n$**

where $s_{\theta,n}$ = #Secondary structures of length $n$ having minimal base-pair distance=$\theta$

1




2

$$S \rightarrow U\,(\,\mathbf{T}\,)\,S \mid U \qquad U \rightarrow \bullet\, U \mid \varepsilon$$

$$\mathbf{T} \rightarrow U\,(\,\mathbf{T}\,)\,S \mid \bullet^{\theta}\, U$$

3

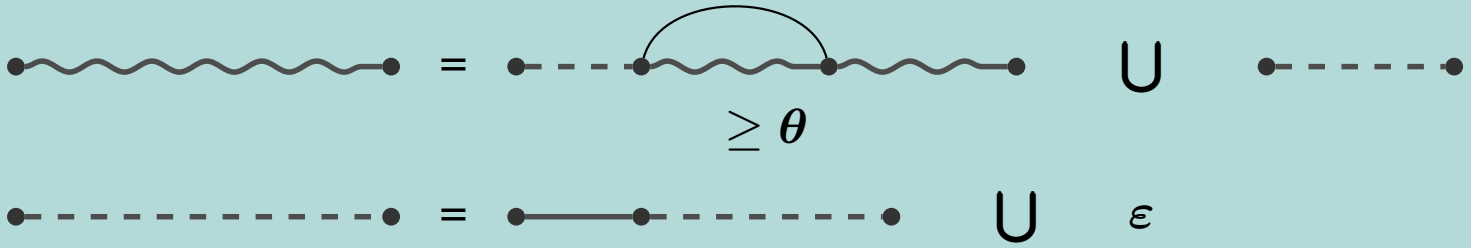
$$S_\theta(z) = \frac{1-2z+2z^2-z^{\theta+2}-\sqrt{1-4z+4z^2-2z^{\theta+2}+4z^{\theta+3}-4z^{\theta+4}+z^{2\theta+4}}}{(1-z)2z^2}$$

# RNA secondary structures ($\theta$ constraint)

**Goal: Generating function $S_\theta(z) = \sum_{n\geq 0} s_{\theta,n}\, z^n$**

where $s_{\theta,n}$ = #Secondary structures of length $n$ having minimal base-pair distance=$\theta$

1




2

$$S \rightarrow U\,(\,\mathbf{T}\,)\,S \mid U \qquad U \rightarrow \bullet\, U \mid \varepsilon$$

$$\mathbf{T} \rightarrow U\,(\,\mathbf{T}\,)\,S \mid \bullet^{\theta}\, U$$

3

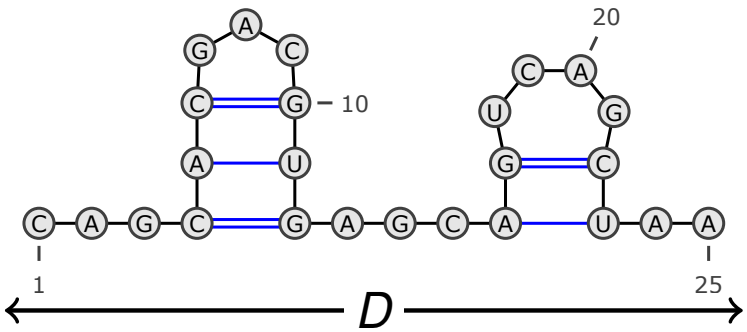
$$S_\theta(z) = \frac{1-2z+2z^2-z^{\theta+2}-\sqrt{1-4z+4z^2-2z^{\theta+2}+4z^{\theta+3}-4z^{\theta+4}+z^{2\theta+4}}}{(1-z)2z^2}$$

4

$$s_n \sim K.\frac{\beta^n}{n\sqrt{n}}(1+\mathcal{O}(1/n))$$

| $\theta$ | 0 | 1 | 3 | 10 |
|---|---|---|---|---|
| $\beta$ | 3. | 2.62 | 2.29 | 2.02 |

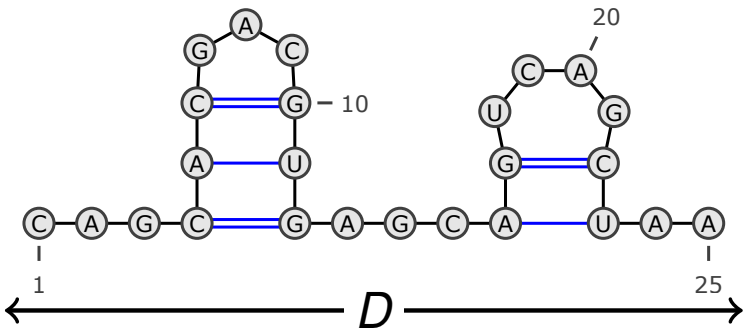
# 5'–3' Distance: Expectation through derivation




**Goal:** $S(z,u) = \sum_{n\geq 0} \sum_{d\geq 0} s_{\theta,n,d}\, z^n\, u^d$

where $s_{\theta,n,d}$ = #Sec. str. of length $n$,
having BP min. dist $= \theta$
and 5'-3' distance $= d$

# 5'–3' Distance: Expectation through derivation



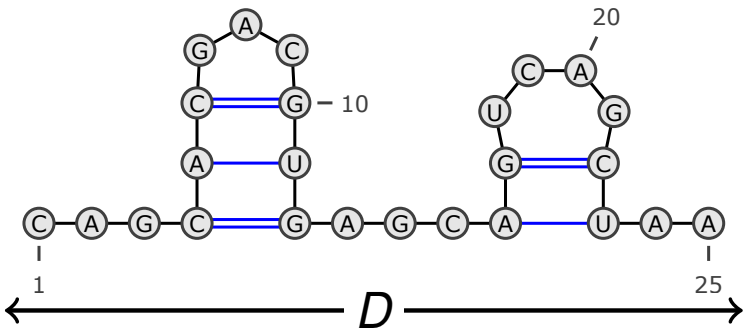

**Goal:** $S(z,u) = \sum_{n\geq 0} \sum_{d\geq 0} s_{\theta,n,d}\, z^n\, u^d$

where $s_{\theta,n,d}$ = #Sec. str. of length $n$,
having BP min. dist $= \theta$
and 5'-3' distance $= d$

2

$$T \to [S_{\geq\theta}]T| \bullet T|\varepsilon \qquad S \to (S_{\geq\theta})S| \circ S|\varepsilon$$
$$S_{\geq\theta} \to (S_{\geq\theta})S| \circ S_{\geq\theta}|\circ^{\theta}$$

# 5'–3' Distance: Expectation through derivation




**Goal:** $S(z,u) = \sum_{n\geq 0}\sum_{d\geq 0} s_{\theta,n,d}\, z^n\, u^d$

where $s_{\theta,n,d}$ = #Sec. str. of length $n$, having BP min. dist $= \theta$ and 5'-3' distance $= d$

2

$$T \to [S_{\geq\theta}]T| \bullet T|\varepsilon \qquad S \to (S_{\geq\theta})S| \circ S|\varepsilon$$
$$S_{\geq\theta} \to (S_{\geq\theta})S| \circ S_{\geq\theta}|\circ^{\theta}$$

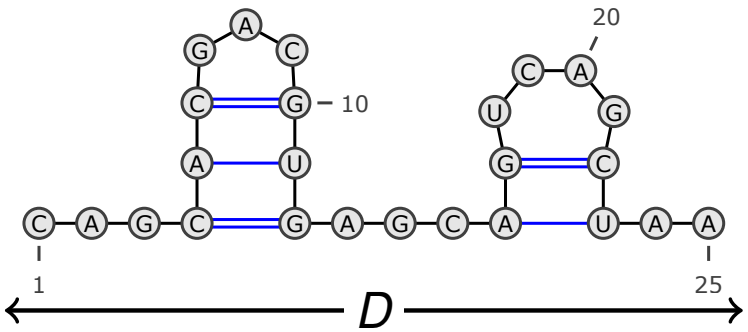
3

$$E_\theta(z) = \left.\frac{\partial T(z,u)}{\partial u}\right|_{u=1} = \frac{\begin{pmatrix} 2-9z+14z^2-8z^3+2z^5 \\ +z^{\theta+2}(-4+10z-10z^2+2z^3)+z^{2\theta+4}(2-z) \\ -(2-5z+4z^2-2z^{\theta+2}+z^{\theta+3})\sqrt{\Delta_\theta} \end{pmatrix}}{2(1-z)^2 z^4}$$

$$\Delta_\theta := 1-4z+4z^2-2z^{\theta+2}+4z^{\theta+3}-4z^{\theta+4}+z^{2\theta+4}$$

[Clote, Ponty, and Steyaert, 2012b]

# 5'–3' Distance: Expectation through derivation




**Goal:** $S(z,u) = \sum_{n\geq 0}\sum_{d\geq 0} s_{\theta,n,d}\, z^n\, u^d$

where $s_{\theta,n,d}$ = #Sec. str. of length $n$, having BP min. dist $= \theta$ and 5'-3' distance $= d$

2

$$T \to [S_{\geq\theta}]T| \bullet T|\varepsilon \qquad S \to (S_{\geq\theta})S| \circ S|\varepsilon$$
$$S_{\geq\theta} \to (S_{\geq\theta})S| \circ S_{\geq\theta}|\circ^{\theta}$$

3

$$E_\theta(z) = \left.\frac{\partial T(z,u)}{\partial u}\right|_{u=1} = \frac{\begin{pmatrix} 2-9z+14z^2-8z^3+2z^5 \\ +z^{\theta+2}(-4+10z-10z^2+2z^3)+z^{2\theta+4}(2-z) \\ -(2-5z+4z^2-2z^{\theta+2}+z^{\theta+3})\sqrt{\Delta_\theta} \end{pmatrix}}{2(1-z)^2 z^4}$$

$$\Delta_\theta := 1-4z+4z^2-2z^{\theta+2}+4z^{\theta+3}-4z^{\theta+4}+z^{2\theta+4}$$

4

$$D_n \sim \frac{2-5\rho+4\rho^2-2\rho^{\theta+2}+\rho^{\theta+3}}{(1-\rho)\rho^2} - 1,\ \rho \text{ smallest root of } \Delta_\theta = 0$$

[Clote, Ponty, and Steyaert, 2012b]

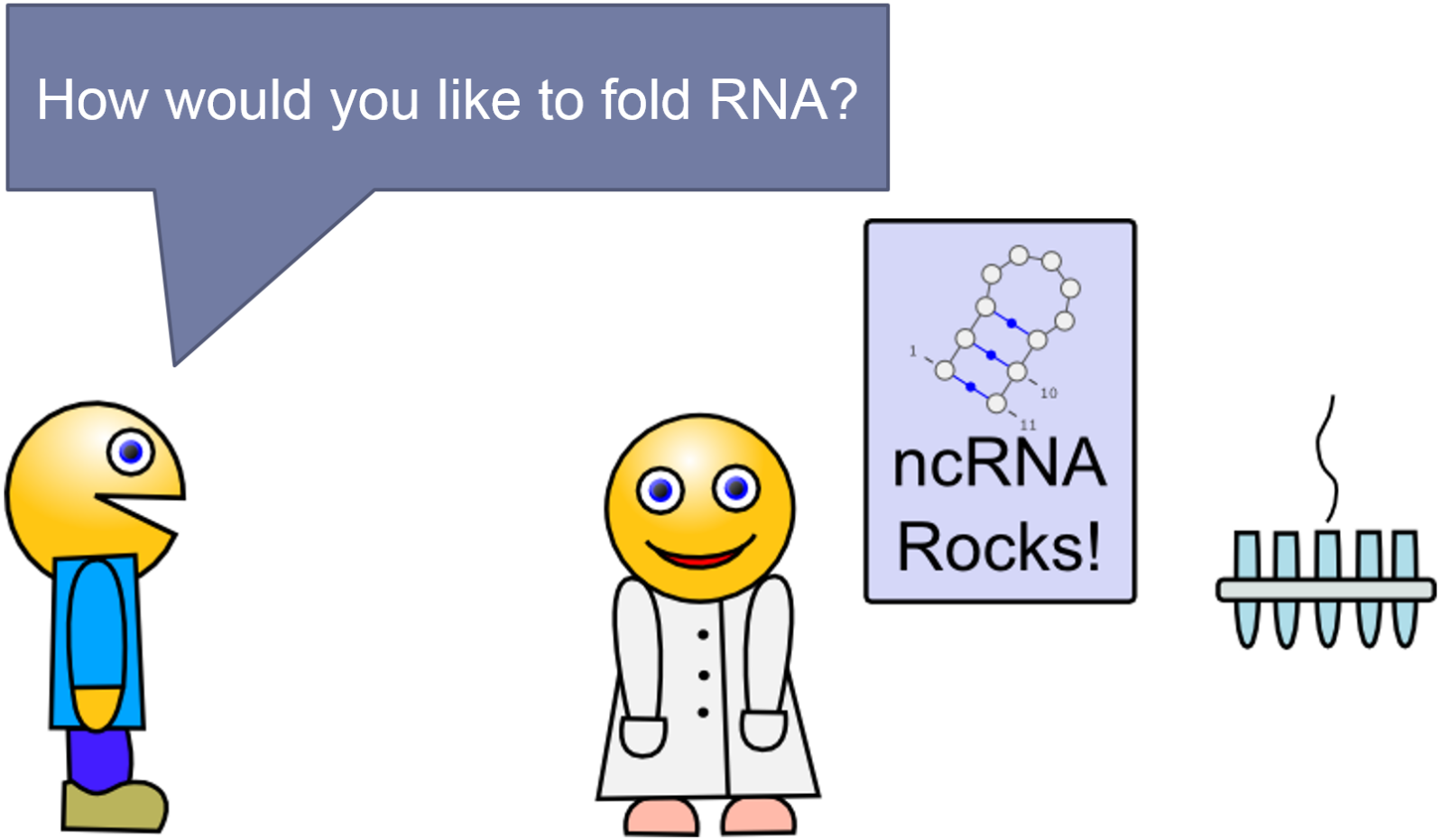
# Intermezzo

“*Traduttore tradittore*”

I can solve graph problems, why not predict RNA $2^{ary}$ structures?

...


CS
rulez!


How would you like to fold RNA?
1
10
11
ncRNA
Rocks!

... ...


CS
rulez!


## Common sense rules:

- Crossing interactions should be allowed
- But restricted to topologically valid structures
- Energy model should be realistic
- Robustness of prediction should be testable

Satisfying these rules makes the problem NP-Hard, but we can still decently approximate it, assuming that …
… APX … greedy … dynamic programming … P=NP(?)…

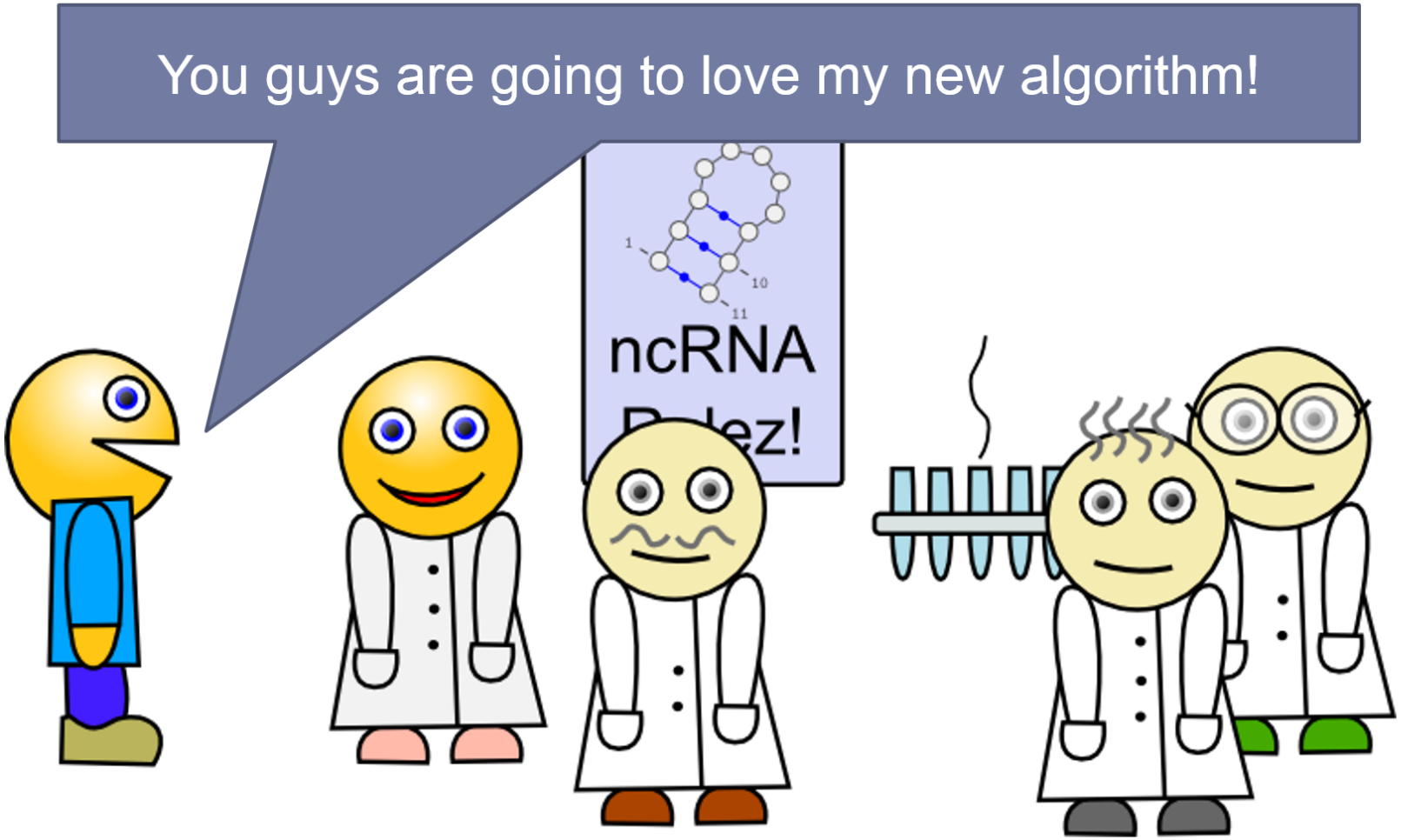
... ... ...

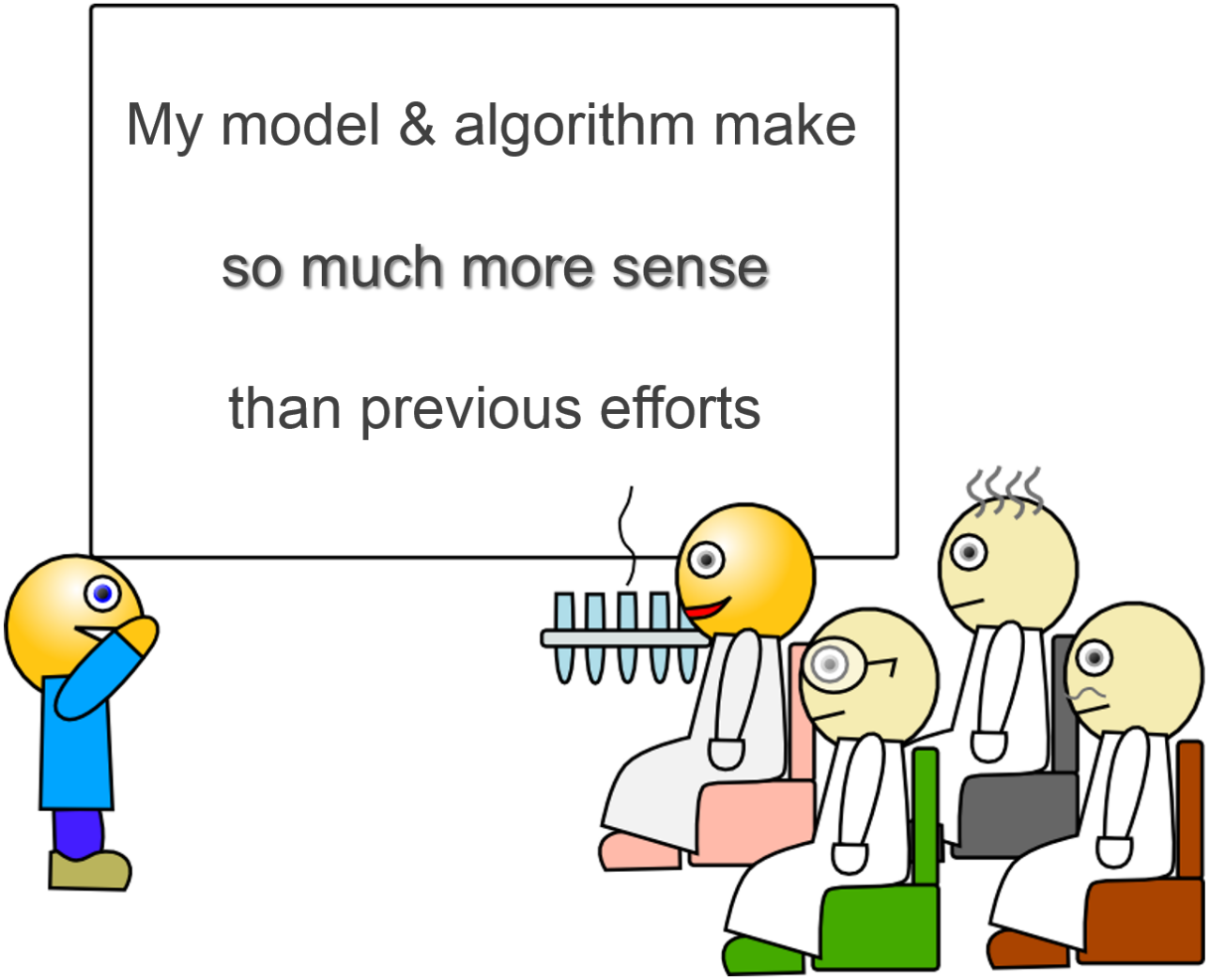

CS
rulez!


Common sense rules:

- Crossing interactions should be allowed
- But restricted to topologically valid structures
- Energy model should be realistic
- Robustness of prediction should be testable

+ Ninja algorithmic skills
+ Hard work
= Pretty decent algorithm

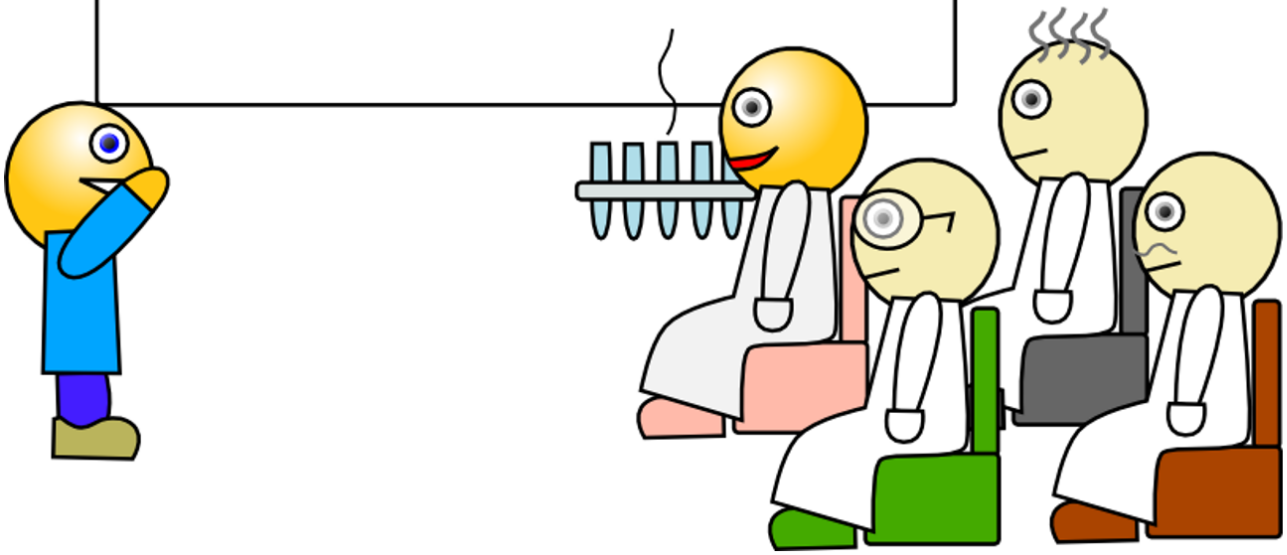

You guys are going to love my new algorithm!
ncRNA


My model & algorithm make
so much more sense
than previous efforts

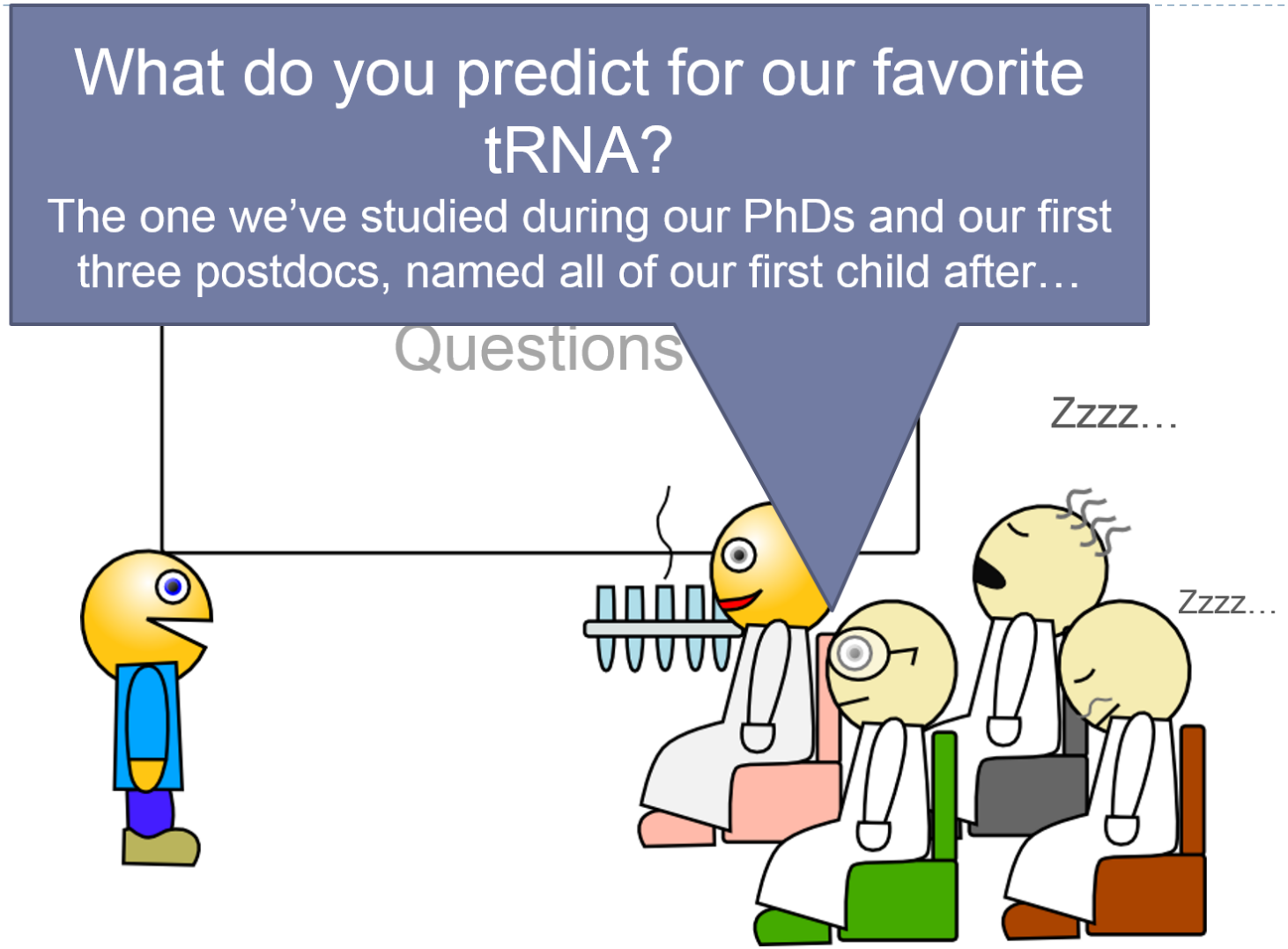
Theorem 35. The easy part

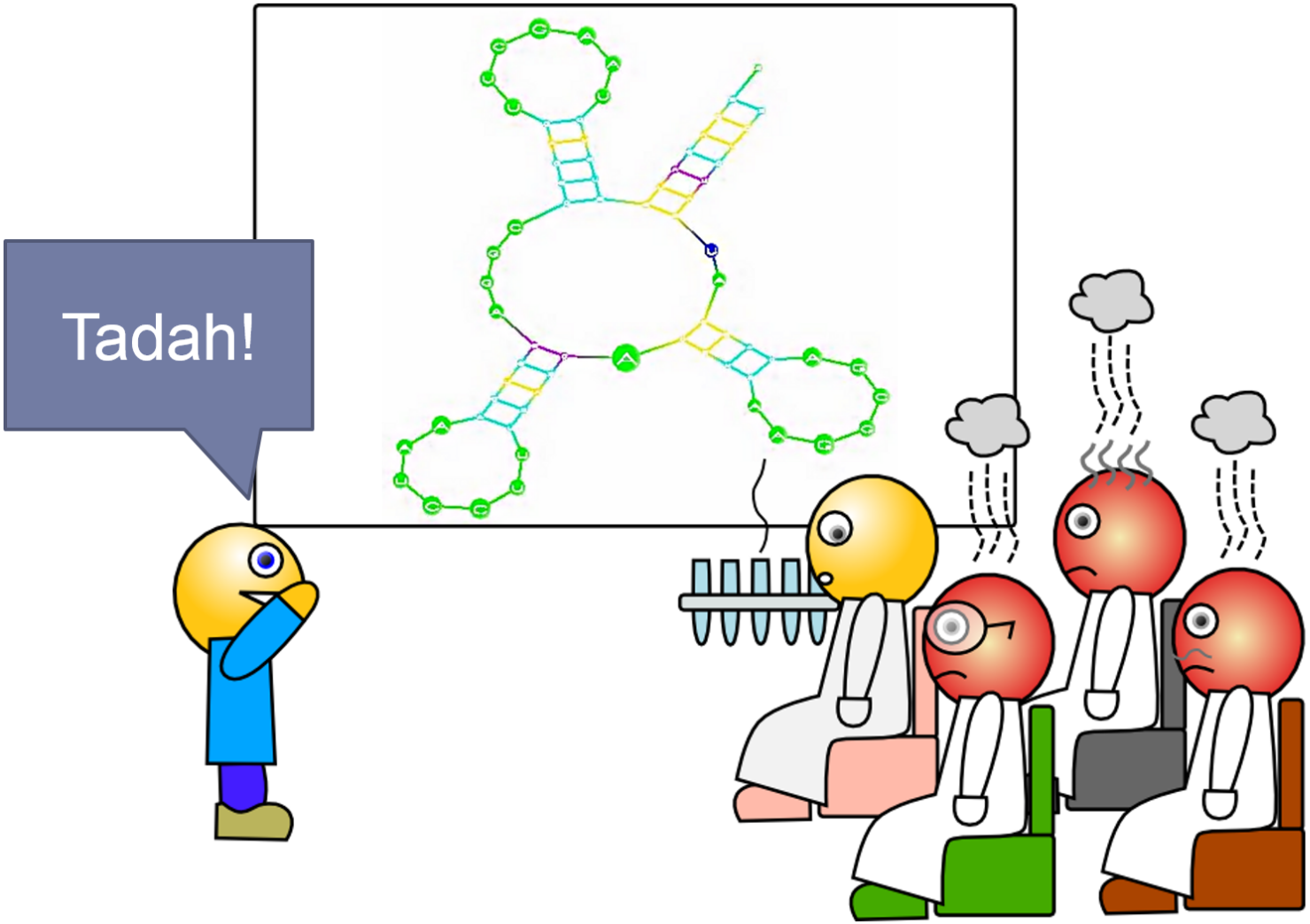
$$(x+a)^n = \sum_{k=0}^{n} \binom{n}{k} x^k a^{n-k}$$

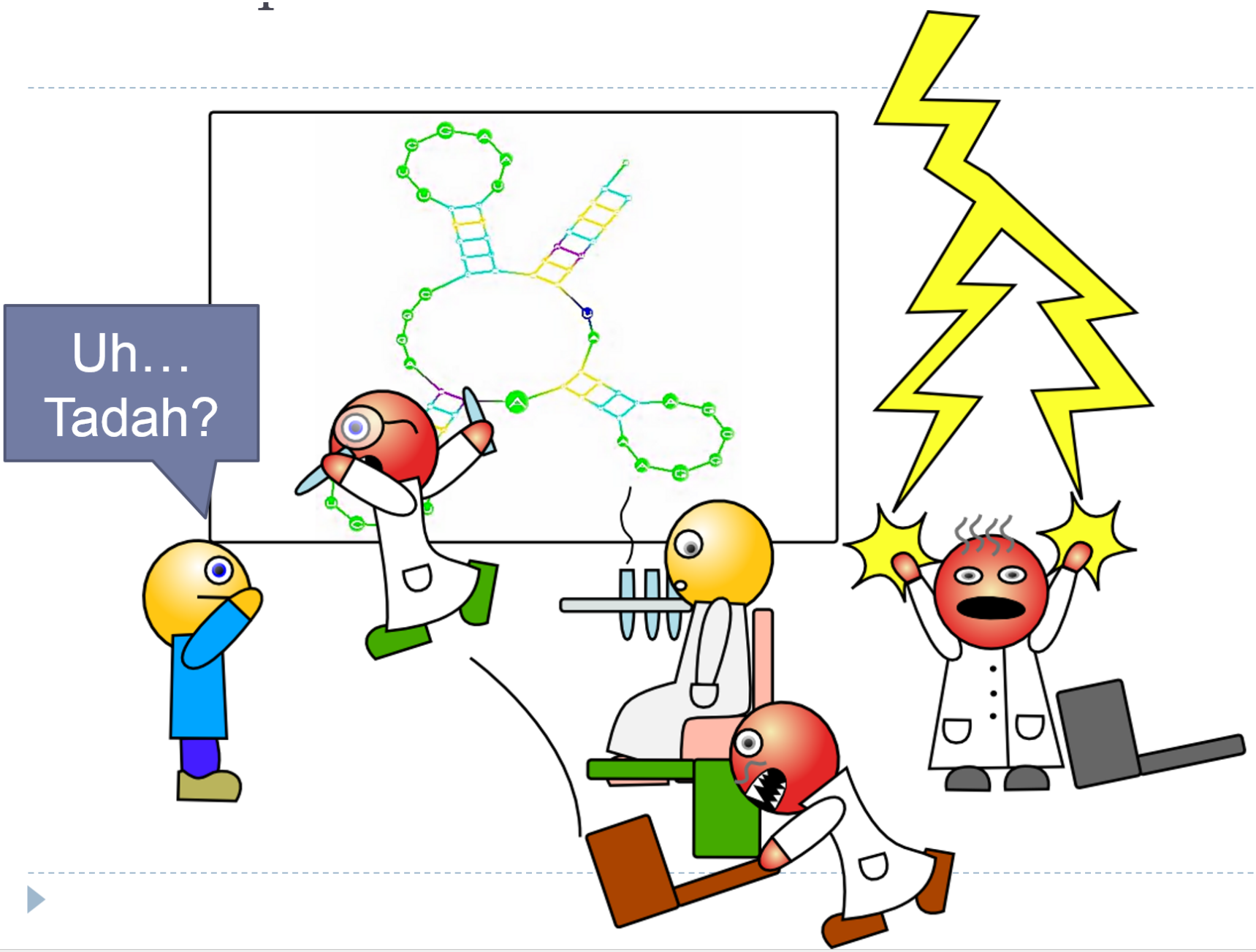
And the rest follows trivially

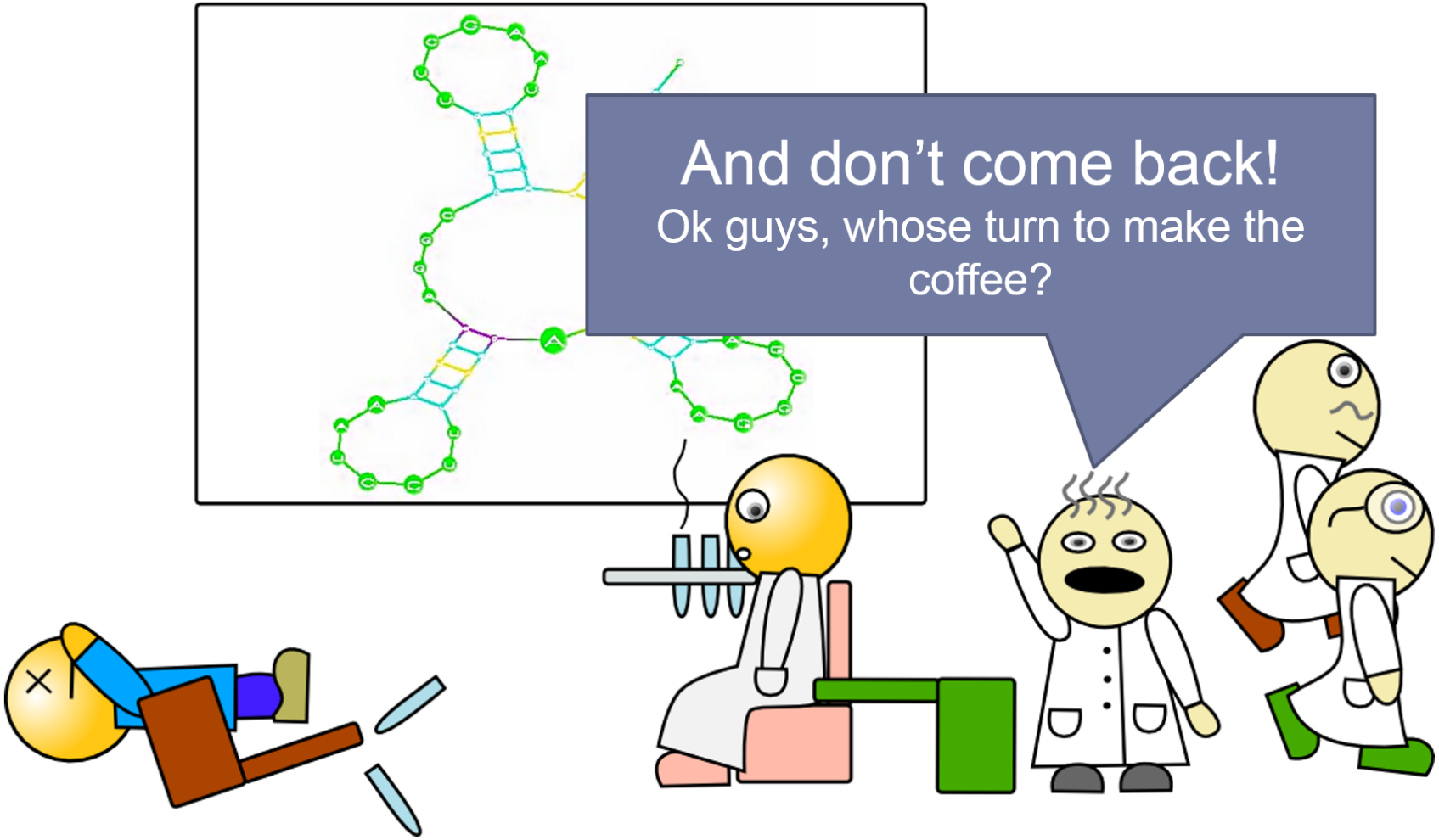


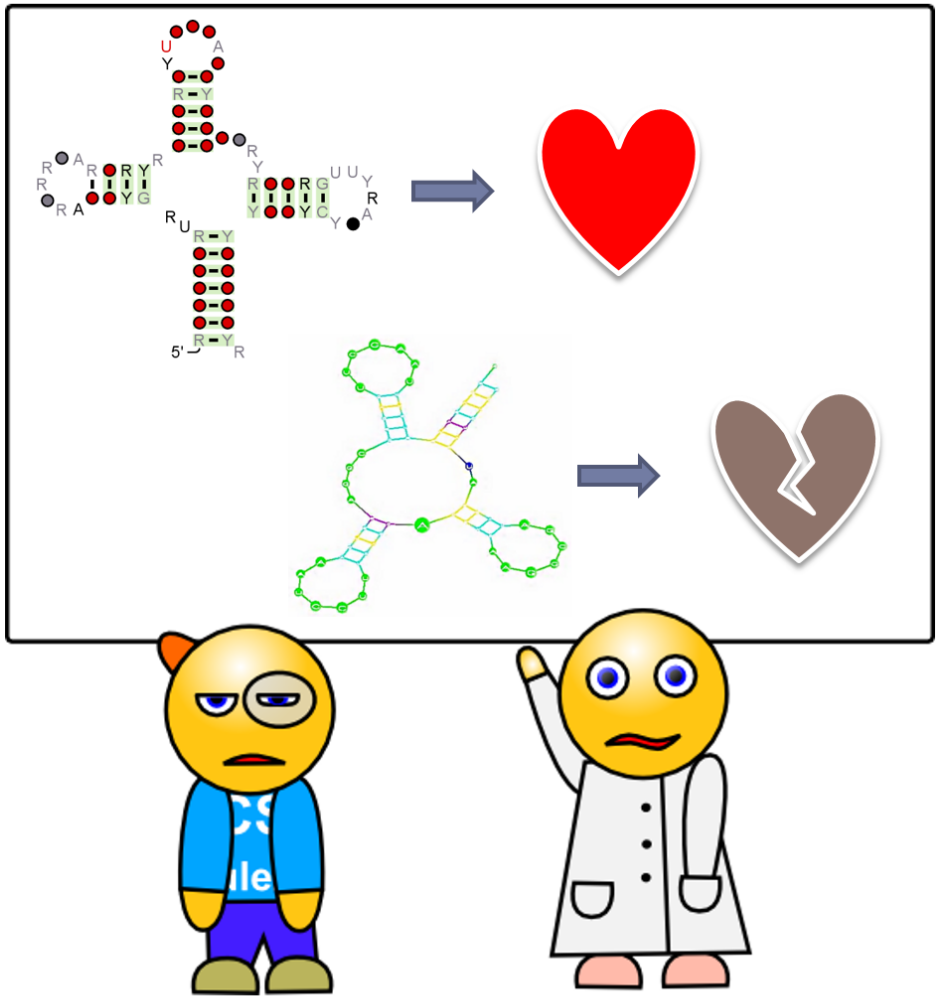

Thanks for listening.
Questions?
Zzzz…
Zzzz…

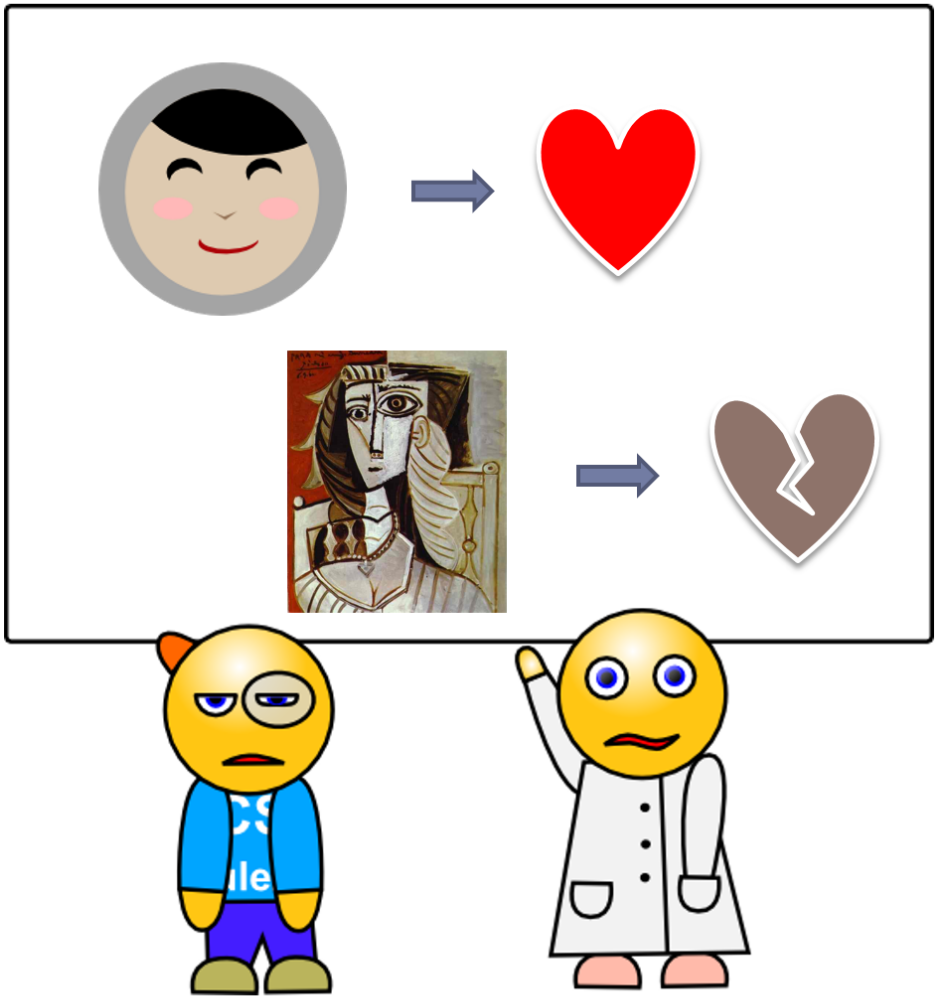

What do you predict for our favorite tRNA?
The one we've studied during our PhDs and our first three postdocs, named all of our first child after…
Questions
Zzzz…
Zzzz…


Tadah!


Uh…
Tadah?


And don’t come back!
Ok guys, whose turn to make the coffee?

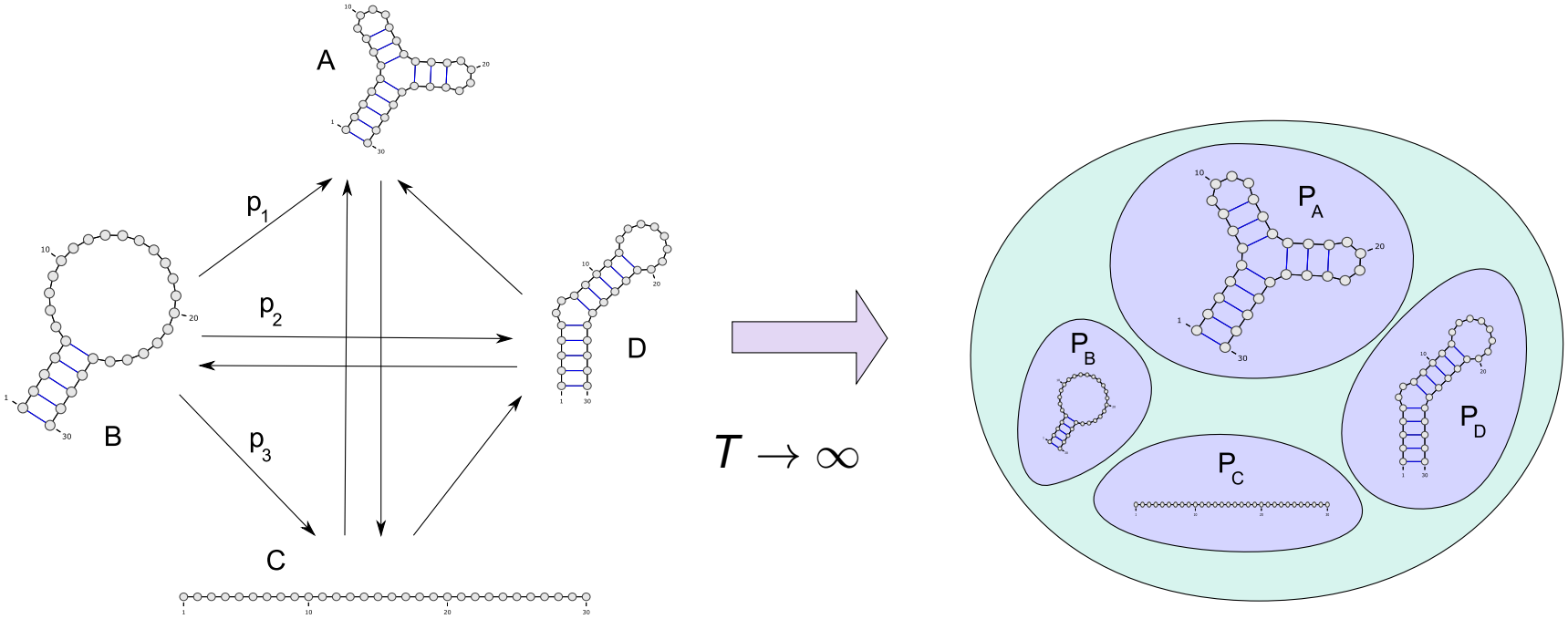




# Part. 2: Predicting how RNA folds

# Thermodynamics view

At the **nanoscale**, **RNA folding** can be adequately viewed as a **Markov process**, whose **stationary distribution** is the **Boltzmann distribution**.




## Definition (Thermodynamic equilibrium)

Each structure $S$ *compatible* with an RNA $w$ observed with probability:

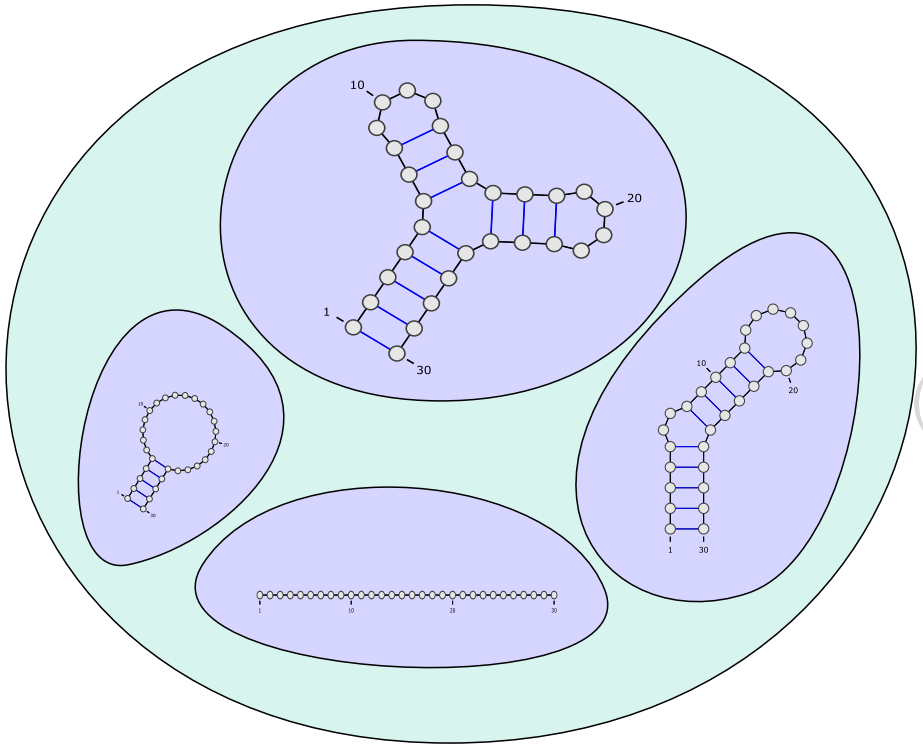
$$\mathbb{P}(S \mid w) = \frac{e^{\frac{-E_w(S)}{kT}}}{\mathcal{Z}_w} \qquad \text{and} \qquad \mathcal{Z}_w \equiv \sum_{S'} e^{\frac{-E_w(S')}{RT}} \quad \{\text{Partition function}\}$$

$E_w(S)$: **free-energy** of $S$ over $w$; $R$: Boltzmann constant; and $T$: temperature.

# Thermodynamics vs Kinetics

## Paradigms for RNA structure prediction

- **1978–1990s** Most probable structure = Minimal Free-Energy (MFE)
- **1990s–2010s** Functional structure(s) = Boltzmann ensemble (partition function)
- **2010s–????** Embracing the kinetics view



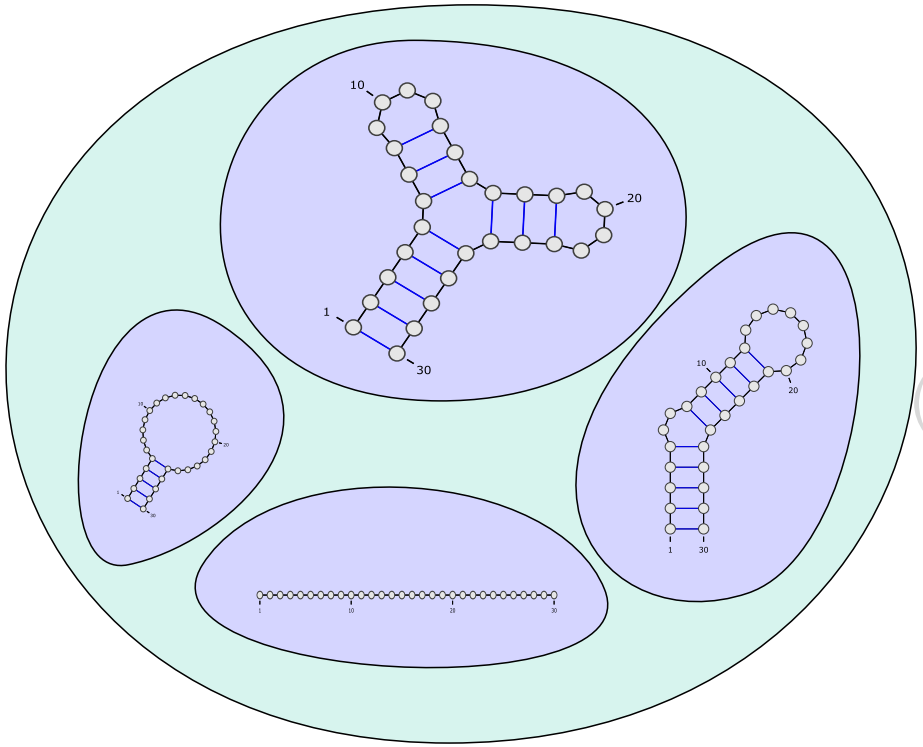

$T \to \infty$

mRNA half-life: $\sim$7h
(Mouse [Sharova et al., 2009])

# Thermodynamics vs Kinetics

## Paradigms for RNA structure prediction

- **1978–1990s** Most probable structure = Minimal Free-Energy (MFE)
- **1990s–2010s** Functional structure(s) = Boltzmann ensemble (partition function)
- **2010s–????** Embracing the kinetics view



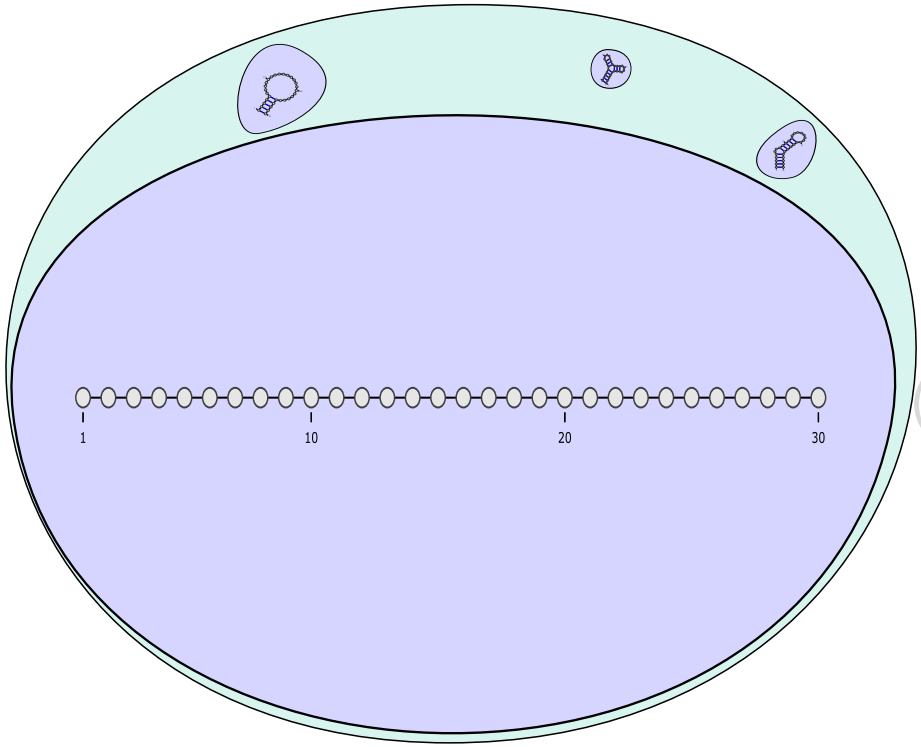

$T \to \infty$

mRNA half-life: $\sim$7h
(Mouse [Sharova et al., 2009])

# Thermodynamics vs Kinetics

## Paradigms for RNA structure prediction

- **1978–1990s** Most probable structure = Minimal Free-Energy (MFE)
- **1990s–2010s** Functional structure(s) = Boltzmann ensemble (partition function)
- **2010s–????** Embracing the kinetics view



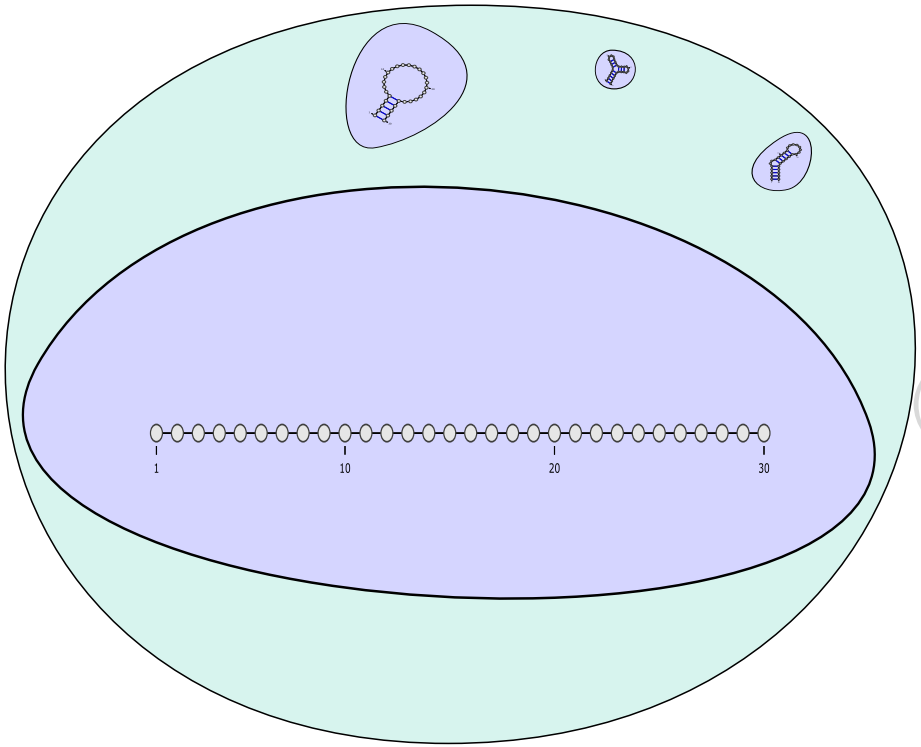

$T = 0$h

mRNA half-life: $\sim$7h
(Mouse [Sharova et al., 2009])

# Thermodynamics vs Kinetics

## Paradigms for RNA structure prediction

- **1978–1990s** Most probable structure = Minimal Free-Energy (MFE)
- **1990s–2010s** Functional structure(s) = Boltzmann ensemble (partition function)
- **2010s–????** Embracing the kinetics view



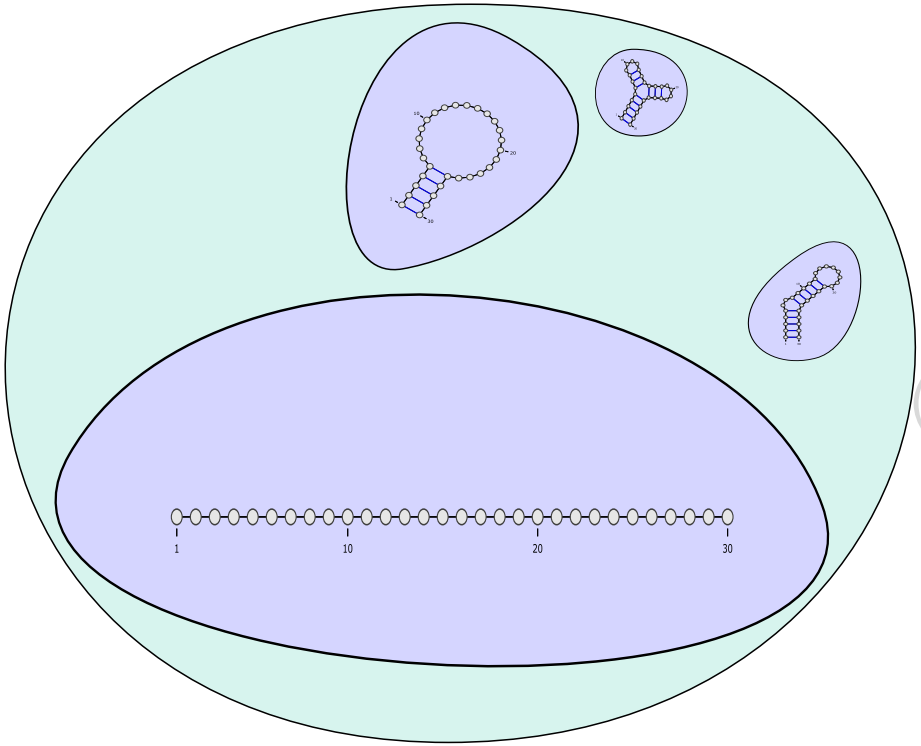

$T = 1\text{h}$

mRNA half-life: $\sim$7h
(Mouse [Sharova et al., 2009])

# Thermodynamics vs Kinetics

## Paradigms for RNA structure prediction

- **1978–1990s** Most probable structure = Minimal Free-Energy (MFE)
- **1990s–2010s** Functional structure(s) = Boltzmann ensemble (partition function)
- **2010s–????** Embracing the kinetics view



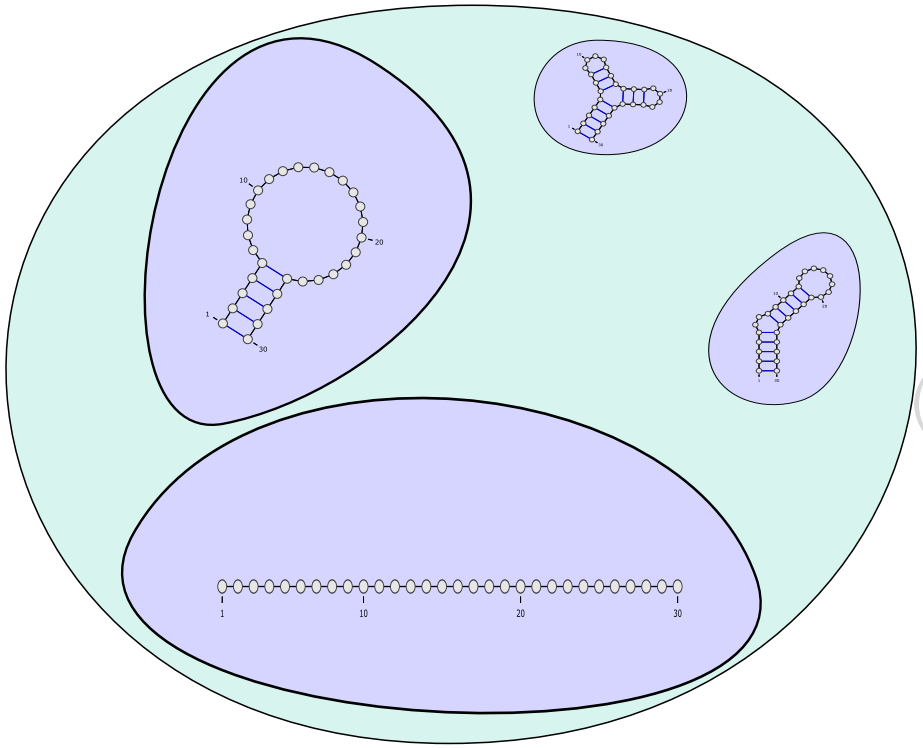

$T = 2\text{h}$

mRNA half-life: $\sim$7h
(Mouse [Sharova et al., 2009])

# Thermodynamics vs Kinetics

## Paradigms for RNA structure prediction

- **1978–1990s** Most probable structure = Minimal Free-Energy (MFE)
- **1990s–2010s** Functional structure(s) = Boltzmann ensemble (partition function)
- **2010s–????** Embracing the kinetics view



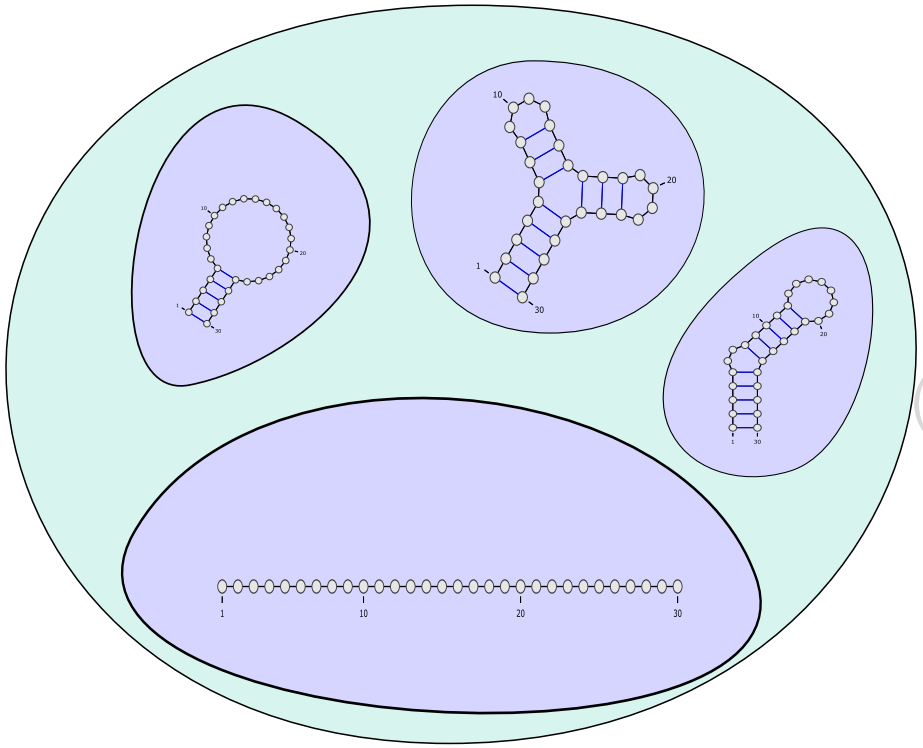

$T = 5$h

mRNA half-life: $\sim$7h
(Mouse [Sharova et al., 2009])

# Thermodynamics vs Kinetics

## Paradigms for RNA structure prediction

- **1978–1990s** Most probable structure = Minimal Free-Energy (MFE)
- **1990s–2010s** Functional structure(s) = Boltzmann ensemble (partition function)
- **2010s–????** Embracing the kinetics view



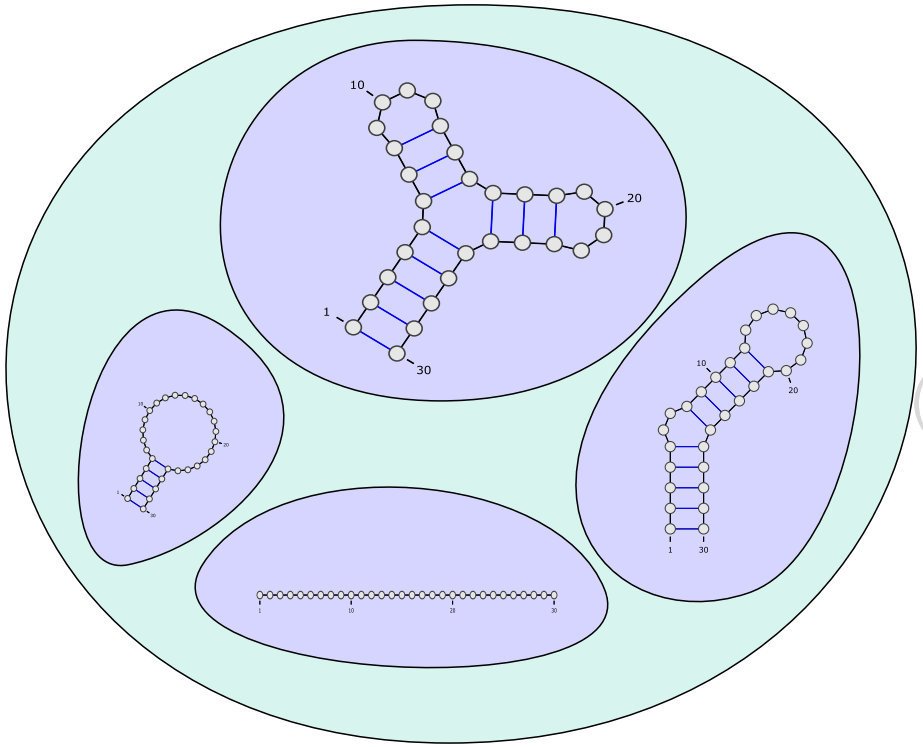
$T = 10$h

mRNA half-life: $\sim$7h
(Mouse [Sharova et al., 2009])

# Thermodynamics vs Kinetics

## Paradigms for RNA structure prediction

- **1978–1990s** Most probable structure = Minimal Free-Energy (MFE)
- **1990s–2010s** Functional structure(s) = Boltzmann ensemble (partition function)
- **2010s–????** Embracing the kinetics view



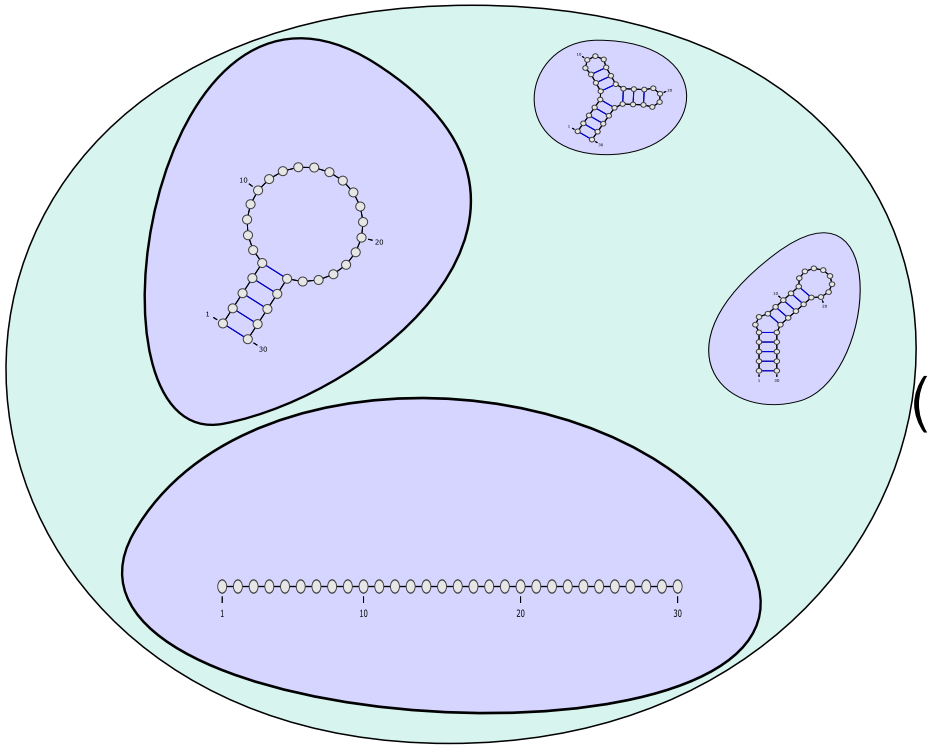

$T \to \infty$

mRNA half-life: $\sim$7h
(Mouse [Sharova et al., 2009])

# Thermodynamics vs Kinetics

## Paradigms for RNA structure prediction

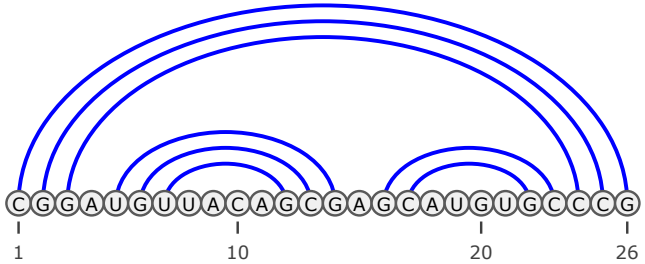
- **1978–1990s** Most probable structure = Minimal Free-Energy (MFE)
- **1990s–2010s** Functional structure(s) = Boltzmann ensemble (partition function)
- **2010s–????** Embracing the kinetics view

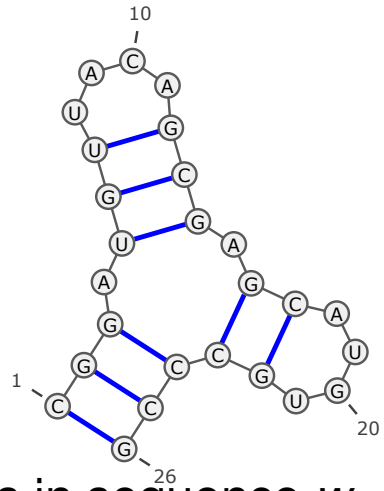


$T = 10$h

mRNA half-life: $\sim$7h
(Mouse [Sharova et al., 2009])

# Free-energy




- **RNA structure $S$:** (Partial) matching of positions in sequence $w$
- **Motifs:** Sequence/structure features (e.g. pairs, Stacking pairs, Nearest neighbor...)
- **Energy model.**

  **Motif** $\rightarrow$ Free-energy contribution $\Delta G(\cdot) \in \mathbb{R}^- \cup \{+\infty\}$

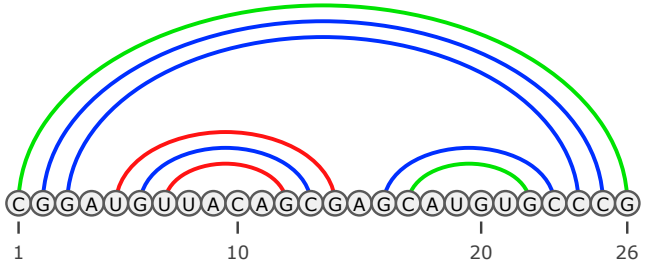
  **Free-Energy $E_w(S)$:** Sum over (independently contributing) motifs in $S$

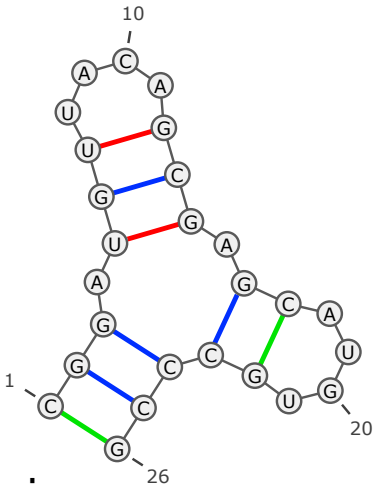
**Nussinov/Jacobson energy model** [Nussinov and Jacobson, 1980]

**Pairs:** $\Delta G(x, y) = \begin{cases} -1 \ (-3/-2/-1) & \text{if } (x, y) = (G \equiv C)/(A = U)/(G - U) \\ +\infty & \text{otherwise.} \end{cases}$

**Rem.:** Structure prediction $\approx$ Energy minimization $\Leftrightarrow$ Base-pair maximization
$\Leftrightarrow$ Max (weighted) independent set in circle graph

# Free-energy




- **RNA structure $S$:** (Partial) matching of positions in sequence $w$
- **Motifs:** Sequence/structure features (e.g. pairs, Stacking pairs, Nearest neighbor...)
- **Energy model.**

  **Motif** $\rightarrow$ Free-energy contribution $\Delta G(\cdot) \in \mathbb{R}^- \cup \{+\infty\}$

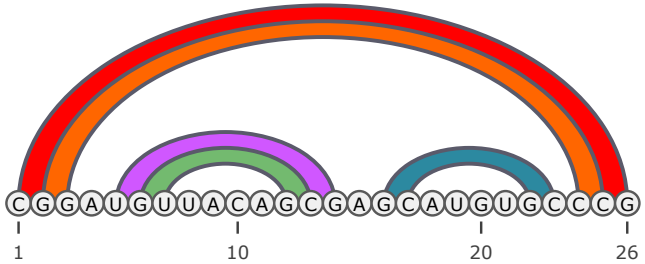
  **Free-Energy $E_w(S)$:** Sum over (independently contributing) motifs in $S$

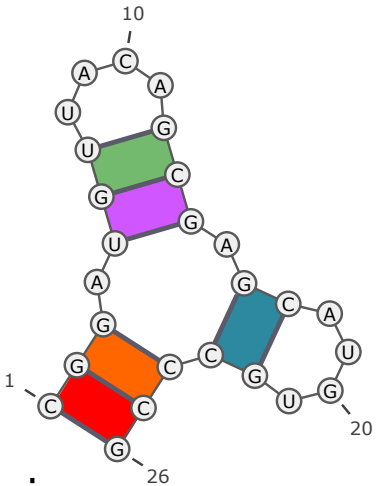
**Nussinov/Jacobson energy model** [Nussinov and Jacobson, 1980]

**Pairs:** $\Delta G(x, y) = \begin{cases} -1 \ (-3/-2/-1) & \text{if } (x, y) = (G \equiv C)/(A = U)/(G - U) \\ +\infty & \text{otherwise.} \end{cases}$

**Rem.:** Structure prediction $\approx$ Energy minimization $\Leftrightarrow$ Base-pair maximization $\Leftrightarrow$ Max (weighted) independent set in circle graph

# Free-energy




- **RNA structure $S$:** (Partial) matching of positions in sequence $w$
- **Motifs:** Sequence/structure features (e.g. pairs, Stacking pairs, Nearest neighbor...)
- **Energy model.**

  **Motif** $\rightarrow$ Free-energy contribution $\Delta G(\cdot) \in \mathbb{R}^- \cup \{+\infty\}$

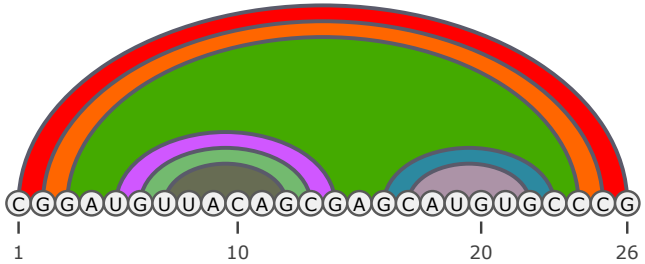
  **Free-Energy $E_w(S)$:** Sum over (independently contributing) motifs in $S$

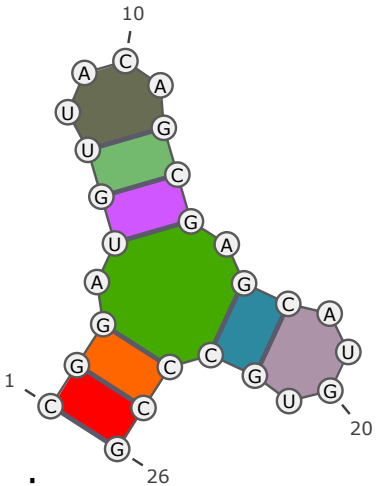
**Nussinov/Jacobson energy model** [Nussinov and Jacobson, 1980]

**Pairs:** $\Delta G(x, y) = \begin{cases} -1 \ (-3/-2/-1) & \text{if } (x, y) = (G \equiv C)/(A = U)/(G - U) \\ +\infty & \text{otherwise.} \end{cases}$

**Rem.:** Structure prediction $\approx$ Energy minimization $\Leftrightarrow$ Base-pair maximization $\Leftrightarrow$ Max (weighted) independent set in circle graph

# Free-energy




- **RNA structure $S$:** (Partial) matching of positions in sequence $w$
- **Motifs:** Sequence/structure features (e.g. pairs, Stacking pairs, Nearest neighbor...)
- **Energy model.**

  **Motif** $\rightarrow$ Free-energy contribution $\Delta G(\cdot) \in \mathbb{R}^- \cup \{+\infty\}$

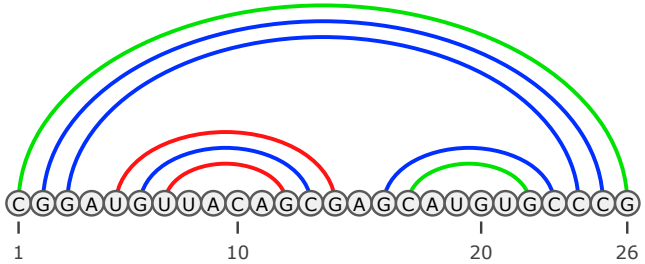
  **Free-Energy $E_w(S)$:** Sum over (independently contributing) motifs in $S$

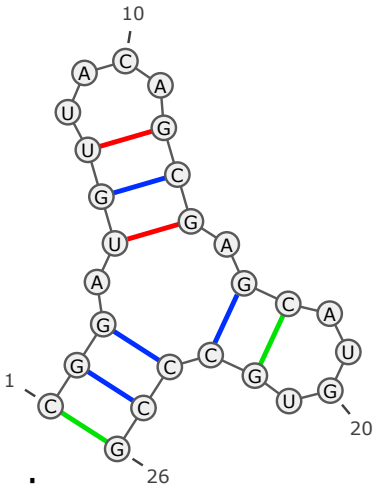
**Nussinov/Jacobson energy model** [Nussinov and Jacobson, 1980]

**Pairs:** $\Delta G(x, y) = \begin{cases} -1 \ (-3/-2/-1) & \text{if } (x, y) = (G \equiv C)/(A = U)/(G - U) \\ +\infty & \text{otherwise.} \end{cases}$

**Rem.:** Structure prediction $\approx$ Energy minimization $\Leftrightarrow$ Base-pair maximization $\Leftrightarrow$ Max (weighted) independent set in circle graph

# Free-energy




- **RNA structure $S$:** (Partial) matching of positions in sequence $w$
- **Motifs:** Sequence/structure features (e.g. pairs, Stacking pairs, Nearest neighbor...)
- **Energy model.**

  **Motif** $\rightarrow$ Free-energy contribution $\Delta G(\cdot) \in \mathbb{R}^- \cup \{+\infty\}$

  **Free-Energy $E_w(S)$:** Sum over (independently contributing) motifs in $S$

**Nussinov/Jacobson energy model** [Nussinov and Jacobson, 1980]

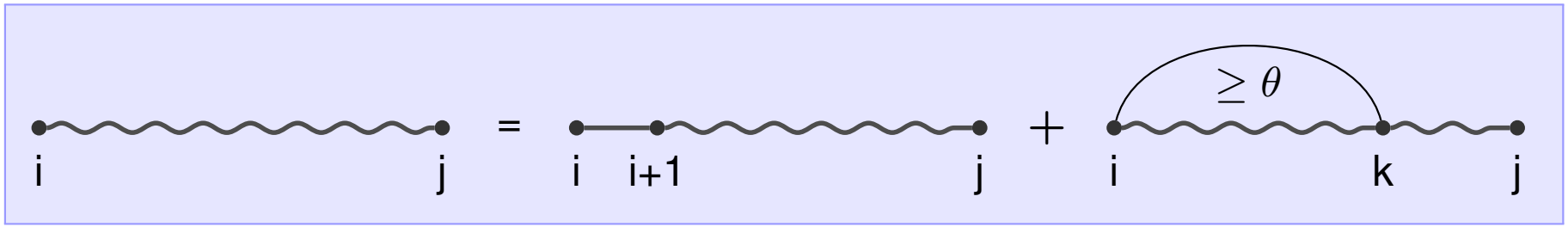
**Pairs:** $\Delta G(x,y) = \begin{cases} \text{-1 (-3/-2/-1)} & \text{if } (x,y) = (G \equiv C)/(A = U)/(G - U) \\ +\infty & \text{otherwise.} \end{cases}$

**Rem.:** Structure prediction $\approx$ Energy minimization $\Leftrightarrow$ Base-pair maximization
$\Leftrightarrow$ Max (weighted) independent set in circle graph

# Dynamic programming (DP) for RNA folding

**Theorem (Nussinov and Jacobson [1980])**

*Max #base-pairs/min weight structure is computable in $\mathcal{O}(n^3)/\mathcal{O}(n^2)$ time/memory*




$E_{i,k}$: Free-energy contribution of base-pair $(i,k)$. $(-1/+\infty$ or $\Delta G(s_i \overset{?}{\equiv} s_k))$

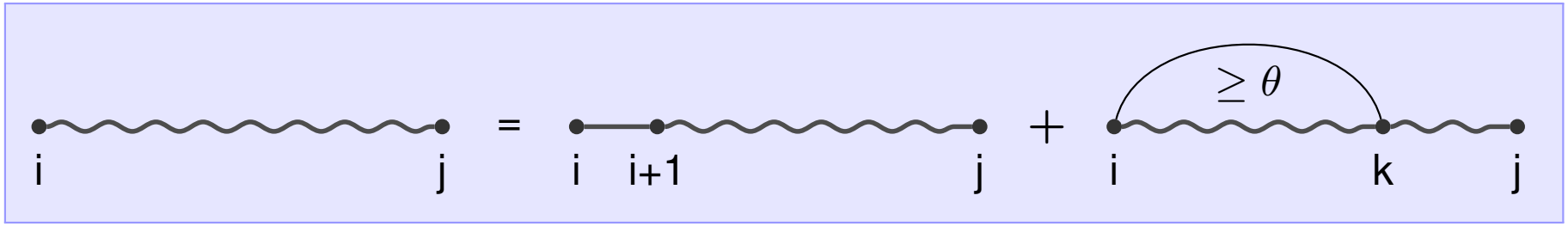
$\boldsymbol{N}_{i,j}$ : Max #base-pairs over interval $[i,j]$

$$\begin{aligned}
\boldsymbol{N}_{i,t} &= 0, \quad \forall t \in [i, i+\theta] \\
\boldsymbol{N}_{i,j} &= \min \begin{cases} \boldsymbol{N}_{i+1,j} & \{i \text{ unpaired}\} \\ \min\limits_{k=i+\theta+1}^{j} E_{i,k} + \boldsymbol{N}_{i+1,k-1} + \boldsymbol{N}_{k+1,j} & \{i \text{ paired to } k\} \end{cases}
\end{aligned}$$

# Dynamic programming (DP) for RNA folding

**Theorem (Nussinov and Jacobson [1980])**

*Max #base-pairs/min weight structure is computable in $\mathcal{O}(n^3)/\mathcal{O}(n^2)$ time/memory*




$E_{i,k}$: Free-energy contribution of base-pair $(i,k)$. $(-1/+\infty$ or $\Delta G(s_i \overset{?}{\equiv} s_k))$

$C_{i,j}$ : Number of secondary structures compatible with interval $[i,j]$

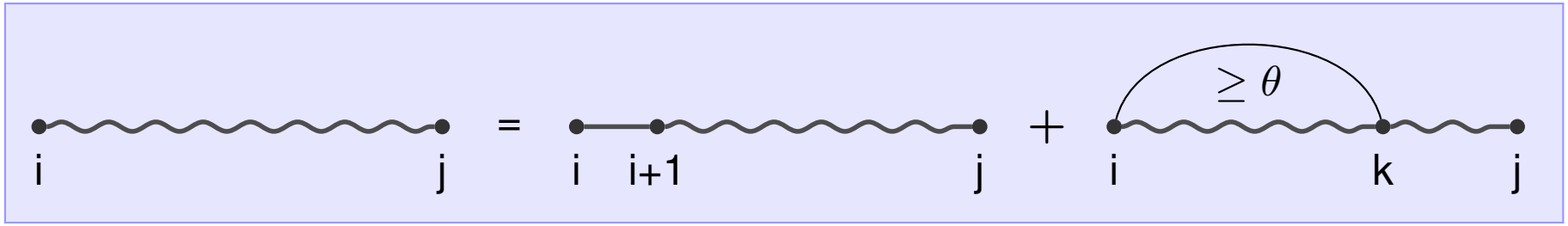
$$C_{i,t} = 1, \quad \forall t \in [i, i+\theta]$$

$$C_{i,j} = \sum \begin{cases} C_{i+1,j} & \{i \text{ unpaired}\} \\ \sum_{k=i+\theta+1}^{j} \mathbb{1}_{\text{comp.}(i,k)} \times C_{i+1,k-1} \times C_{k+1,j} & \{i \text{ paired to } k\} \end{cases}$$

# Dynamic programming (DP) for RNA folding

**Theorem (Nussinov and Jacobson [1980])**

*Max #base-pairs/min weight structure is computable in $\mathcal{O}(n^3)/\mathcal{O}(n^2)$ time/memory*




$E_{i,k}$: Free-energy contribution of base-pair $(i,k)$. $(-1/+\infty$ or $\Delta G(s_i \overset{?}{\equiv} s_k))$

$\mathcal{Z}_{i,j} = \sum_{\substack{S \text{ comp.} \\ \text{with } w_{[i,j]}}} e^{\frac{-E_W(S)}{RT}}$ = Partition function of structures compatible with interval $[i,j]$

$$\mathcal{Z}_{i,t} = \mathbf{1}, \quad \forall t \in [i, i+\theta]$$

$$\mathcal{Z}_{i,j} = \sum \begin{cases} \mathcal{Z}_{i+1,j} & \{i \text{ unpaired}\} \\ \sum_{k=i+\theta+1}^{j} e^{\frac{-E_{i,k}}{RT}} \times \mathcal{Z}_{i+1,k-1} \times \mathcal{Z}_{k+1,j} & \{i \text{ paired to } k\} \end{cases}$$

# Dynamic programming (DP) for RNA folding

**Many extensions:**

- Nearest-neighbor/Turner energy model [Zuker and Stiegler, 1981]
- Comparative folding [Sankoff, 1985]
- Equilibrium base-pairing probabilities [McCaskill, 1990]
- Moments of additive features [Miklós et al., 2005; Ponty and Saule, 2011]
- $\Delta$ kcal.mol$^{-1}$ suboptimal structures of MFE [Wuchty et al., 1999]
- Basic crossing structures [Rivas and Eddy, 1999]...
- Exact sampling in Boltzmann distr. [Ding and Lawrence, 2003; Ponty, 2008]
- Moments of additive features [Miklós et al., 2005; Ponty and Saule, 2011]
- Maximum expected accuracy structure [Do et al., 2006]
- Distance-classified partitioning of Boltzmann ens. [Freyhult et al., 2007]

**Made possible by:**

- **Completeness**/**Unambiguity** of decomposition
  $\exists$ energy-preserving bijection between **derivations of DP scheme** and **search space**
- Objective function **additive** with respect to DP scheme

$\Rightarrow$ **Combinatorial Dynamic Programming**

# Part. 3: Combinatorial Dynamic Programming

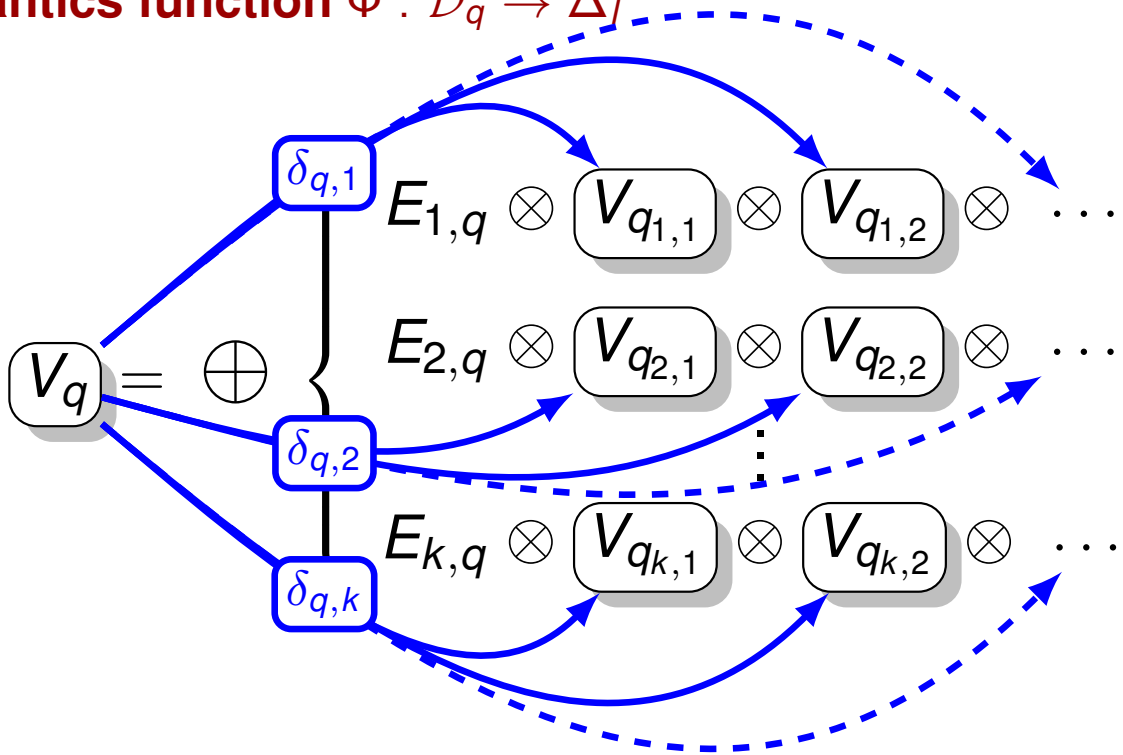
# Reasoning structurally about dynamic programming

- $I$ : Instance (aka problem)
- $Q$ : State space for dyn. prog. scheme (LHS terms, $I$ initial state)
- $\Delta_q$ : Search space for state $q$
- $\mathcal{D}_q$ : Derivations of DP scheme from state $q \in Q_i$
- **Semantics function** $\Phi : \mathcal{D}_q \to \Delta_I$

$$V_q = \bigoplus \begin{cases} E_{1,q} \otimes V_{q_{1,1}} \otimes V_{q_{1,2}} \otimes \cdots \\ E_{2,q} \otimes V_{q_{2,1}} \otimes V_{q_{2,2}} \otimes \cdots \\ \vdots \\ E_{k,q} \otimes V_{q_{k,1}} \otimes V_{q_{k,2}} \otimes \cdots \end{cases}$$

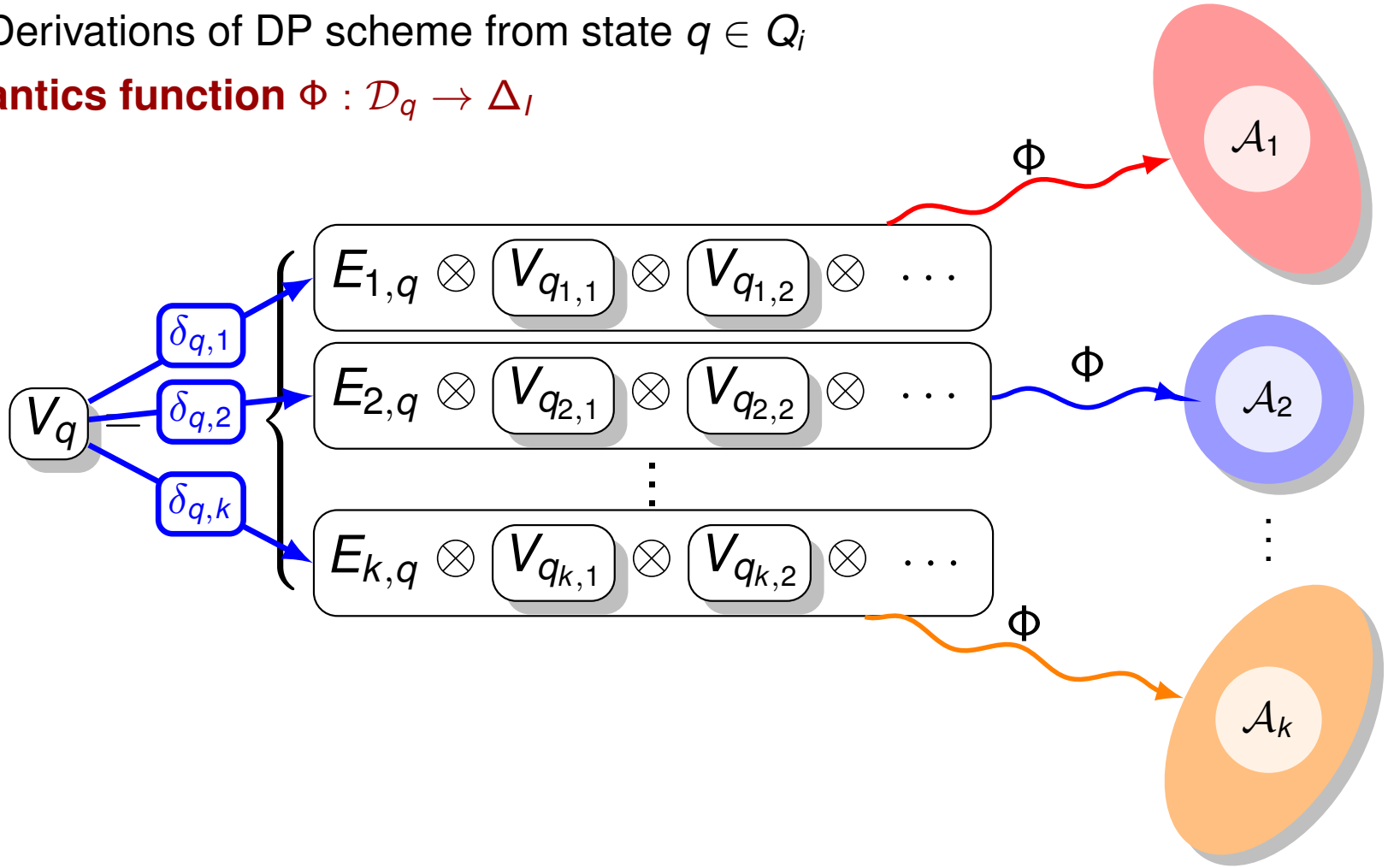
# Reasoning structurally about dynamic programming

- $I$ : Instance (aka problem)
- $Q$ : State space for dyn. prog. scheme (LHS terms, $I$ initial state)
- $\Delta_q$ : Search space for state $q$
- $\mathcal{D}_q$ : Derivations of DP scheme from state $q \in Q_i$
- **Semantics function** $\Phi : \mathcal{D}_q \to \Delta_I$



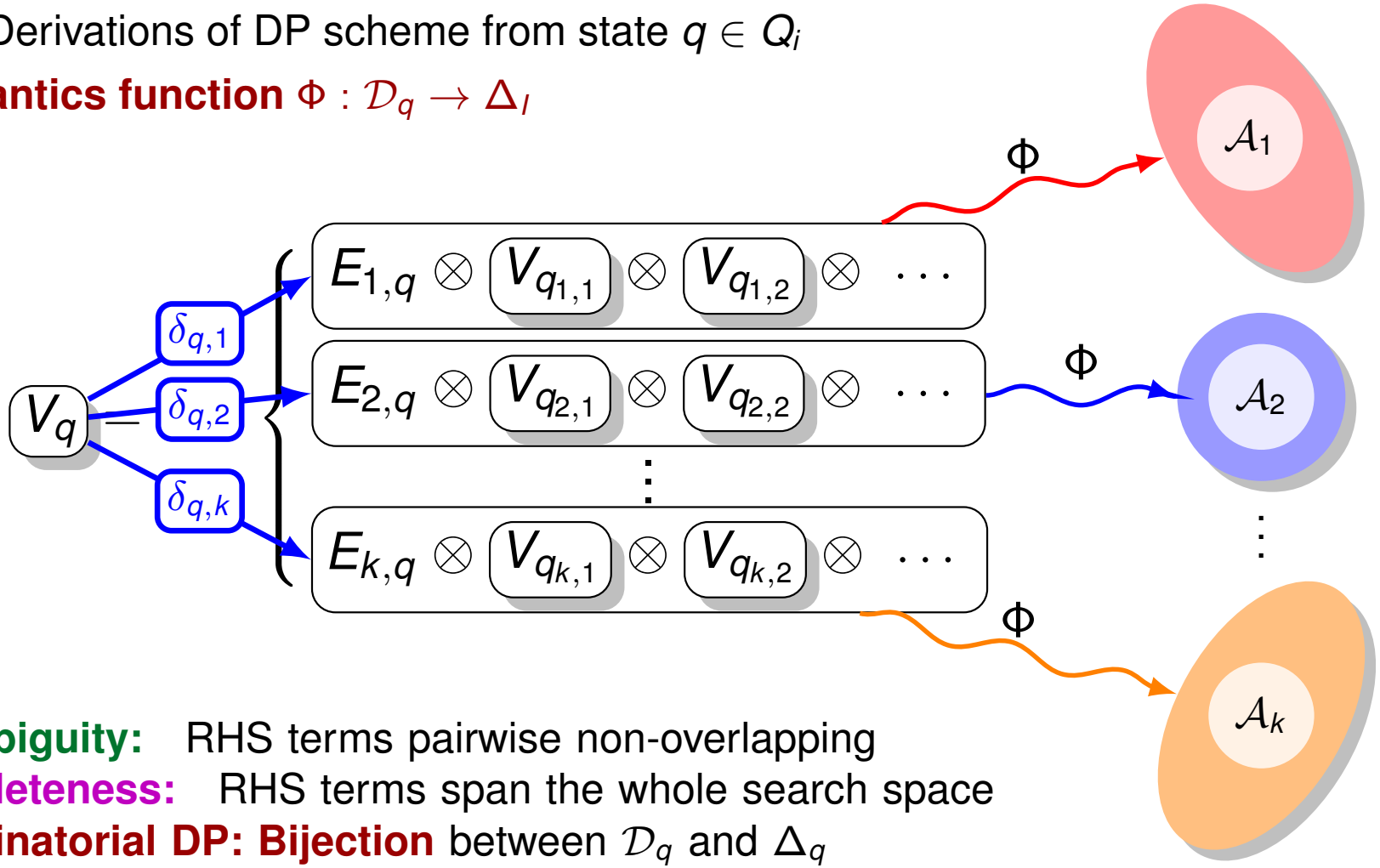
# Reasoning structurally about dynamic programming

- $I$ : Instance (aka problem)
- $Q$ : State space for dyn. prog. scheme (LHS terms, $I$ initial state)
- $\Delta_q$ : Search space for state $q$
- $\mathcal{D}_q$ : Derivations of DP scheme from state $q \in Q_i$
- **Semantics function** $\Phi : \mathcal{D}_q \to \Delta_I$



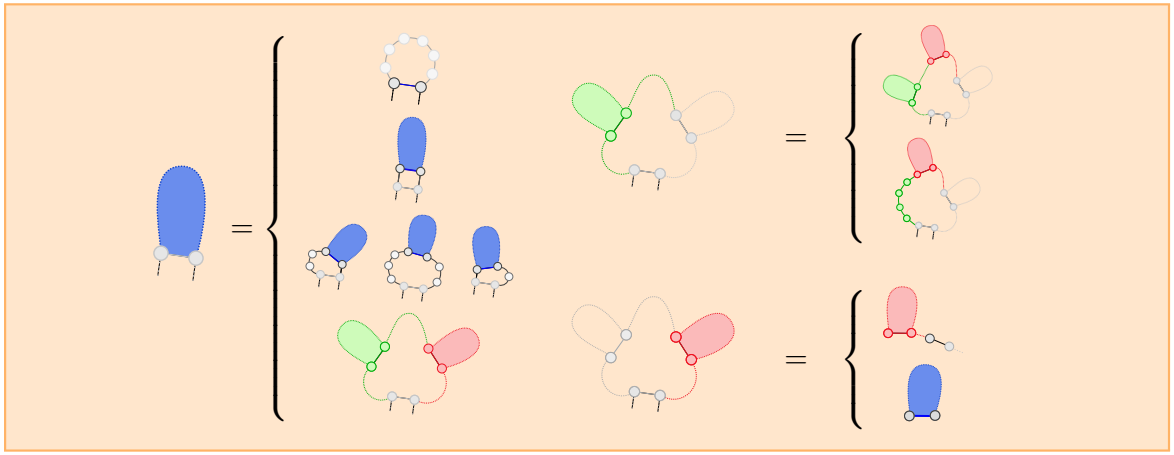
# Reasoning structurally about dynamic programming

- $I$ : Instance (aka problem)
- $Q$ : State space for dyn. prog. scheme (LHS terms, $I$ initial state)
- $\Delta_q$ : Search space for state $q$
- $\mathcal{D}_q$ : Derivations of DP scheme from state $q \in Q_i$
- **Semantics function** $\Phi : \mathcal{D}_q \to \Delta_I$



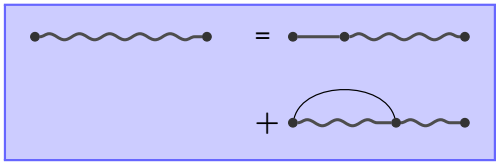

**Unambiguity:** RHS terms pairwise non-overlapping
**Completeness:** RHS terms span the whole search space
**Combinatorial DP: Bijection** between $\mathcal{D}_q$ and $\Delta_q$

# Combinatorics help in the design of DP schemes



MFold DP scheme [Zuker and Stiegler, 1981]

$\overset{??}{\Longleftrightarrow}$



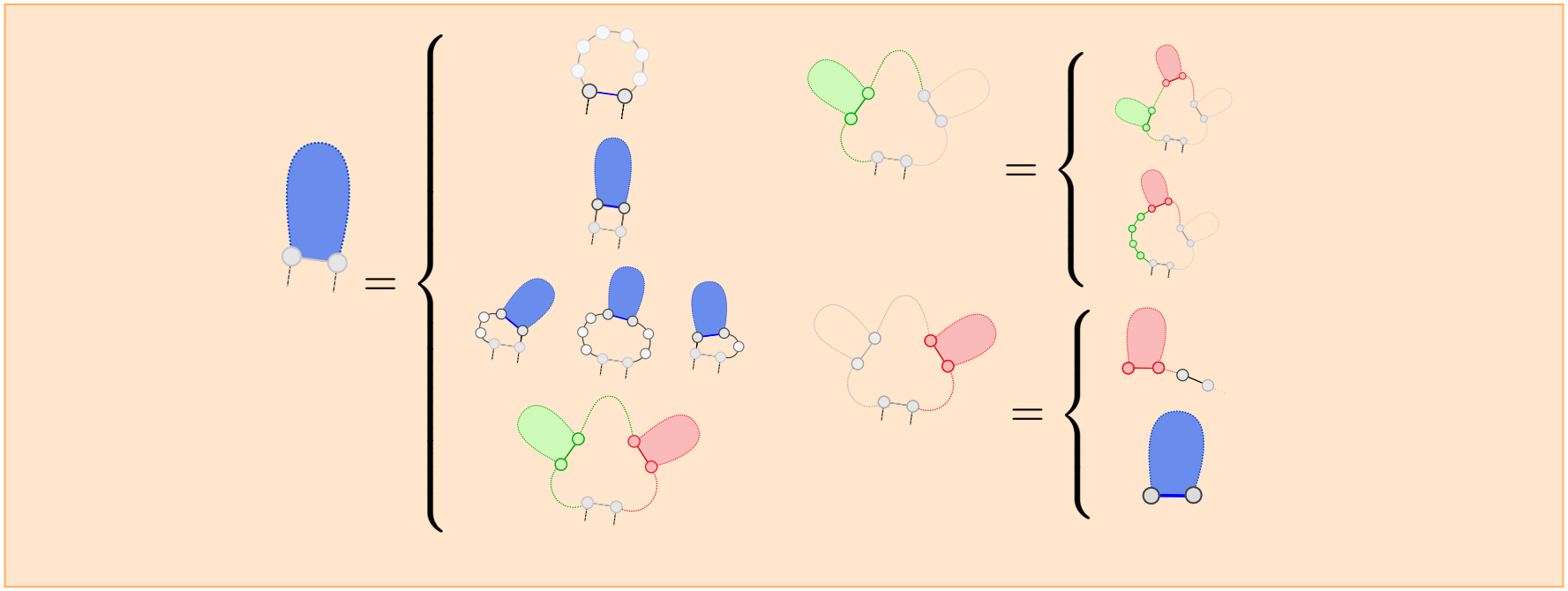
**Unambiguous** (pairwise non-overlapping generated search spaces)
**Completeness**? Use generating functions...

# Combinatorics help in the design of DP schemes

**Reminder:** Generating function of secondary structures [Waterman, 1978]

$$S(z) := \sum_{n \geq 0} s_n \, z^n = \frac{1 - z + z^2 - \sqrt{1 - 2z - z^2 - 2z^3 + z^4}}{2z^2}$$

- DP scheme **unambiguous** ($\Phi$ injective);
- **Completeness** can be established by cardinality argument ($|\Phi(\mathcal{D}_n)| = s_n$)



# Combinatorics help in the design of DP schemes

**Reminder:** Generating function of secondary structures [Waterman, 1978]

$$S(z) := \sum_{n \geq 0} s_n \, z^n = \frac{1 - z + z^2 - \sqrt{1 - 2z - z^2 - 2z^3 + z^4}}{2z^2}$$

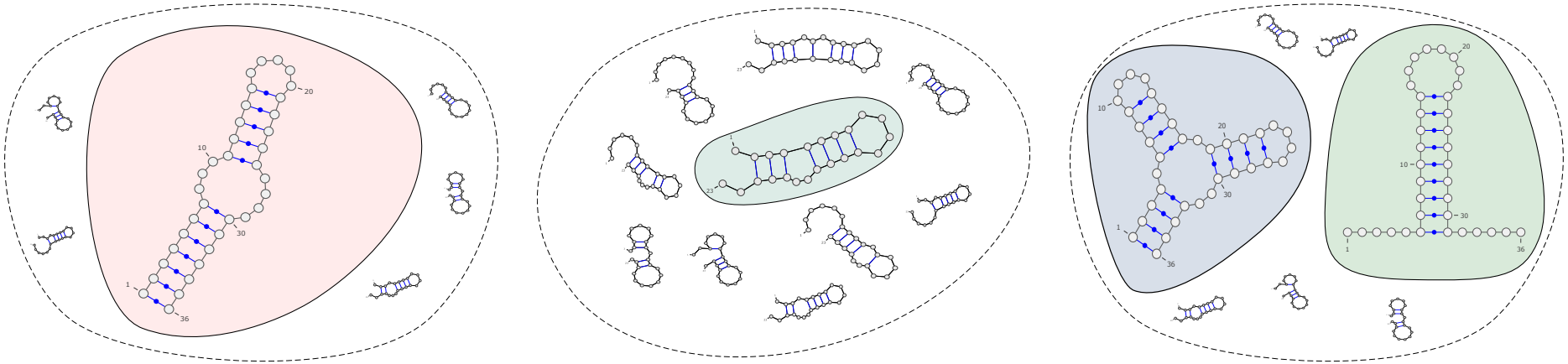
- DP scheme **unambiguous** ($\Phi$ injective);
- **Completeness** can be established by cardinality argument ($|\Phi(\mathcal{D}_n)| = s_n$)

$$A(z) = \begin{cases} \mathrm{Seq}(z) \\ z^2 \, A(z) \\ z \, \mathrm{Seq}(z) \, z^2 \, A(z) + z^2 \, A(z) \, \mathrm{Seq}(z) \, z \\ \quad + z \, \mathrm{Seq}(z) \, z^2 \, A(z) \, \mathrm{Seq}(z) z \\ B(z) \, C(z) \end{cases} \qquad B(z) = \begin{cases} B(z) \, C(z) \\ \mathrm{Seq}(z) \, B(z) \end{cases} \qquad C(z) = \begin{cases} C(z) \, z \\ z^2 \, A(z) \end{cases}$$

$$\mathrm{Seq}(z) = 1 + z \, \mathrm{Seq}(z)$$

$$\begin{aligned} A(z) &= \frac{1 - z - z^2 - \sqrt{1 - 2z - z^2 - 2z^3 + z^4}}{2z^2} \\ &= W(z) - 1 \quad \text{(OMG! The } \textit{empty} \text{ secondary structure is missing...)} \end{aligned}$$

# Motivation



**Observation:** **Large diversity of pictures** for the Boltzmann ensemble, associated with specific **functions**.

How to extract **functional information** from the Boltzmann ensemble?

**Idea:** Observe the distribution of additive structural parameters.

(Examples: #BPs, Free-energy, #Multiloops. . . )

**However,** exact computation is costly + Mean, Variance. . . often sufficient.

. . . How to efficiently compute the **moments** of the distribution?

# Distribution of discrete additive features

**Discrete feature function** $\alpha : \mathcal{D}_I \to [0, M]$, additively associated with derivations.

**Random variable** $A \in \mathcal{D}_I$**:** Uniformly/Boltzmann-distributed random derivation.

**Problem:** Compute **explicit** distribution of $V$, defined as:

$$\forall m \in [0, M] : \mathbb{P}(\alpha(A) = m \mid I) \equiv \sum_{\substack{d \in \mathcal{D}_I \\ \alpha(d) = m}} \frac{1}{|\mathcal{D}_I|} = \frac{\overbrace{|\{d \in \mathcal{D}_I \mid \alpha(d) = m\}|}^{V_q^m}}{|\mathcal{D}_I|}$$

**Naive solution:** Explicit convolution products

$$\forall q \in Q, \quad V_q = \sum \begin{cases} V_{q_{1,1}} \times V_{q_{1,2}} \times \cdots \\ V_{q_{2,1}} \times V_{q_{2,2}} \times \cdots \\ \vdots \\ V_{q_{k,1}} \times V_{q_{k,2}} \times \cdots \end{cases}$$

# Distribution of discrete additive features

**Discrete feature function** $\alpha : \mathcal{D}_I \to [0, M]$, additively associated with derivations.

**Random variable** $A \in \mathcal{D}_I$**:** Uniformly/Boltzmann-distributed random derivation.

**Problem:** Compute **explicit** distribution of $V$, defined as:

$$\forall m \in [0, M] : \mathbb{P}(\alpha(A) = m \mid I) \equiv \sum_{\substack{d \in \mathcal{D}_I \\ \alpha(d) = m}} \frac{1}{|\mathcal{D}_I|} = \frac{\overbrace{|\{d \in \mathcal{D}_I \mid \alpha(d) = m\}|}^{V_q^m}}{|\mathcal{D}_I|}$$

**Naive solution:** Explicit convolution products $\to$ Time: $\mathcal{O}(M^2\, k\, |Q|)$/Mem.: $\Theta(M\, |Q|)$

$$\forall m \in [0, M], \forall q \in Q, \quad V_q^m = \sum \begin{cases} \displaystyle\sum_{m_1 + m_2 + \ldots = m - \delta_{q,1}} V_{q_{1,1}}^{m_1} \times V_{q_{1,2}}^{m_2} \times \cdots \\ \displaystyle\sum_{m_1 + m_2 + \ldots = m - \delta_{q,2}} V_{q_{2,1}}^{m_2} \times V_{q_{2,2}}^{m_2} \times \cdots \\ \vdots \\ \displaystyle\sum_{m_1 + m_2 + \ldots = m - \delta_{q,k}} V_{q_{k,1}}^{m_2} \times V_{q_{k,2}}^{m_2} \times \cdots \end{cases}$$

# Distribution of discrete additive features

**Discrete feature function** $\alpha : \mathcal{D}_I \to [0, M]$, additively associated with derivations.

**Random variable** $A \in \mathcal{D}_I$**:** Uniformly/Boltzmann-distributed random derivation.

**Problem:** Compute **explicit** distribution of $V$, defined as:

$V_q^m$

$$\forall m \in [0, M] : \mathbb{P}(\alpha(A) = m \mid I) \equiv \sum_{\substack{d \in \mathcal{D}_I \\ \alpha(d) = m}} \frac{1}{|\mathcal{D}_I|} = \frac{|\{d \in \mathcal{D}_I \mid \alpha(d) = m\}|}{|\mathcal{D}_I|}$$

**Naive solution:** Explicit convolution products $\to$ Time: $\mathcal{O}(M^2\, k\, |Q|)$/Mem.: $\Theta(M\, |Q|)$

**Interpolation** [Waldispühl and Ponty, 2011] **:**

- Consider polynomials $V_q(z) = \sum_{m=0}^{M} V_q^m \cdot z^m$;
- Evaluation of $V_q(z)$ possible in $\Theta(k\, |Q|)/\Theta(|Q|)$ for any given $z \in \mathbb{R}^+$;

$$\forall q \in Q, \quad V_q(z) = \sum \begin{cases} z^{\alpha(\delta q, 1)} \times V_{q_{1,1}}(z) \times V_{q_{1,2}}(z) \times \cdots \\ \vdots \\ z^{\alpha(\delta q, k)} \times V_{q_{k,1}}(z) \times V_{q_{k,2}}(z) \times \cdots \end{cases}$$

- Compute $V_{q_0}(z)$ on $M + 1$ distinct values $(z_1, z_2, \cdots, z_{M+1})$;
- **Interpolate** coeff. $V_q^m$ $\to$ DFT [Senter, Sheikh, Dotu, Ponty, and Clote, 2012]: $\Theta(M \log(M))$.

# Distribution of discrete additive features

**Discrete feature function** $\alpha : \mathcal{D}_I \to [0, M]$, additively associated with derivations.

**Random variable** $A \in \mathcal{D}_I$**:** Uniformly/Boltzmann-distributed random derivation.

**Problem:** Compute **explicit** distribution of $V$, defined as:

$$\forall m \in [0, M] : \mathbb{P}(\alpha(A) = m \mid I) \equiv \sum_{\substack{d \in \mathcal{D}_I \\ \alpha(d) = m}} \frac{1}{|\mathcal{D}_I|} = \frac{\overbrace{|\{d \in \mathcal{D}_I \mid \alpha(d) = m\}|}^{V_q^m}}{|\mathcal{D}_I|}$$

**Naive solution:** Explicit convolution products $\to$ Time: $\mathcal{O}(M^2\, k\, |Q|)$/Mem.: $\Theta(M\, |Q|)$

**Interpolation** [Waldispühl and Ponty, 2011] **:** $\to$ Time: $\mathcal{O}(M\, k\, |Q|)$/Mem.: $\Theta(|Q|)$

- Consider polynomials $V_q(z) = \sum_{m=0}^{M} V_q^m \cdot z^m$;
- Evaluation of $V_q(z)$ possible in $\Theta(k\, |Q|)/\Theta(|Q|)$ for any given $z \in \mathbb{R}^+$;
- Compute $V_{q_0}(z)$ on $M + 1$ distinct values $(z_1, z_2, \cdots, z_{M+1})$;
- **Interpolate** coeff. $V_q^m$ $\to$ DFT [Senter, Sheikh, Dotu, Ponty, and Clote, 2012]: $\Theta(M \log(M))$.

# Computing the moments of additive features

**Discrete feature function** $\alpha : \mathcal{D}_I \to [0, M]$, additively associated with derivations.

**Random variable** $A \in \mathcal{D}_I$**:** Uniformly/Boltzmann-distributed random derivation.

**Problem:** Given an instance $I$, compute $1^{\text{st}}, \ldots, p^{\text{th}}$ **moment** of $A$ :

$$\mathbb{E}(\alpha(A)^p \mid I) = \sum_{d \in \mathcal{D}_I} \mathbb{P}(d \mid I) \cdot \alpha(d) = \frac{\boxed{\sum_{d \in \mathcal{D}_I} \alpha(d)^p}}{|\mathcal{D}_I|} \longrightarrow \mathcal{X}_{i,j}$$

# Computing the moments of additive features

**Discrete feature function** $\alpha : \mathcal{D}_I \to [0, M]$, additively associated with derivations.

**Random variable** $A \in \mathcal{D}_I$**:** Uniformly/Boltzmann-distributed random derivation.

**Problem:** Given an instance $I$, compute $1^{\text{st}}, \ldots, p^{\text{th}}$ **moment** of $A$ :

$$\mathbb{E}(\alpha(A)^p \mid I) = \sum_{d \in \mathcal{D}_I} \mathbb{P}(d \mid I) \cdot \alpha(d) = \frac{\boxed{\sum_{d \in \mathcal{D}_I} \alpha(d)^p}}{|\mathcal{D}_I|} \longrightarrow \mathcal{X}_{i,j}$$

**Why?**

- $1^{\text{st}}$ **moment:** Average free-energy, # base-pairs $\mu := \mathbb{E}(A)$
- $2^{\text{nd}}$ **moment:** Variance/standard dev., correlations...

# Computing the moments of additive features

**Discrete feature function** $\alpha : \mathcal{D}_I \to [0, M]$, additively associated with derivations.

**Random variable** $A \in \mathcal{D}_I$**:** Uniformly/Boltzmann-distributed random derivation.

**Problem:** Given an instance $I$, compute $1^{\text{st}}, \ldots, p^{\text{th}}$ **moment** of $A$ :

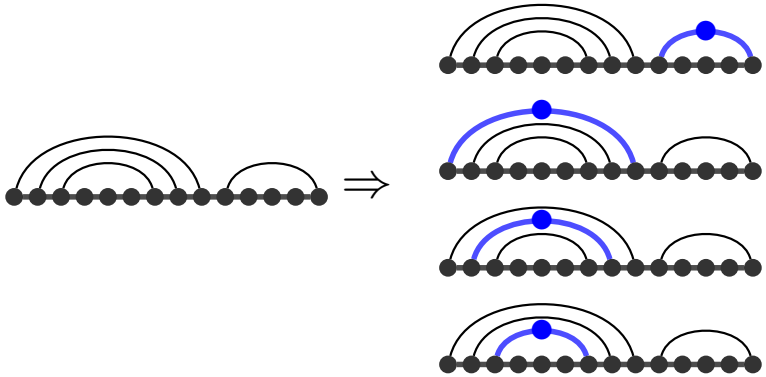
$$\mathbb{E}(\alpha(A)^p \mid I) = \sum_{d \in \mathcal{D}_I} \mathbb{P}(d \mid I) \cdot \alpha(d) = \frac{\boxed{\sum_{d \in \mathcal{D}_I} \alpha(d)^p}}{|\mathcal{D}_I|} \longrightarrow \mathcal{X}_{i,j}$$

**Why?**

- $1^{\text{st}}$ **moment:** Average free-energy, # base-pairs $\mu := \mathbb{E}(A)$
- $2^{\text{nd}}$ **moment:** Variance/standard dev., correlations...

**Pointing derivations (formal derivative)** [Ponty and Saule, 2011]**:** $\to \Theta(2^p\, k\, |Q|)/\Theta(p\, |Q|)$

- **Transform equation** to generate derivations **pointed** on #RHS→LHS transitions.
- **Weight** each pointed transition with contribution to $\alpha$.



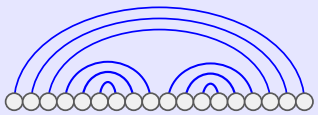
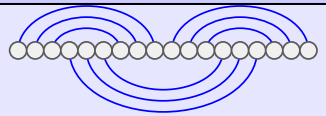
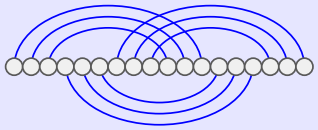
Counting in this decomposition
$\Leftrightarrow$ Compute $\mathcal{X}_{i,j}$

# Conclusion

# We need your help!


- **Crossing interactions (pseudoknots):** Finding the right parameter
- **Kinetics:** Markov process. . . computing energy barrier is hard! [Thachuk et al., 2010]
- **RNA Inverse folding/Design:** Complexity unknown, largely open!
- Constructing combinatorial DP scheme for classic problems

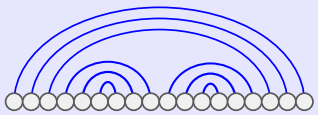
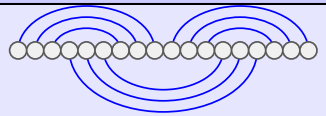
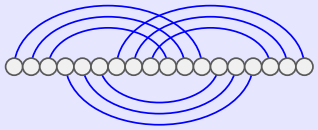
# Folding including crossing interactions

| | | Base-pairs | Stacking-Pairs | Nearest-Neighbor |
|---|---|---|---|---|
|  | Comp. | P [Nussinov and Jacobson, 1980] | P [Ieong et al., 2003] | P [Zuker and Stiegler, 1981] |
| Non-crossing | Approx. | – | – | – |
|  | Comp. | ??? | NP-Hard [Ieong et al., 2003] | NP-Hard [Ieong et al., 2003] |
| Planar | Approx. | 2-approx. ≈[Ieong et al., 2003] | 2-approx. [Ieong et al., 2003] | ??? |
|  | Comp. | P [Tabaska et al., 1998] | NP-Hard [Lyngsø, 2004; Sheikh, Backofen, and Ponty, 2012] (any* Δ model) | NP-Hard [Lyngsø and Pedersen, 2000; Akutsu, 2000] |
| General | Approx. | – | $\varepsilon$-approx. $\in \mathcal{O}(n^{4^{1/\varepsilon}})$ [Lyngsø, 2004; Sheikh, Backofen, and Ponty, 2012] 1/5 (any Δ model) | APX-Hard [Sheikh, Backofen, and Ponty, 2012] |

**Idea:** Parameterized complexity approach.

- **Bounded map genus** → Avoid **finite** set of substructures! [Vernizzi et al., 2006; Reidys et al., 2011]
- **Bounded tree-width** → Promising. . . but what geometric relevance? [Ding et al., 2014]
- **Bounded page number** → Already hard for two pages [Clote et al., 2012a]
- **Bounded wave number** [Akutsu, 2000]

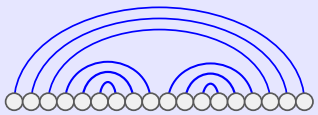
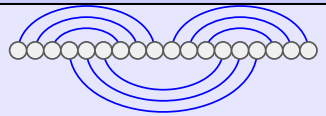
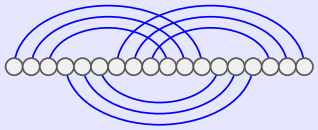
# Folding including crossing interactions

| | | Base-pairs | Stacking-Pairs | Nearest-Neighbor |
|---|---|---|---|---|
|  | Comp. | P [Nussinov and Jacobson, 1980] | P [Ieong et al., 2003] | P [Zuker and Stiegler, 1981] |
| Non-crossing | Approx. | – | – | – |
|  | Comp. | ??? | NP-Hard [Ieong et al., 2003] | NP-Hard [Ieong et al., 2003] |
| Planar | Approx. | 2-approx. ≈[Ieong et al., 2003] | 2-approx. [Ieong et al., 2003] | ??? |
|  | Comp. | P [Tabaska et al., 1998] | NP-Hard [Lyngsø, 2004; Sheikh, Backofen, and Ponty, 2012] (any* Δ model) | NP-Hard [Lyngsø and Pedersen, 2000; Akutsu, 2000] |
| General | Approx. | – | $\varepsilon$-approx. $\in \mathcal{O}(n^{4^{1/\varepsilon}})$ [Lyngsø, 2004; Sheikh, Backofen, and Ponty, 2012] 1/5 (any Δ model) | APX-Hard [Sheikh, Backofen, and Ponty, 2012] |

**Idea:** Parameterized complexity approach.

- **Bounded map genus** → Avoid **finite** set of substructures! [Vernizzi et al., 2006; Reidys et al., 2011]
- **Bounded tree-width** → Promising. . . but what geometric relevance? [Ding et al., 2014]
- **Bounded page number** → Already hard for two pages [Clote et al., 2012a]
- **Bounded wave number** [Akutsu, 2000]

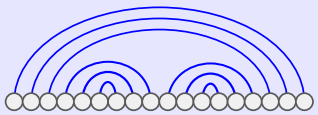
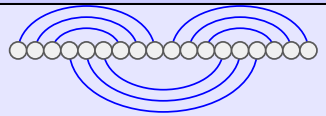
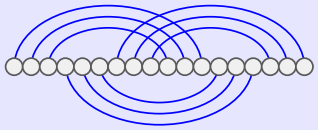
# Folding including crossing interactions

| | | Base-pairs | Stacking-Pairs | Nearest-Neighbor |
|---|---|---|---|---|
|  | Comp. | P [Nussinov and Jacobson, 1980] | P [Ieong et al., 2003] | P [Zuker and Stiegler, 1981] |
| Non-crossing | Approx. | – | – | – |
|  | Comp. | ??? | NP-Hard [Ieong et al., 2003] | NP-Hard [Ieong et al., 2003] |
| Planar | Approx. | 2-approx. ≈[Ieong et al., 2003] | 2-approx. [Ieong et al., 2003] | ??? |
|  | Comp. | P [Tabaska et al., 1998] | NP-Hard [Lyngsø, 2004; Sheikh, Backofen, and Ponty, 2012] (any* Δ model) | NP-Hard [Lyngsø and Pedersen, 2000; Akutsu, 2000] |
| General | Approx. | – | $\varepsilon$-approx. $\in \mathcal{O}(n^{4^{1/\varepsilon}})$ [Lyngsø, 2004; Sheikh, Backofen, and Ponty, 2012] 1/5 (any Δ model) | APX-Hard [Sheikh, Backofen, and Ponty, 2012] |

**Idea:** Parameterized complexity approach.

- **Bounded map genus** → Avoid **finite** set of substructures! [Vernizzi et al., 2006; Reidys et al., 2011]
- **Bounded tree-width** → Promising. . . but what geometric relevance? [Ding et al., 2014]
- **Bounded page number** → Already hard for two pages [Clote et al., 2012a]
- **Bounded wave number** [Akutsu, 2000]

# Folding including crossing interactions

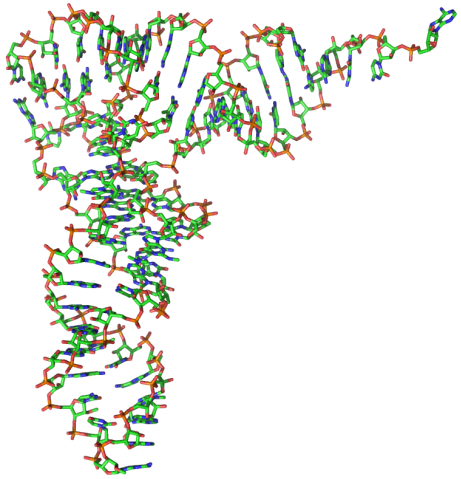
| | | Base-pairs | Stacking-Pairs | Nearest-Neighbor |
|---|---|---|---|---|
|  | Comp. | P [Nussinov and Jacobson, 1980] | P [Ieong et al., 2003] | P [Zuker and Stiegler, 1981] |
| Non-crossing | Approx. | – | – | – |
|  | Comp. | ??? | NP-Hard [Ieong et al., 2003] | NP-Hard [Ieong et al., 2003] |
| Planar | Approx. | 2-approx. ≈[Ieong et al., 2003] | 2-approx. [Ieong et al., 2003] | ??? |
|  | Comp. | P [Tabaska et al., 1998] | NP-Hard [Lyngsø, 2004; Sheikh, Backofen, and Ponty, 2012] (any* Δ model) | NP-Hard [Lyngsø and Pedersen, 2000; Akutsu, 2000] |
| General | Approx. | – | $\varepsilon$-approx. $\in \mathcal{O}(n^{4^{1/\varepsilon}})$ [Lyngsø, 2004; Sheikh, Backofen, and Ponty, 2012] 1/5 (any Δ model) | APX-Hard [Sheikh, Backofen, and Ponty, 2012] |

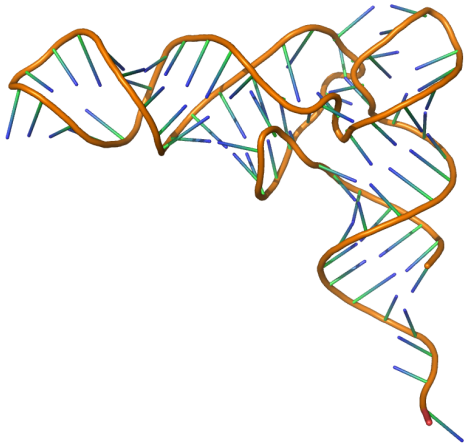
**Idea:** Parameterized complexity approach.

- **Bounded map genus** → Avoid **finite** set of substructures! [Vernizzi et al., 2006; Reidys et al., 2011]
- **Bounded tree-width** → Promising. . . but what geometric relevance? [Ding et al., 2014]
- **Bounded page number** → Already hard for two pages [Clote et al., 2012a]
- **Bounded wave number** [Akutsu, 2000]

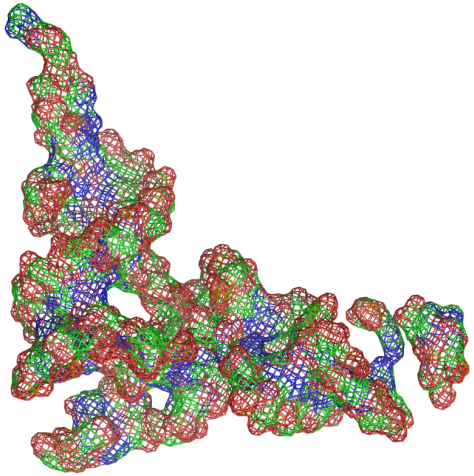
# We need your help!

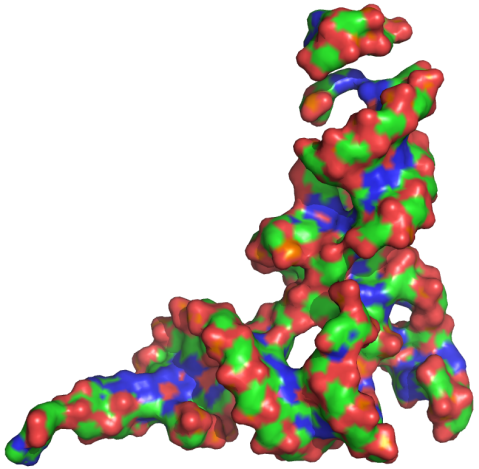

- **Crossing interactions (pseudoknots):** Finding the right parameter
- **Kinetics:** Markov process. . . computing energy barrier is hard! [Thachuk et al., 2010]
- **RNA Inverse folding/Design:** Complexity unknown, largely open!
- Constructing combinatorial DP scheme for classic problems





# Thank you for your attention





# References I


Tatsuya Akutsu. Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discrete Appl. Math.*, 104(1-3):45–62, 2000. ISSN 0166-218X. doi: 10.1016/S0166-218X(00)00186-4.

Peter Clote, Stefan Dobrev, Ivan Dotu, Evangelos Kranakis, Danny Krizanc, and Jorge Urrutia. On the page number of rna secondary structures with pseudoknots. *J Math Biol*, 65(6-7):1337–1357, Dec 2012a. doi: 10.1007/s00285-011-0493-6. URL http://dx.doi.org/10.1007/s00285-011-0493-6.

Peter Clote, Yann Ponty, and Jean-Marc Steyaert. Expected distance between terminal nucleotides of rna secondary structures. *Journal of Mathematical Biology*, 65(3):581–599, 2012b.

Liang Ding, Xingran Xue, Sal Lamarca, Mohammad Mohebbi, Abdul Samad, Russell L. Malmberg, and Liming Cai. Ab initio prediction of rna nucleotide interactions with backbone *k*-tree model. In Fabrice Jossinet, Yann Ponty, and Jérôme Waldispühl, editors, *Proceedings of 1st Workshop on Computational Methods for Structural RNAs (CMSR'14)*, volume 1, pages 25–42, Strasbourg, France, September 2014. doi: 10.15455/CMSR.2014.0003. URL http://dx.doi.org/10.15455/CMSR.2014.0003.

Y. Ding and E. Lawrence. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Research*, 31(24):7280–7301, 2003.

Chuong B. Do, Daniel A. Woods, and Serafim Batzoglou. Contrafold: Rna secondary structure prediction without physics-based models. *Bioinformatics*, 22 (14):e90–e98, Jul 2006. doi: 10.1093/bioinformatics/btl246. URL http://dx.doi.org/10.1093/bioinformatics/btl246.

Philippe Flajolet and Robert Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2009.

E. Freyhult, V. Moulton, and P .Clote. Boltzmann probability of RNA structural neighbors and riboswitch detection. *Bioinformatics*, 23(16):2054–2062, 2007.

Samuel Ieong, Ming yang Kao, Tak wah Lam, Wing kin Sung, and Siu ming Yiu. Predicting rna secondary structures with arbitrary pseudoknots by maximizing the number of stacking pairs. *Journal Of Computational Biology*, 10(6):981–995, 2003.

R. B. Lyngsø and C. N. S. Pedersen. RNA pseudoknot prediction in energy-based models. *Journal of Computational Biology*, 7(3-4):409–427, 2000.

Rune Lyngsø. Complexity of pseudoknot prediction in simple models. In *Proceedings of ICALP*, 2004.

J.S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29:1105–1119, 1990.

István Miklós, Irmtraud M. Meyer, and Borbála Nagy. Moments of the boltzmann distribution for RNA secondary structures. *Bull Math Biol*, 67(5):1031–1047, Sep 2005. doi: 10.1016/j.bulm.2004.12.003. URL http://dx.doi.org/10.1016/j.bulm.2004.12.003.

R. Nussinov and A.B. Jacobson. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc Natl Acad Sci U S A*, 77:6903–13, 1980.

Yann Ponty. Efficient sampling of RNA secondary structures from the boltzmann ensemble of low-energy: The boustrophedon method. *Journal of Mathematical Biology*, 56(1-2):107–127, Jan 2008.

Yann Ponty and Cédric Saule. A combinatorial framework for designing (pseudoknotted) RNA algorithms. In Teresa M. Przytycka and Marie-France Sagot, editors, *Algorithms in Bioinformatics - 11th International Workshop, WABI 2011, Saarbrücken, Germany, September 5-7, 2011. Proceedings*, volume 6833 of *Lecture Notes in Computer Science*, pages 250–269. Springer, 2011. doi: 10.1007/978-3-642-23038-7_22. URL http://dx.doi.org/10.1007/978-3-642-23038-7_22.

# References II


Christian M. Reidys, Fenix W D. Huang, Jørgen E. Andersen, Robert C. Penner, Peter F. Stadler, and Markus E. Nebel. Topology and prediction of rna pseudoknots. *Bioinformatics*, 27(8):1076–1085, Apr 2011. doi: 10.1093/bioinformatics/btr090. URL http://dx.doi.org/10.1093/bioinformatics/btr090.

E. Rivas and S.R. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *Journal of Molecular Biology*, 285: 2053–2068, 1999.

David Sankoff. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM Journal on Applied Mathematics*, 45(5):pp. 810–825, 1985. ISSN 00361399. URL http://www.jstor.org/stable/2101630.

Evan Senter, Saad Sheikh, Ivan Dotu, Yann Ponty, and Peter Clote. Using the fast fourier transform to accelerate the computational search for rna conformational switches. *PLoS One*, 7(12):e50506, 2012. doi: 10.1371/journal.pone.0050506. URL http://dx.doi.org/10.1371/journal.pone.0050506.

Lioudmila V. Sharova, Alexei A. Sharov, Timur Nedorezov, Yulan Piao, Nabeebi Shaik, and Minoru S H. Ko. Database for mrna half-life of 19 977 genes obtained by dna microarray analysis of pluripotent and differentiating mouse embryonic stem cells. *DNA Res*, 16(1):45–58, Feb 2009. doi: 10.1093/dnares/dsn030. URL http://dx.doi.org/10.1093/dnares/dsn030.

Saad Sheikh, Rolf Backofen, and Yann Ponty. Impact of the energy model on the complexity of rna folding with pseudoknots. In Juha Karkkainen and Jens Stoye, editors, *Combinatorial Pattern Matching*, volume 7354 of *Lecture Notes in Computer Science*, pages 321–333. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-31264-9. doi: 10.1007/978-3-642-31265-6_26. URL http://dx.doi.org/10.1007/978-3-642-31265-6_26.

J. E. Tabaska, R. B. Cary, H. N. Gabow, and G. D. Stormo. An RNA folding method capable of identifying pseudoknots and base triples. *Bioinformatics*, 14 (8):691–699, 1998.

Chris Thachuk, Ján Manuch, Arash Rafiey, Leigh-Anne Mathieson, Ladislav Stacho, and Anne Condon. An algorithm for the energy barrier problem without pseudoknots and temporary arcs. *Pac Symp Biocomput*, pages 108–119, 2010.

M. Vauchaussade de Chaumont and X.G. Viennot. Enumeration of RNA secondary structures by complexity. In V. Capasso, E. Grosso, and S.L. Paven-Fontana, editors, *Mathematics in Medecine and Biology*, volume 57 of *Lecture Notes in Biomathematics*, pages 360–365, 1985.

G. Vernizzi, P. Ribeca, H. Orland, and A. Zee. Topology of pseudoknotted homopolymers. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 73(3):031902, 2006.

Jérôme Waldispühl and Yann Ponty. An unbiased adaptive sampling algorithm for the exploration of rna mutational landscapes under evolutionary pressure. *J Comput Biol*, 18(11):1465–1479, Nov 2011. doi: 10.1089/cmb.2011.0181. URL http://dx.doi.org/10.1089/cmb.2011.0181.

M. S. Waterman. Secondary structure of single stranded nucleic acids. *Advances in Mathematics Supplementary Studies*, 1(1):167–212, 1978.

S. Wuchty, W. Fontana, I.L. Hofacker, and P. Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49: 145–164, 1999.

M. Zuker and D. Sankoff. Rna secondary structures and their prediction. *Bull Math Bio*, 46:591–621, 1984.

M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequencesusing thermodynamics and auxiliary information. *Nucleic Acids Res.*, 9:133–148, 1981.