Repliement de l'ARN

Un point de vue combinatoire

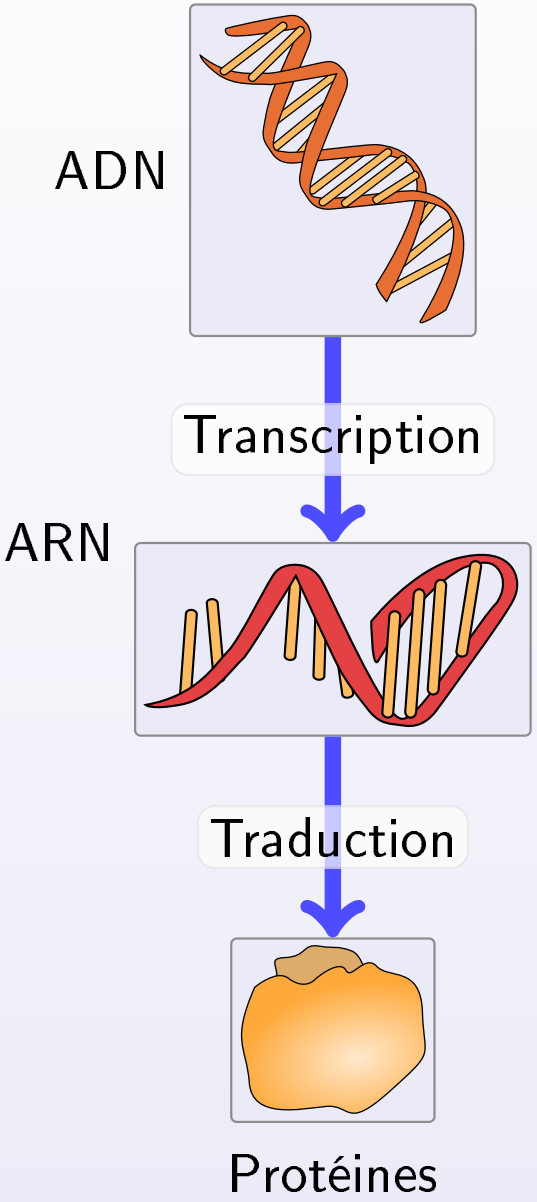
Yann Ponty

Équipe Bioinformatique LIX/AMIB INRIA
École Polytechnique/CNRS/INRIA – France

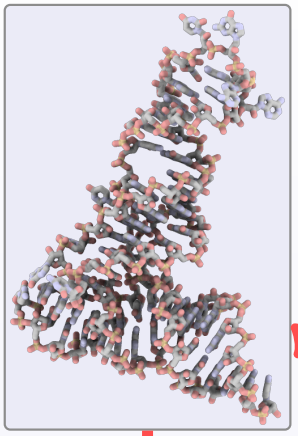
15 Décembre 2009

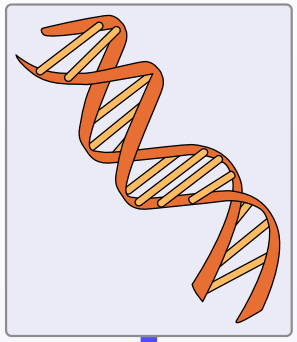
Plan

Dogme fondamental de la biologie moléculaire

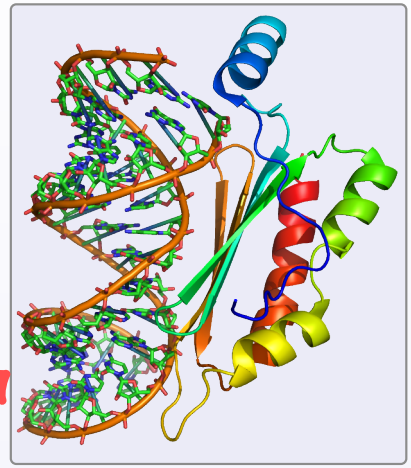


Dogme fondamental de la biologie moléculaire





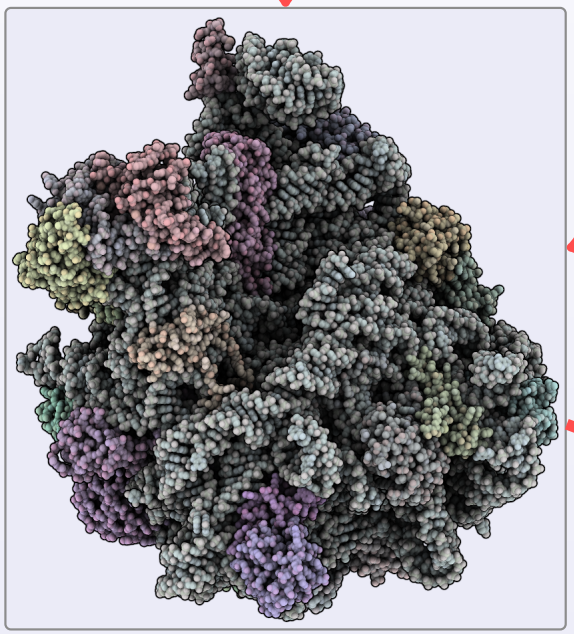
ADN



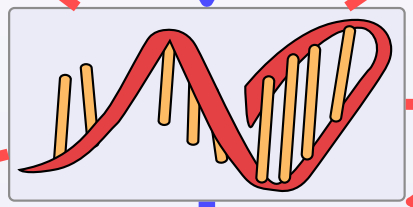
Transfert

Transcription

Maturation



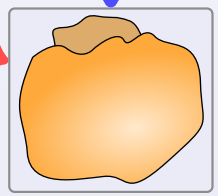
ARN



Régulation

Traduction

Synthèse



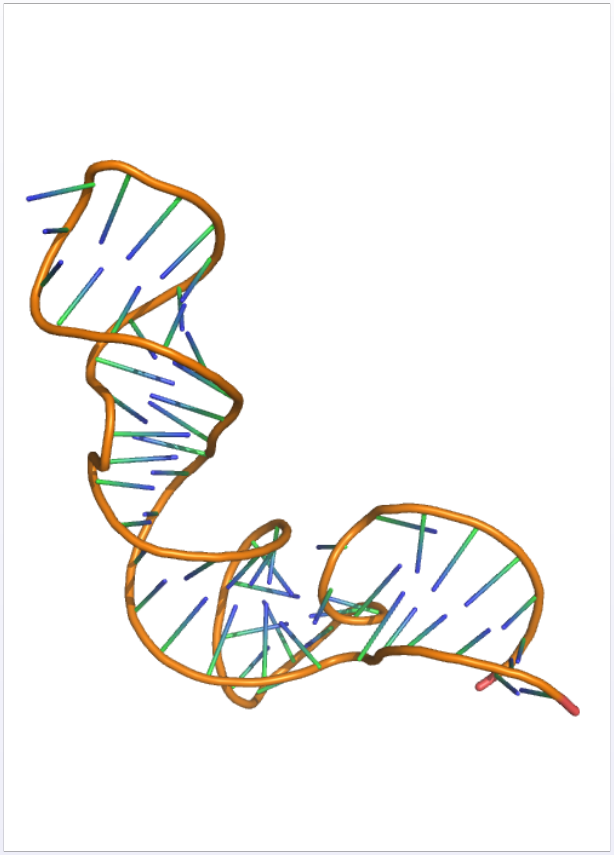
Protéines

Rôles de l'ARN :

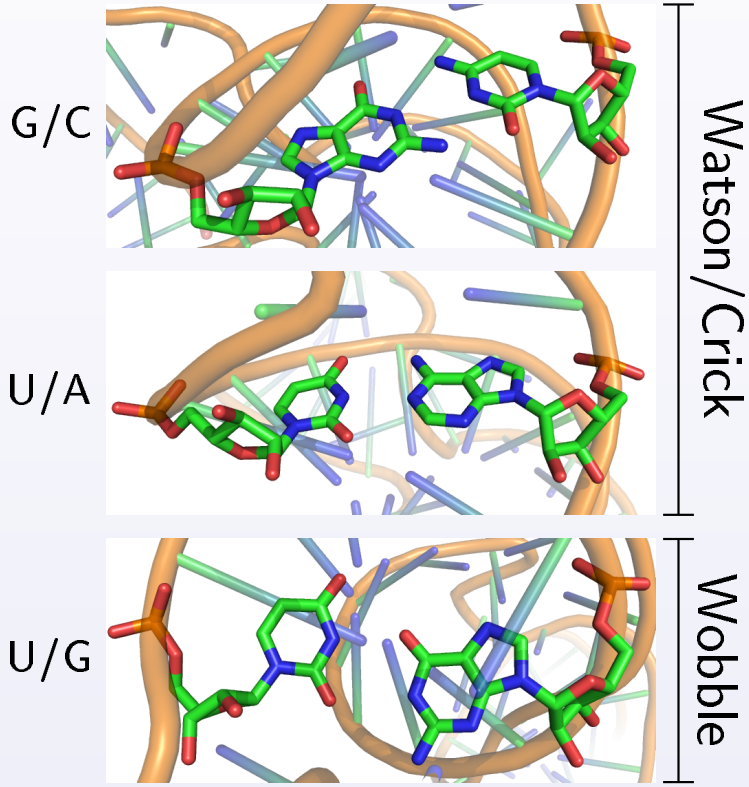
- Messager
- Machinerie traductionelle
- Régulation
- Modification
- Catalythique
- ...

Repliement de l'ARN

ARN = Polymère linéaire composé de nucléotides (A,C,G,U)



Appariements canoniques

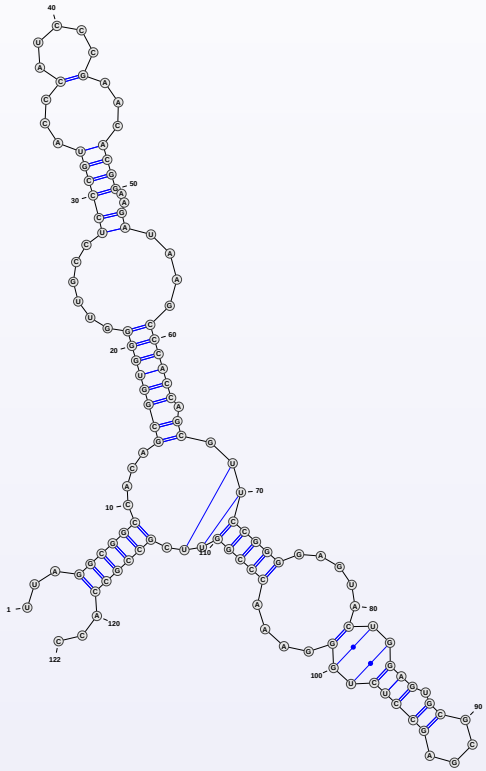


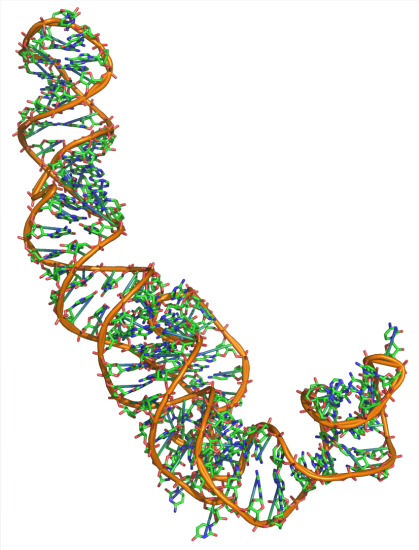
Repliement de l'ARN = Processus stochastique continu dirigé par (résultant en) un appariement des nucléotides.

Structure(s) de l'ARN

Trois niveaux de représentation[1] :

UUAGGCGGCCACAGC
GGUGGGGUUGCCUCC
CGUACCCAUCCCGAA
CACGGAAGAUAAGCC
CACCAGCGUUCCGGG
GAGUACUGGAGUGCG
CGAGCCUCUGGGAAA
CCCGGUUCGCCGCCA
CC





Source : 5s rRNA

(PDBID : 1K73 :B)

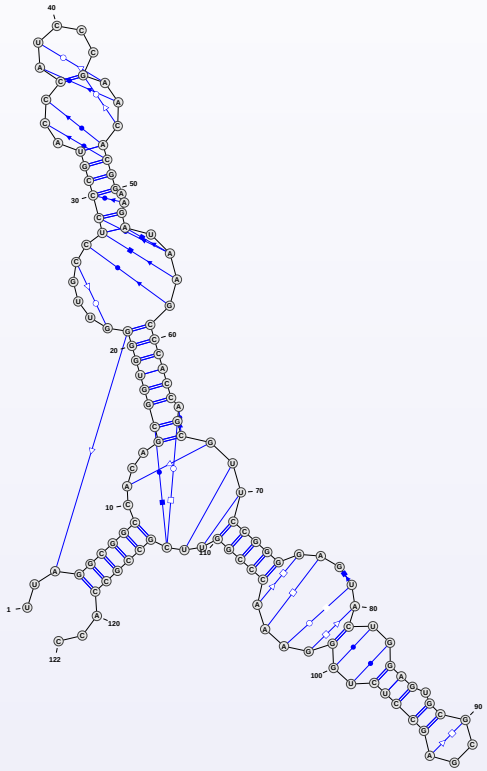
Structure primaire | Structure secondaire | Structure tertiaire

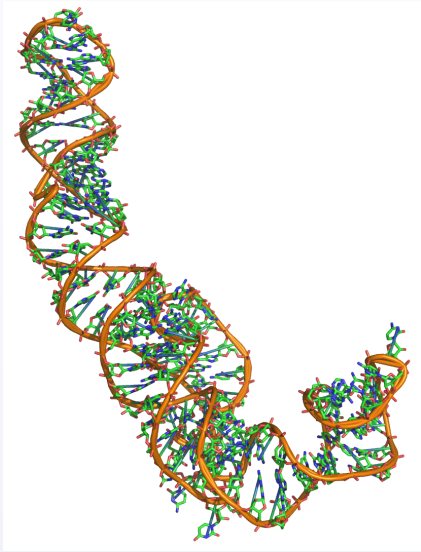
[1] Ou presque ...

Structure(s) de l'ARN

Trois niveaux de représentation[1] :

UUAGGCGGCCACAGC
GGUGGGGUUGCCUCC
CGUACCCAUCCCGAA
CACGGAAGAUAAGCC
CACCAGCGUUCCGGG
GAGUACUGGAGUGCG
CGAGCCUCUGGGAAA
CCCGGUUCGCCGCCA
CC





Source : 5s rRNA

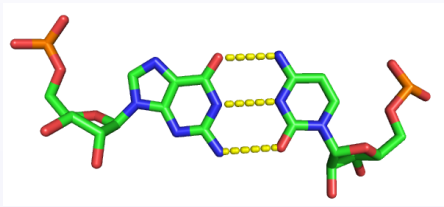
(PDBID : 1K73 :B)

Structure primaire | Structure secondaire+ | Structure tertiaire

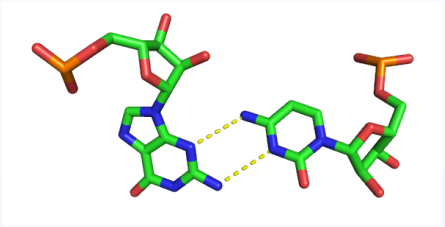
[1] Ou presque ...

Exclus dans la structure secondaire

- Appariements non-canoniques :
  Toute paire de base autre que {(A-U), (C-G), (G-U)}
  Ou interagissant sur un bord non-standard (WC/WC-Cis) [LW01].

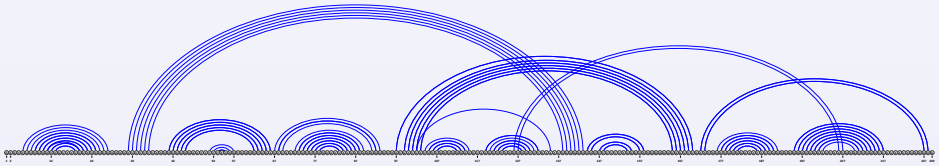


Paire CG canonique (WC/WC-Cis)



Paire CG non canonique (Sucre/WC-Trans)

- Pseudonoeuds :

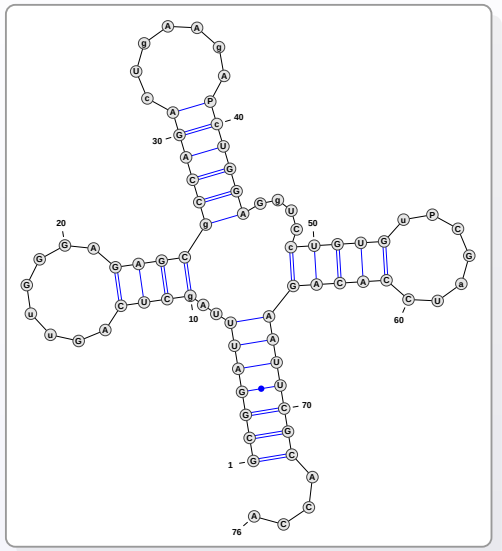


Structure pseudonoeud d'un Ribozyme du Groupe I (PDBID : 1Y0Q :A)

Plus expressif, mais repliement général *in silico* avec pseudonoeud :
$\Rightarrow$ NP-Complet [LP00] ...
... mais polynomial pour des classes restreintes [CDR+04].

Plan

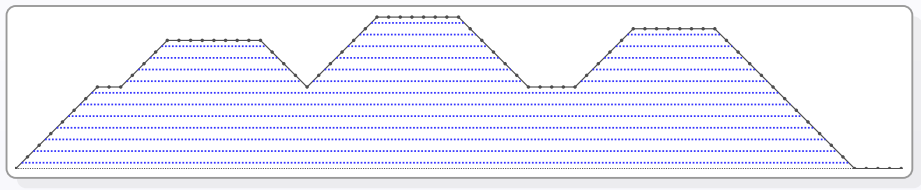
Diverses représentations d'une molécule omniprésente



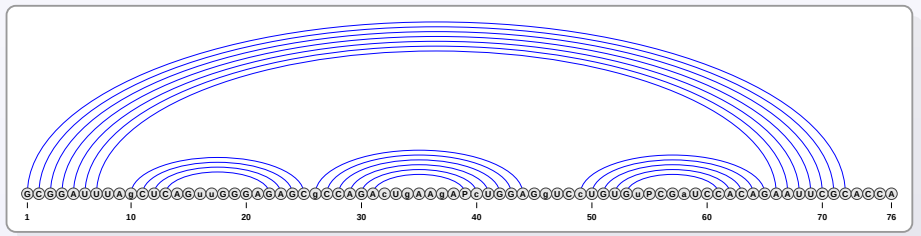
Graphe planaire (outer planar)

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

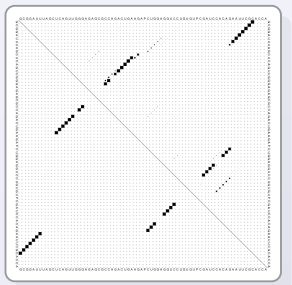
Expression bien parenthésée



Mountain view



Linéaire



Dot plot

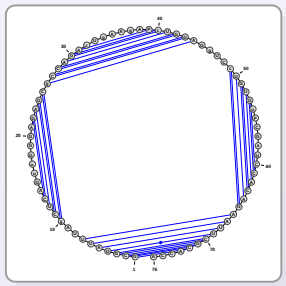


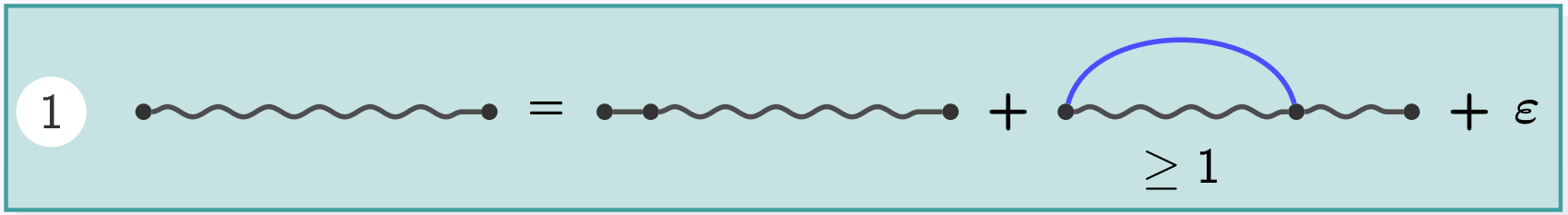
Diagramme de Feynman

Représentation différentes
mais
Structure combinatoire commune

Énumération des structures secondaires

Modèle homopolymère : Appariement possible pour tout couple de bases.

Waterman [Wat78] énumère les structures secondaires.

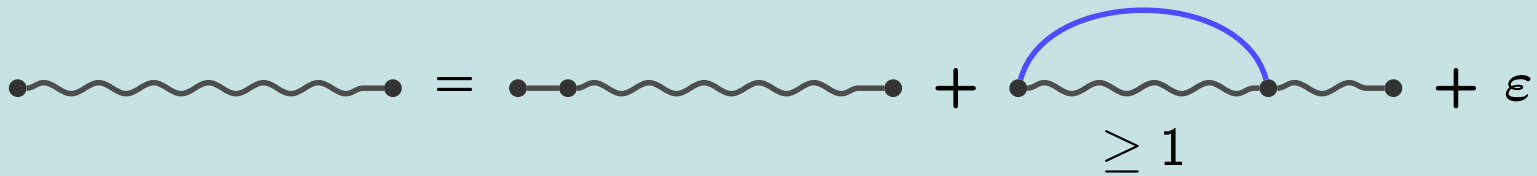


Énumération des structures secondaires

Modèle homopolymère : Appariement possible pour tout couple de bases.

Waterman [Wat78] énumère les structures secondaires.

1



2

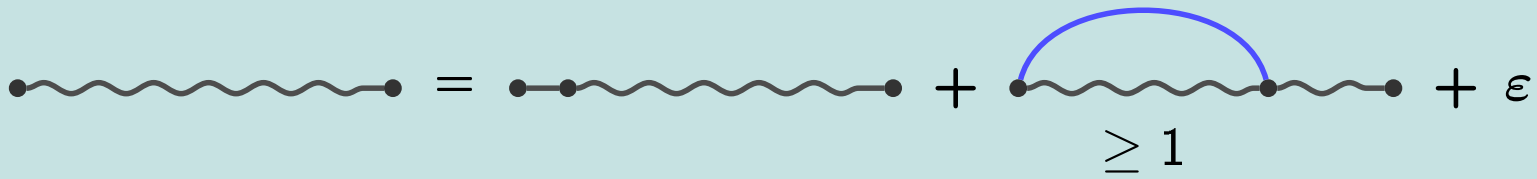
$$S \rightarrow \bullet\, S \mid (\, S_{>0}\,)\, S \mid \varepsilon$$

Énumération des structures secondaires

Modèle homopolymère : Appariement possible pour tout couple de bases.

Waterman [Wat78] énumère les structures secondaires.

1



2

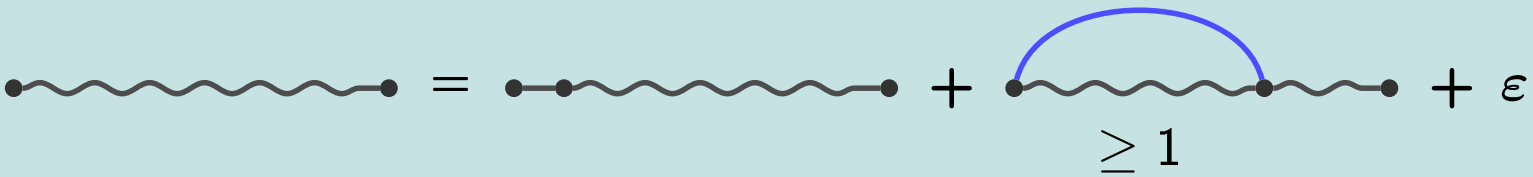
$$
\begin{aligned}
S &\rightarrow \bullet\, S \mid (\ \mathbf{T}\ )\, S \mid \varepsilon \\
\mathbf{T} &\rightarrow \bullet\, S \mid (\ \mathbf{T}\ )\, S
\end{aligned}
$$

Énumération des structures secondaires

Modèle homopolymère : Appariement possible pour tout couple de bases.

Waterman [Wat78] énumère les structures secondaires.

1



2

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

3

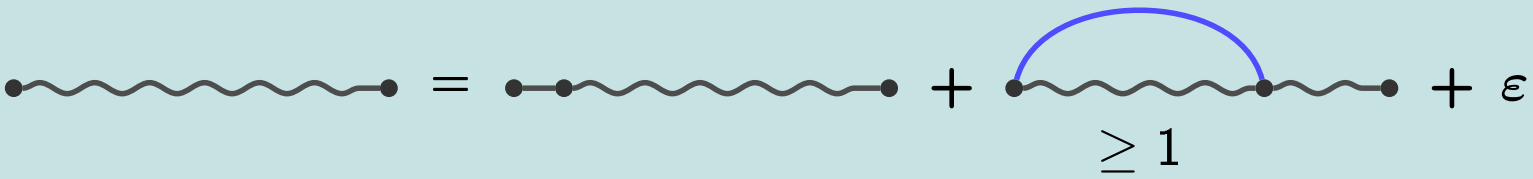
$$S(z) = \frac{1-z+z^2-\sqrt{1-2z-z^2-2z^3+z^4}}{2z^2}$$

Énumération des structures secondaires

Modèle homopolymère : Appariement possible pour tout couple de bases.

Waterman [Wat78] énumère les structures secondaires.

1



2

$$\begin{aligned} S &\to \bullet\, S \mid (\,\mathbf{T}\,)\, S \mid \varepsilon \\ \mathbf{T} &\to \bullet\, S \mid (\,\mathbf{T}\,)\, S \end{aligned}$$

3

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

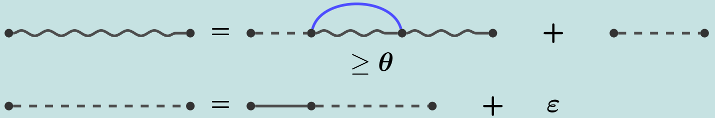
4

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

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

Contrainte stérique : Nombre min. de bases θ entre deux bases appariées.

1



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

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

4 $s_n \sim K.\frac{\beta^n}{n\sqrt{n}}(1+\mathcal{O}(1/n))$

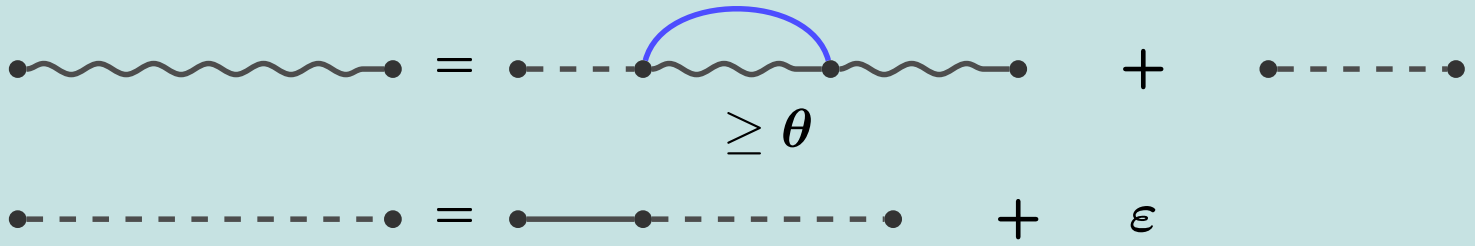
θ	0	1	3	10
β	3.	2.62	2.29	2.02

Empiriquement, $1.4^n/n\sqrt{n}$ struct. compatibles avec appariements [ZS84].

Énumération des structures secondaires

Contrainte stérique : Nombre min. de bases θ entre deux bases appariées.

1



2

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

3

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

4

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

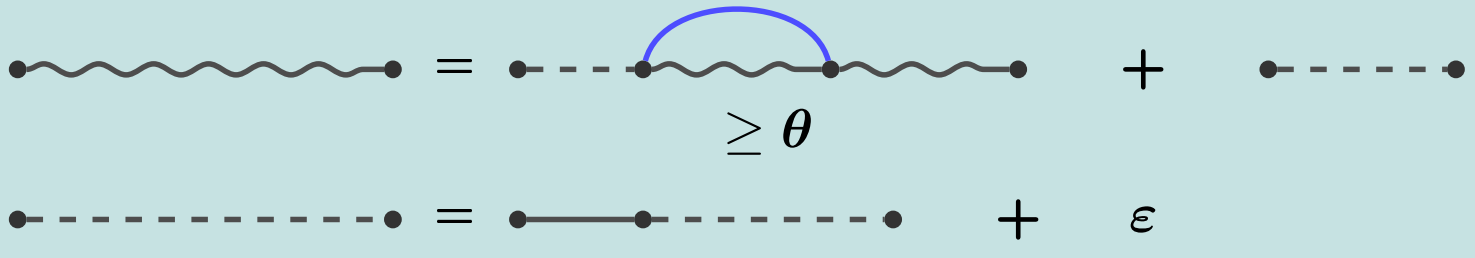
θ	0	1	3	10
β	3.	2.62	2.29	2.02

Empiriquement, $1.4^n/n\sqrt{n}$ struct. compatibles avec appariements [ZS84].

Énumération des structures secondaires

Contrainte stérique : Nombre min. de bases θ entre deux bases appariées.

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

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

4

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

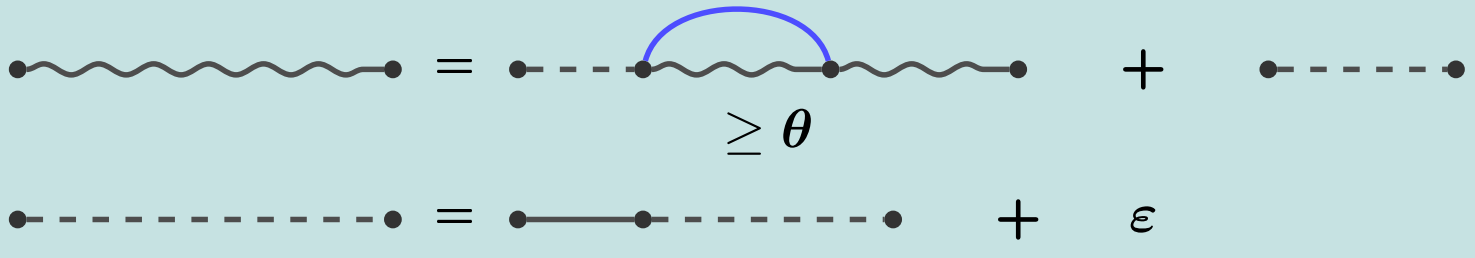
θ	0	1	3	10
β	3.	2.62	2.29	2.02

Empiriquement, $1.4^n/n\sqrt{n}$ struct. compatibles avec appariements [ZS84].

Énumération des structures secondaires

Contrainte stérique : Nombre min. de bases θ entre deux bases appariées.

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

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

4

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

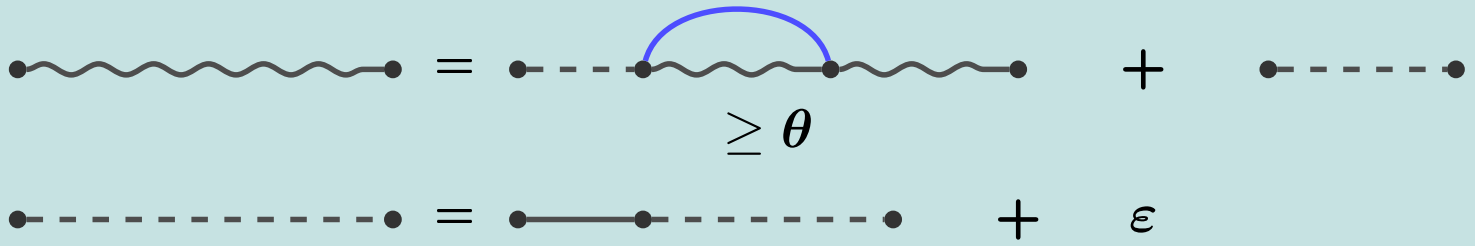
θ	0	1	3	10
β	3.	2.62	2.29	2.02

Empiriquement, $1.4^n/n\sqrt{n}$ struct. compatibles avec appariements [ZS84].

Énumération des structures secondaires

Contrainte stérique : Nombre min. de bases θ entre deux bases appariées.

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

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

4

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

θ	0	1	3	10
β	3.	2.62	2.29	2.02

Empiriquement, $1.4^n/n\sqrt{n}$ struct. compatibles avec appariements [ZS84].

1978–2004 : Paradigme énergie libre minimale

Paradigme énergie libre minimale (MFE)

Repliement fonctionnel d'un ARN = Structure d'énergie libre minimale

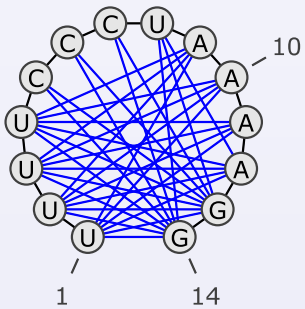
Modèle d'énergie de Nussinov/Jacobson : *Plus proche voisins* simple.

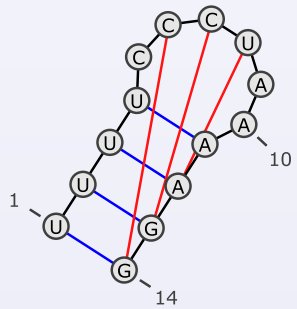
Énergie libre = $-$# Paires de bases

Repliement dans NJ $\Leftrightarrow$ Maximisation du nombre de paires de bases.
($\Leftrightarrow$ Stable max dans graphe circulaire $\Leftrightarrow$ CYK pondéré...)

Exemple :

UUUUCCCCUAAAAGG





Variante : Pondérer les paires selon leur nombre de liaisons hydrogènes
$\Delta G(G \equiv C) = -3$ $\Delta G(A = U) = -2$ $\Delta G(G - U) = -1$

Minimisation de l'énergie libre Nussinov/Jacobson

Récurrence sur l'énergie libre minimale $N_{1,n}$ d'un ARN :

$$N_{i,t} = 0, \quad \forall t \in [i, i+\theta]$$

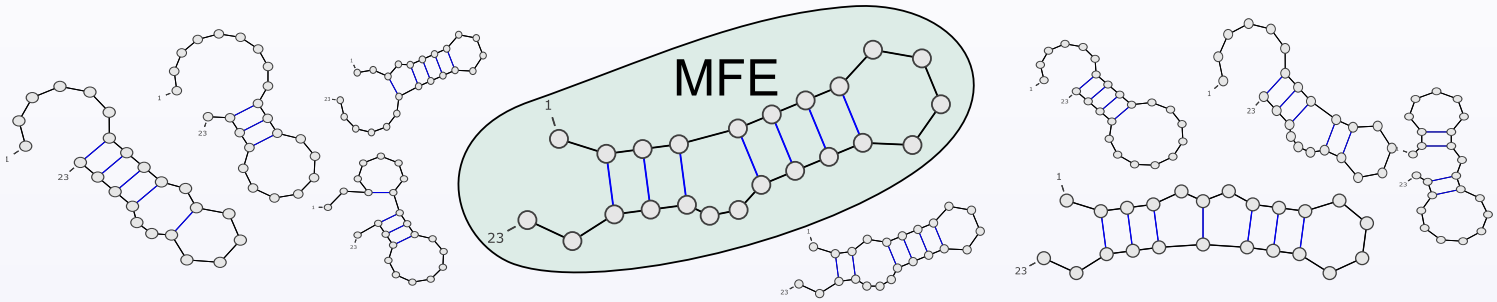
$$N_{i,j} = \min \begin{cases} N_{i+1,j} & (i \text{ non apparié}) \\ \min_{k=i+\theta+1}^{j} E_{i,j} + N_{i+1,k-1} + N_{k+1,j} & (i \text{ comp. avec } k) \end{cases}$$

À partir de N, les contributions aux min permettent de reconstruire récursivement le repliement d'énergie libre minimal (Backtrack).

$\Rightarrow$ Programmation dynamique pour le repliement [NJ80].

2004– : Paradigme ensemble de Boltzmann

L'ARN *respire* ⇒ Il n'existe pas UNE unique conformation fonctionnelle.



Paradigme ensemble de Boltzmann

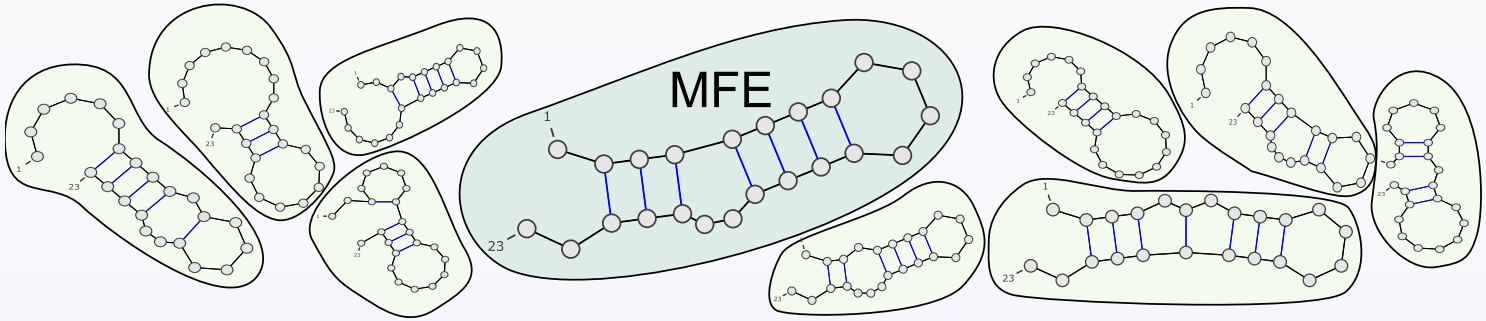
Les repliements d'un ARN co-existent dans un équilibre de Boltzmann.

Conséquence : La probabilité de la MFE peut être négligeable.

⇒ Les structures natives (fonctionnelles) sont alors à chercher dans les sous-optimales, où des structures très proches peuvent se *grouper*.

2004– : Paradigme ensemble de Boltzmann

L'ARN *respire* ⇒ Il n'existe pas UNE unique conformation fonctionnelle.



Paradigme ensemble de Boltzmann

Les repliements d'un ARN co-existent dans un équilibre de Boltzmann.

Conséquence : La probabilité de la MFE peut être négligeable.

⇒ Les structures natives (fonctionnelles) sont alors à chercher dans les sous-optimales, où des structures très proches peuvent se *grouper*.

Distribution de Boltzmann : Définitions

Une distribution de Bolzmann pondère chaque structure S d'un ARN ω par un facteur de Boltzmann $e^{\frac{-E_{S,\omega}}{RT}}$ où :

- $E_{S,\omega}$ est l'énergie libre de S (kCal.mol^{-1})
- T est la température (K)
- R est la constante des gaz parfaits ($1.986.10^{-3}$ kCal.K^{-1}.mol^{-1})

Distribution renormalisée sur S_ω par la fonction de partition

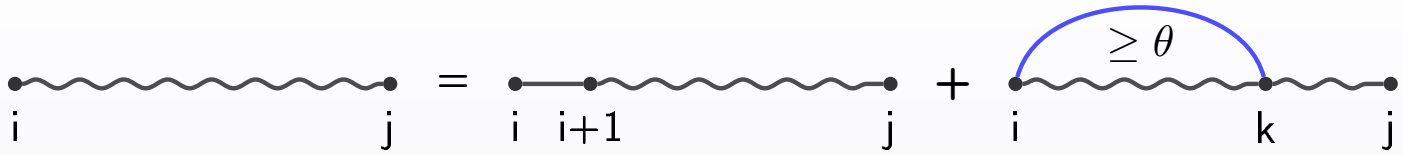
$$\mathcal{Z}_\omega = \sum_{S \in \mathcal{S}_\omega} e^{\frac{-E_{S,\omega}}{RT}}.$$

où $\mathcal{S}_\omega$ est l'ensemble des conformations compatibles avec ω.

La probabilité de Boltzmann d'une structure S est alors donnée par

$$P_{S,\omega} = \frac{e^{\frac{-E_{S,\omega}}{RT}}}{\mathcal{Z}_\omega}.$$

Fonction de partition



Récurrence sur l'énergie minimale d'un repliement :

$$N_{i,t} = 0, \quad \forall t \in [i, i+\theta]$$

$$N_{i,j} = \min \begin{cases} N_{i+1,j} & (i \text{ non apparié}) \\ \min_{k=i+\theta+1}^{j} E_{i,j} + N_{i+1,k-1} + N_{k+1,j} & (i \text{ comp. avec } k) \end{cases}$$

$\Rightarrow$ Récurrence sur la fonction de partition :

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

$$\mathcal{Z}_{i,j} = \sum \begin{cases} \mathcal{Z}_{i+1,j} & (i \text{ non apparié}) \\ \sum_{k=i+\theta+1}^{j} e^{\frac{1}{RT}} \times \mathcal{Z}_{i+1,k-1} \times \mathcal{Z}_{k+1,j} & (i \text{ comp. avec } k) \end{cases}$$

+ Quelques calculs un peu techniques ...

= Probabilité des paires de bases dans l'ensemble de Boltzmann [McC90].

Échantillonnage statistique

Termes du calcul de $\mathcal{Z} \Leftrightarrow$ Poids total des struct. accessibles sur ce choix.

$\Rightarrow$ Backtrack stochastique engendre des structures compatibles sous contraintes d'appariements selon une distribution de Boltzmann [DL03].

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

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

$$A \,|\, B_{i+\theta+1} \,|\, B_{i+\theta+2} \,|\, B_{i+\theta+3} \,|\, \ldots \,|\, B_{j-2} \,|\, B_{j-1} \,|\, B_j$$

$\Rightarrow$ Complexité du cas au pire en $\mathcal{O}(n^2 k)$ pour k échantillons.

Échantillonnage statistique

Termes du calcul de $\mathcal{Z} \Leftrightarrow$ Poids total des struct. accessibles sur ce choix.

$\Rightarrow$ Backtrack stochastique engendre des structures compatibles sous contraintes d'appariements selon une distribution de Boltzmann [DL03].

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

$$\mathcal{Z}_{i,j} \equiv \begin{cases} ??? \dashrightarrow \mathcal{Z}_{i+1,j} & \text{A} \\ \sum_{k=i+\theta+1}^{j} e^{\frac{1}{RT}} \times \mathcal{Z}_{i+1,k-1} \times \mathcal{Z}_{k+1,j} & \text{B}_k \end{cases}$$

$$A \mid B_{i+\theta+1} \mid B_{i+\theta+2} \mid B_{i+\theta+3} \mid \ldots \mid B_{j-2} \mid B_{j-1} \mid B_j$$

$\Rightarrow$ Complexité du cas au pire en $\mathcal{O}(n^2 k)$ pour k échantillons.

Échantillonnage statistique

Termes du calcul de $\mathcal{Z}$ $\Leftrightarrow$ Poids total des struct. accessibles sur ce choix.

$\Rightarrow$ Backtrack stochastique engendre des structures compatibles sous contraintes d'appariements selon une distribution de Boltzmann [DL03].

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

$$\mathcal{Z}_{i,j} = \mathcal{Z}_{i+1,j} \times \mathcal{Z}_{i+1,j}/\mathcal{Z}_{i,j} + \sum_{k=i+\theta+1}^{j} e^{\frac{1}{RT}} \times \mathcal{Z}_{i+1,k-1} \times \mathcal{Z}_{k+1,j} \times \frac{e^{\frac{1}{RT}} \mathcal{Z}_{i+1,k-1} \mathcal{Z}_{k+1,j}}{\mathcal{Z}_{i,j}}$$

A

B_k

r

$A \mid B_{i+\theta+1} \mid B_{i+\theta+2} \mid B_{i+\theta+3} \mid \ldots \mid B_{j-2} \mid B_{j-1} \mid B_j$

$\Rightarrow$ Complexité du cas au pire en $\mathcal{O}(n^2 k)$ pour k échantillons.

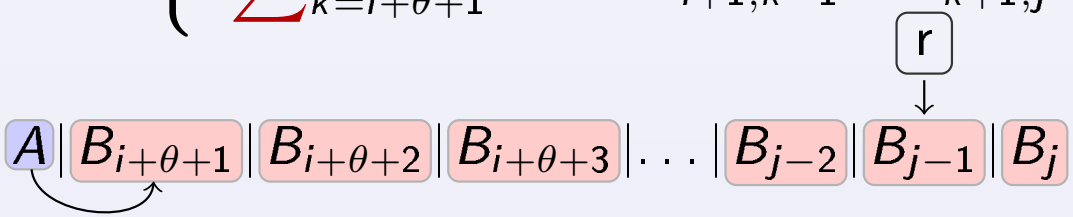
Échantillonnage statistique

Termes du calcul de $\mathcal{Z} \Leftrightarrow$ Poids total des struct. accessibles sur ce choix.

$\Rightarrow$ Backtrack stochastique engendre des structures compatibles sous contraintes d'appariements selon une distribution de Boltzmann [DL03].

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

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



$\Rightarrow$ Complexité du cas au pire en $\mathcal{O}(n^2 k)$ pour k échantillons.

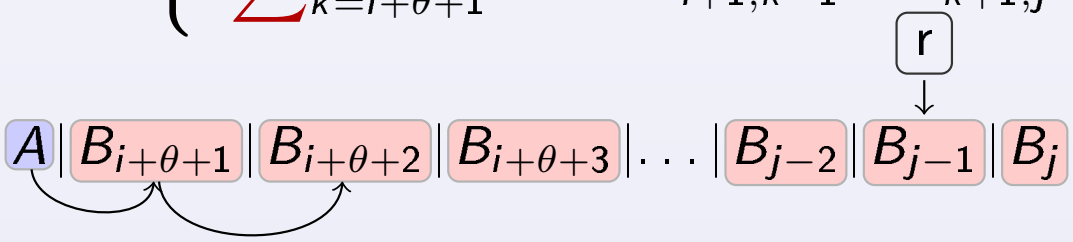
Échantillonnage statistique

Termes du calcul de $\mathcal{Z} \Leftrightarrow$ Poids total des struct. accessibles sur ce choix.

$\Rightarrow$ Backtrack stochastique engendre des structures compatibles sous contraintes d'appariements selon une distribution de Boltzmann [DL03].

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

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



$\Rightarrow$ Complexité du cas au pire en $\mathcal{O}(n^2 k)$ pour k échantillons.

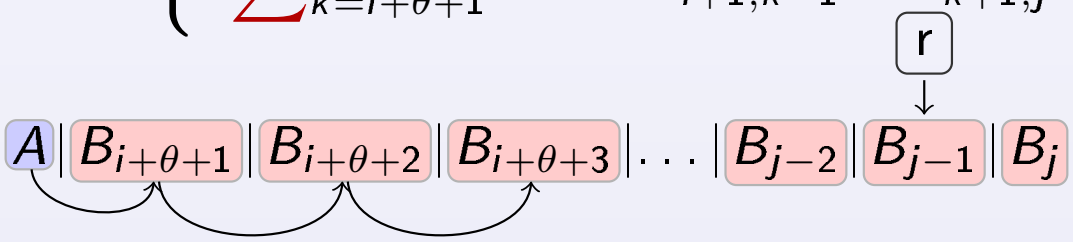
Échantillonnage statistique

Termes du calcul de $\mathcal{Z} \Leftrightarrow$ Poids total des struct. accessibles sur ce choix.

$\Rightarrow$ Backtrack stochastique engendre des structures compatibles sous contraintes d'appariements selon une distribution de Boltzmann [DL03].

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

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



$\Rightarrow$ Complexité du cas au pire en $\mathcal{O}(n^2 k)$ pour k échantillons.

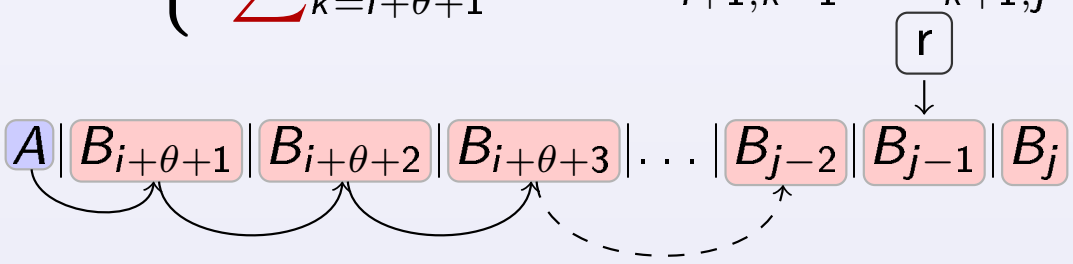
Échantillonnage statistique

Termes du calcul de $\mathcal{Z} \Leftrightarrow$ Poids total des struct. accessibles sur ce choix.

$\Rightarrow$ Backtrack stochastique engendre des structures compatibles sous contraintes d'appariements selon une distribution de Boltzmann [DL03].

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

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



$\Rightarrow$ Complexité du cas au pire en $\mathcal{O}(n^2 k)$ pour k échantillons.

Échantillonnage statistique

Termes du calcul de $\mathcal{Z} \Leftrightarrow$ Poids total des struct. accessibles sur ce choix.

$\Rightarrow$ Backtrack stochastique engendre des structures compatibles sous contraintes d'appariements selon une distribution de Boltzmann [DL03].

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

$$\mathcal{Z}_{i,j} = \sum \left\{ \begin{array}{l} \mathcal{Z}_{i+1,j} \\ \sum_{k=i+\theta+1}^{j} e^{\frac{1}{RT}} \times \mathcal{Z}_{i+1,k-1} \times \mathcal{Z}_{k+1,j} \end{array} \right.$$

A

B_k

r

$A \mid B_{i+\theta+1} \mid B_{i+\theta+2} \mid B_{i+\theta+3} \mid \ldots \mid B_{j-2} \mid B_{j-1} \mid B_j$

$\Rightarrow$ Complexité du cas au pire en $\mathcal{O}(n^2 k)$ pour k échantillons.

Résumé

Un point de vue combinatoire sur l'espace des conformations ...

- ... permet de quantifier sa cardinalité asymptotique.
- ... unifie l'algorithmique des paradigmes MFE/Boltzmann.
- ... aide à prouver la correction d'un schéma de prog. dyn.
- ... donne un cadre pour l'analyse en moyenne des algorithmes.

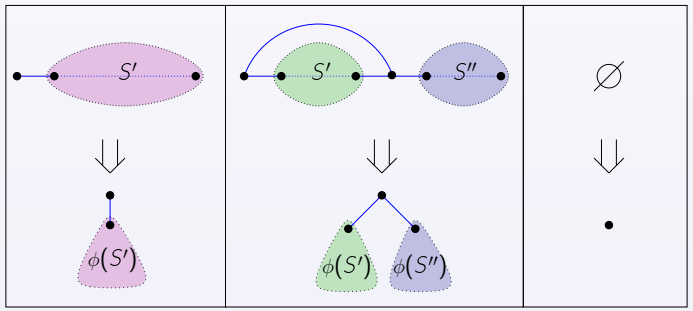
Plan

Analyse en moyenne

Modèle Homopolymère : Toutes paires de bases autorisées.
Distribution de Boltzmann : Basée sur une énergie de Nussinov/Jacobson.

Théorème

Soit n la longueur de l'ARN et k le nombre d'échantillons engendrés.
La complexité en moyenne de l'échantillonnage est en $\Theta(n^3 + kn\sqrt{n})$.



Remarque 1 : Le coût de génération $c(S)$ d'une structure S est dominé par la somme des tailles des fils gauches $t \in \mathcal{T}_g(S)$ de l'arbre unaire/binaire associé :

$$c(S) = n - \#bp(S)/2 + \sum_{t \in \mathcal{T}_g(S)} |t| - \theta$$

Analyse en moyenne

Modèle Homopolymère : Toutes paires de bases autorisées.
Distribution de Boltzmann : Basée sur une énergie de Nussinov/Jacobson.

Théorème

Soit n la longueur de l'ARN et k le nombre d'échantillons engendrés.
La complexité en moyenne de l'échantillonnage est en $\Theta(n^3 + kn\sqrt{n})$.

On s'intéresse à l'espérance de la v.a. X *coût d'une génération*. On a alors

$$\mathbb{E}(X \mid n) = \sum_{|S|=n} c(S) \cdot \frac{e^{\frac{bp(S)}{RT}}}{\mathcal{Z}_n} = \frac{\sum_{|S|=n} c(S) \cdot e^{\frac{bp(S)}{RT}}}{\sum_{|S|=n} e^{\frac{bp(S)}{RT}}}.$$

On sépare notre étude en introduisant deux s.g. $C(z)$ et $P_f(z)$ telles que

$$C(z) = \sum_{S \in \mathcal{S}} e^{\frac{bp(S)}{RT}} c(S) z^{|S|} \quad \text{et} \quad P_f(z) = \sum_{S \in \mathcal{S}} e^{\frac{bp(S)}{RT}} z^{|S|}.$$

On note alors que $\mathbb{E}(X \mid n) = [z^n]C(z)/[z^n]P_f(z)$.

Analyse en moyenne

Modèle Homopolymère : Toutes paires de bases autorisées.
Distribution de Boltzmann : Basée sur une énergie de Nussinov/Jacobson.

Théorème

Soit n la longueur de l'ARN et k le nombre d'échantillons engendrés.
La complexité en moyenne de l'échantillonnage est en $\Theta(n^3 + kn\sqrt{n})$.

$C(z)$ et $P_f(z)$ sont alors les solutions positives du système

$$\begin{aligned} C(z) &= z\left(P_f(z) + C(z)\right) + z^2 e^{\frac{1}{RT}}(1-\theta)P_f^{\geq\theta}(z)P_f(z) \\ &+ z^3 e^{\frac{1}{RT}} \frac{\partial P_f^{\geq\theta}(z)}{\partial z} P_f(z) + z^2 e^{\frac{1}{RT}} C^{\geq\theta}(z)P_f(z) \\ &+ z^2 e^{\frac{1}{RT}} P_f^{\geq\theta}(z)C(z) \\ P_f(z) &= z^2 e^{\frac{1}{RT}} P_f^{\geq\theta}(z)P_f(z) + zP_f(z) + 1 \end{aligned}$$

où $S^{\geq\theta}(z)$ est la s.g. $S(z)$ privée de ses termes de degré $< \theta$.

Analyse en moyenne

Modèle Homopolymère : Toutes paires de bases autorisées.
Distribution de Boltzmann : Basée sur une énergie de Nussinov/Jacobson.

Théorème

Soit n la longueur de l'ARN et k le nombre d'échantillons engendrés.
La complexité en moyenne de l'échantillonnage est en $\Theta(n^3 + kn\sqrt{n})$.

$\Rightarrow$ Grosses séries génératrices un peu complexes (Merci Maple !)...

+ Analyse de singularités :

$$[z^n]P_f(z) \sim \frac{\kappa}{\rho^n n\sqrt{n}}(1+\mathcal{O}(1/n)) \qquad [z^n]C(z) \sim \frac{\kappa'}{\rho^n}(1+\mathcal{O}(1/\sqrt{n}))$$

On obtient ainsi un équivalent asymptotique pour $\mathbb{E}(X \mid n)$:

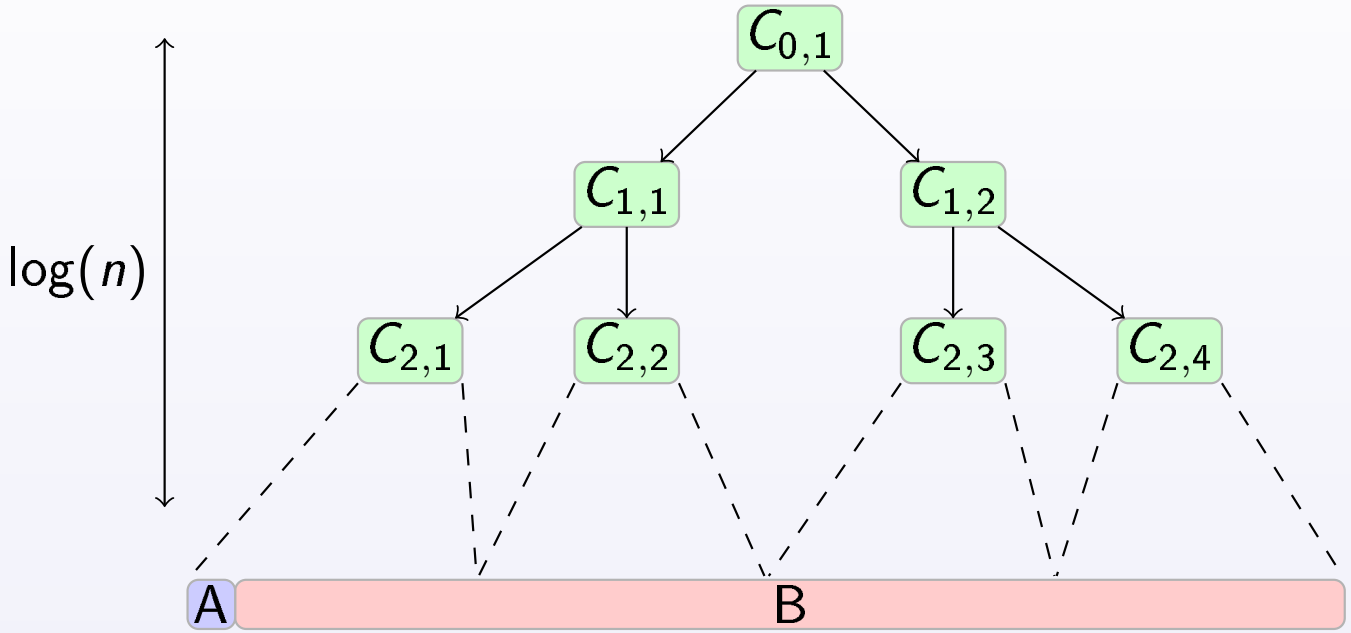
$$\mathbb{E}(X \mid n) = \frac{[z^n]C(z)}{[z^n]P_f(z)} \sim \frac{\kappa'}{\kappa} n\sqrt{n}(1+\mathcal{O}(1/\sqrt{n})).$$

□

Parcours Boustrophédon

En moyenne, on perd du temps à tester des décompositions infructeuses !

- Organisation hiérarchique des contributions



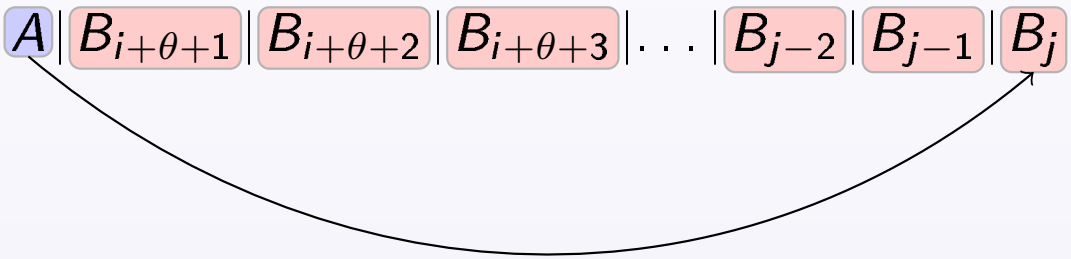
$\Rightarrow$ Génération en $\mathcal{O}(n^3 + kn\log(n))$, mémoire en $\Theta(n^3)$.

- Parcours Boustrophédon

Parcours Boustrophédon

En moyenne, on perd du temps à tester des décompositions infructeuses !

- Organisation hiérarchique des contributions
  $\Rightarrow$ Génération en $\mathcal{O}(n^3 + kn\log(n))$, mémoire en $\Theta(n^3)$.
- Parcours Boustrophédon



Décomposition inégale
$\Rightarrow$ Coût *immédiat* faible, récursion sur intervalle de taille $\mathcal{O}(j\text{-}i\text{-}c)$.
Décomposition égale $\Rightarrow$ Coût *immédiat* élevé, diviser pour régner.

Complexité au pire solution de

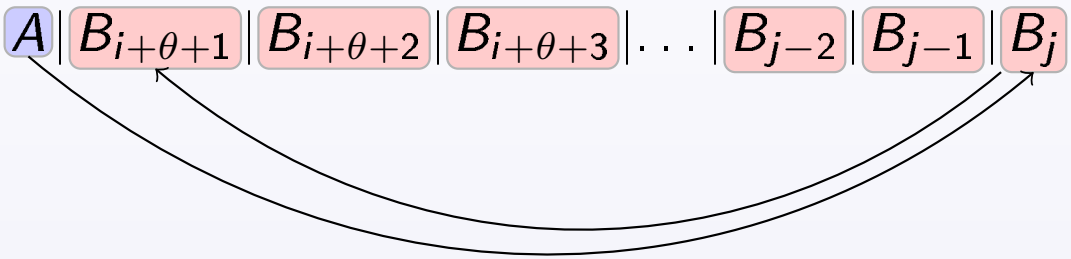
$$f(n) = \max_{k\in[1,n-1]} (f(k)+f(n-k)+2\min(k,n-k)) \in \mathcal{O}(n\log(n))$$ [FZV94].

$\Rightarrow$ Génération en $\mathcal{O}(n^3 + kn\log(n))$ sans précalcul supplémentaire.

Parcours Boustrophédon

En moyenne, on perd du temps à tester des décompositions infructeuses !

- Organisation hiérarchique des contributions
  $\Rightarrow$ Génération en $\mathcal{O}(n^3 + kn\log(n))$, mémoire en $\Theta(n^3)$.
- Parcours Boustrophédon



Décomposition inégale
$\Rightarrow$ Coût *immédiat* faible, récursion sur intervalle de taille $\mathcal{O}(j\text{-}i\text{-}c)$.
Décomposition égale $\Rightarrow$ Coût *immédiat* élevé, diviser pour régner.

Complexité au pire solution de

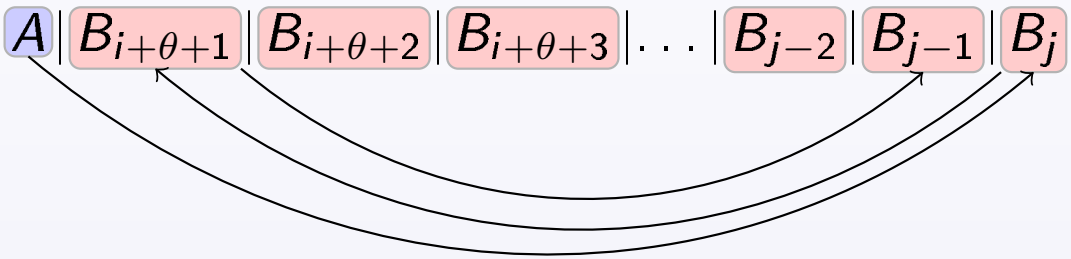
$$f(n) = \max_{k\in[1,n-1]} (f(k)+f(n-k)+2\min(k,n-k)) \in \mathcal{O}(n\log(n))$$ [FZV94].

$\Rightarrow$ Génération en $\mathcal{O}(n^3 + kn\log(n))$ sans précalcul supplémentaire.

Parcours Boustrophédon

En moyenne, on perd du temps à tester des décompositions infructeuses !

- Organisation hiérarchique des contributions
  $\Rightarrow$ Génération en $\mathcal{O}(n^3 + kn\log(n))$, mémoire en $\Theta(n^3)$.
- Parcours Boustrophédon



Décomposition inégale
$\Rightarrow$ Coût *immédiat* faible, récursion sur intervalle de taille $\mathcal{O}(j\text{-}i\text{-}c)$.
Décomposition égale $\Rightarrow$ Coût *immédiat* élevé, diviser pour régner.

Complexité au pire solution de

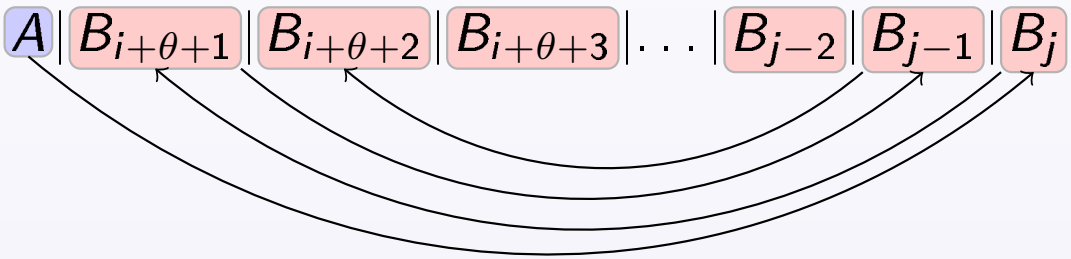
$$f(n) = \max_{k\in[1,n-1]} (f(k)+f(n-k)+2\min(k,n-k)) \in \mathcal{O}(n\log(n))$$ [FZV94].

$\Rightarrow$ Génération en $\mathcal{O}(n^3 + kn\log(n))$ sans précalcul supplémentaire.

Parcours Boustrophédon

En moyenne, on perd du temps à tester des décompositions infructeuses !

- Organisation hiérarchique des contributions
  $\Rightarrow$ Génération en $\mathcal{O}(n^3 + kn\log(n))$, mémoire en $\Theta(n^3)$.
- Parcours Boustrophédon



Décomposition inégale
$\Rightarrow$ Coût *immédiat* faible, récursion sur intervalle de taille $\mathcal{O}(j\text{-}i\text{-}c)$.
Décomposition égale $\Rightarrow$ Coût *immédiat* élevé, diviser pour régner.

Complexité au pire solution de

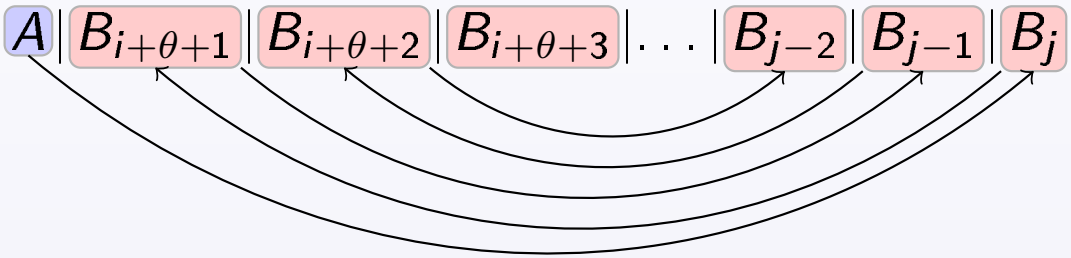
$$f(n) = \max_{k\in[1,n-1]} (f(k)+f(n-k)+2\min(k,n-k)) \in \mathcal{O}(n\log(n))$$ [FZV94].

$\Rightarrow$ Génération en $\mathcal{O}(n^3 + kn\log(n))$ sans précalcul supplémentaire.

Parcours Boustrophédon

En moyenne, on perd du temps à tester des décompositions infructeuses !

- Organisation hiérarchique des contributions
  $\Rightarrow$ Génération en $\mathcal{O}(n^3 + kn\log(n))$, mémoire en $\Theta(n^3)$.
- Parcours Boustrophédon



Décomposition inégale
$\Rightarrow$ Coût *immédiat* faible, récursion sur intervalle de taille $\mathcal{O}(j\text{-}i\text{-}c)$.
Décomposition égale $\Rightarrow$ Coût *immédiat* élevé, diviser pour régner.

Complexité au pire solution de

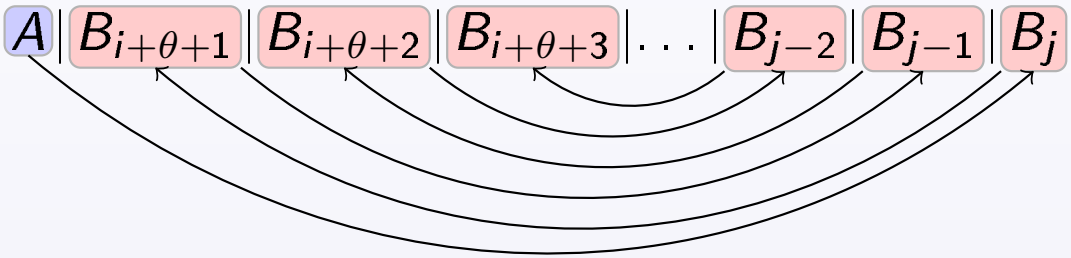
$$f(n) = \max_{k\in[1,n-1]} (f(k)+f(n-k)+2\min(k,n-k)) \in \mathcal{O}(n\log(n))$$ [FZV94].

$\Rightarrow$ Génération en $\mathcal{O}(n^3 + kn\log(n))$ sans précalcul supplémentaire.

Parcours Boustrophédon

En moyenne, on perd du temps à tester des décompositions infructeuses !

- Organisation hiérarchique des contributions
  $\Rightarrow$ Génération en $\mathcal{O}(n^3 + kn\log(n))$, mémoire en $\Theta(n^3)$.
- Parcours Boustrophédon



Décomposition inégale
$\Rightarrow$ Coût *immédiat* faible, récursion sur intervalle de taille $\mathcal{O}(j\text{-}i\text{-}c)$.

Décomposition égale $\Rightarrow$ Coût *immédiat* élevé, diviser pour régner.

Complexité au pire solution de

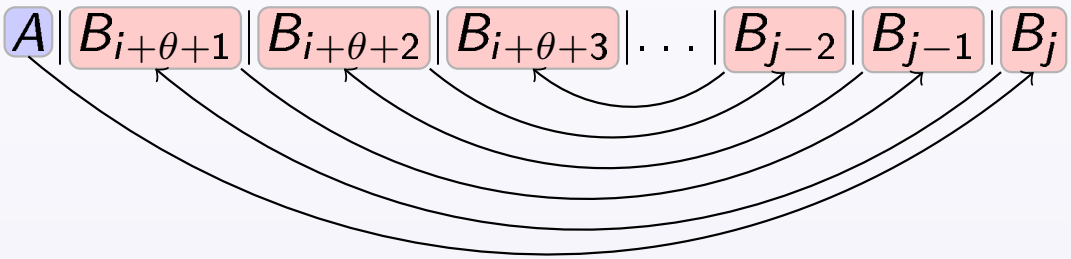
$$f(n) = \max_{k\in[1,n-1]} (f(k)+f(n-k)+2\min(k,n-k)) \in \mathcal{O}(n\log(n))$$ [FZV94].

$\Rightarrow$ Génération en $\mathcal{O}(n^3 + kn\log(n))$ sans précalcul supplémentaire.

Parcours Boustrophédon

En moyenne, on perd du temps à tester des décompositions infructeuses !

- Organisation hiérarchique des contributions
  $\Rightarrow$ Génération en $\mathcal{O}(n^3 + kn\log(n))$, mémoire en $\Theta(n^3)$.
- Parcours Boustrophédon



Décomposition inégale
$\Rightarrow$ Coût *immédiat* faible, récursion sur intervalle de taille $\mathcal{O}(j\text{-}i\text{-}c)$.
Décomposition égale $\Rightarrow$ Coût *immédiat* élevé, diviser pour régner.

Complexité au pire solution de

$$f(n) = \max_{k\in[1,n-1]} (f(k)+f(n-k)+2\min(k,n-k)) \in \mathcal{O}(n\log(n))$$ [FZV94].

$\Rightarrow$ Génération en $\mathcal{O}(n^3 + kn\log(n))$ sans précalcul supplémentaire.

Plan

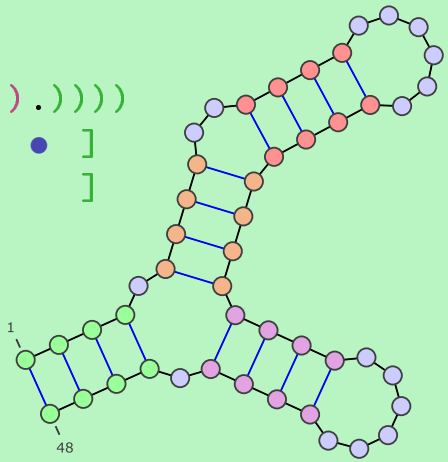
Definition (RNA shapes [GVR04])

Hiérarchie de représentation *gros-grain* pour la struct. sec. d'ARN.

Basée sur l'arbre *sous-jacent* à la structure secondaire.

Exemple

```
Sec. str. ((((.((((..((((......))))))))((((......)))).))))
π′-shape  [ • [ • [   •   ]  ]  [   •   ] • ]
π-shape   [   [     -       -  ]  [       ]   ]
```



Definition (RNA shapes [GVR04])

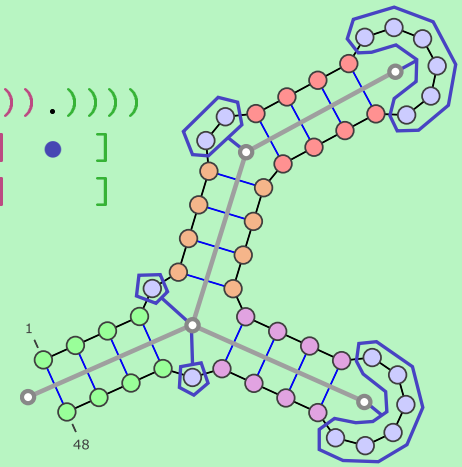
Hiérarchie de représentation *gros-grain* pour la struct. sec. d'ARN.

Basée sur l'arbre *sous-jacent* à la structure secondaire.

Exemple

Sec. str. ((((.((((..((((.....))))))))((((.....)))).))))
π'-shape [• [• [•]] [•] •]
π-shape [[- -] [ ]]

Contracter les caractères identiques consécutifs



Definition (RNA shapes [GVR04])

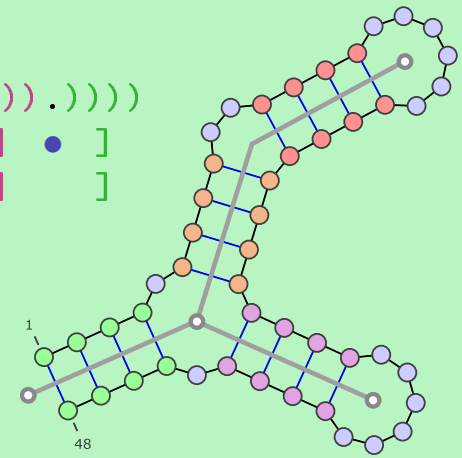
Hiérarchie de représentation *gros-grain* pour la struct. sec. d'ARN.

Basée sur l'arbre *sous-jacent* à la structure secondaire.

Exemple

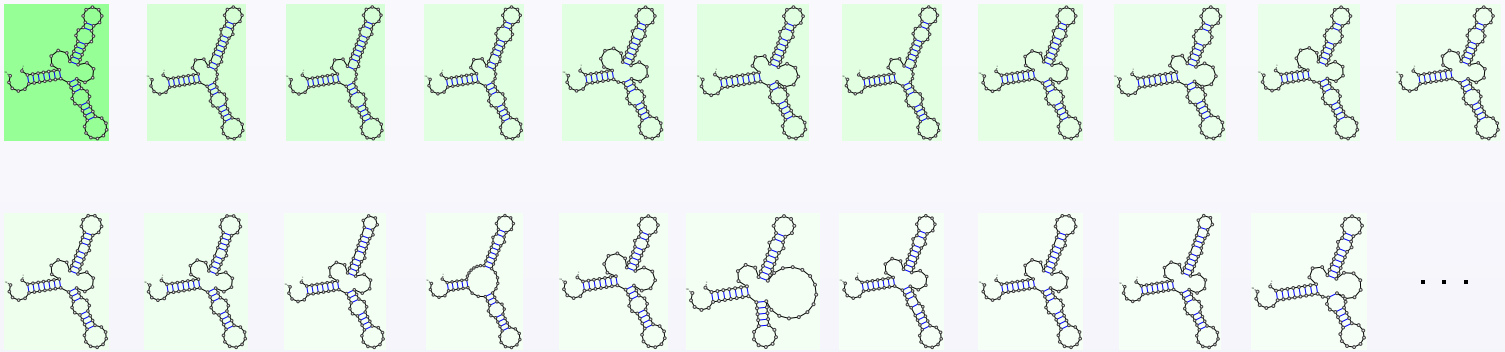
Sec. str. ((((.((((..((((.....))))))))((((.....)))).))))
π'-shape [• [• [•]] [•] •]
π-shape [[- -] [ ]]

Occulter les régions non-appariées
Contracter les hélices imbriquées



Motivation

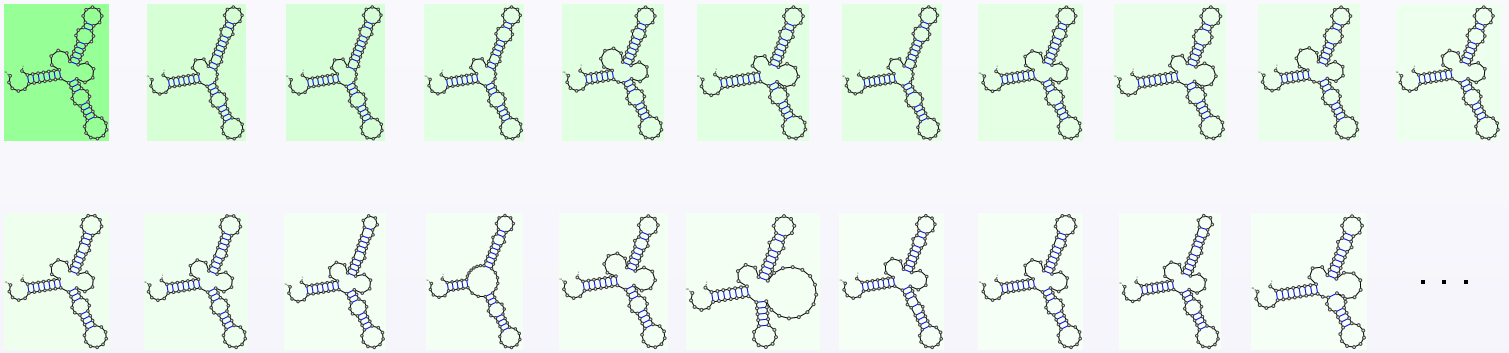
Les shapes permettent une exploration hiérarchique de l'ensemble de Boltzmann.



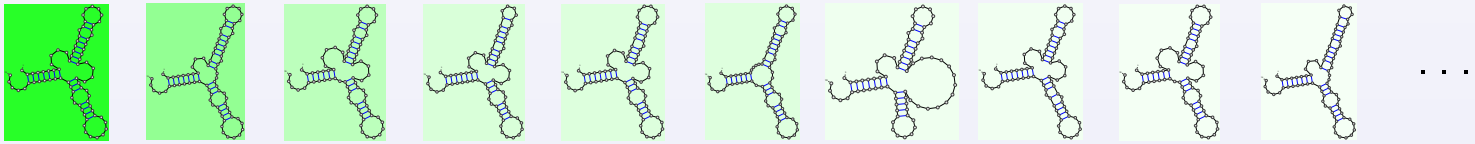
Génération de 10000 échantillons $\Rightarrow$ 1727 Struct. sec. ...

Motivation

Les shapes permettent une exploration hiérarchique de l'ensemble de Boltzmann.



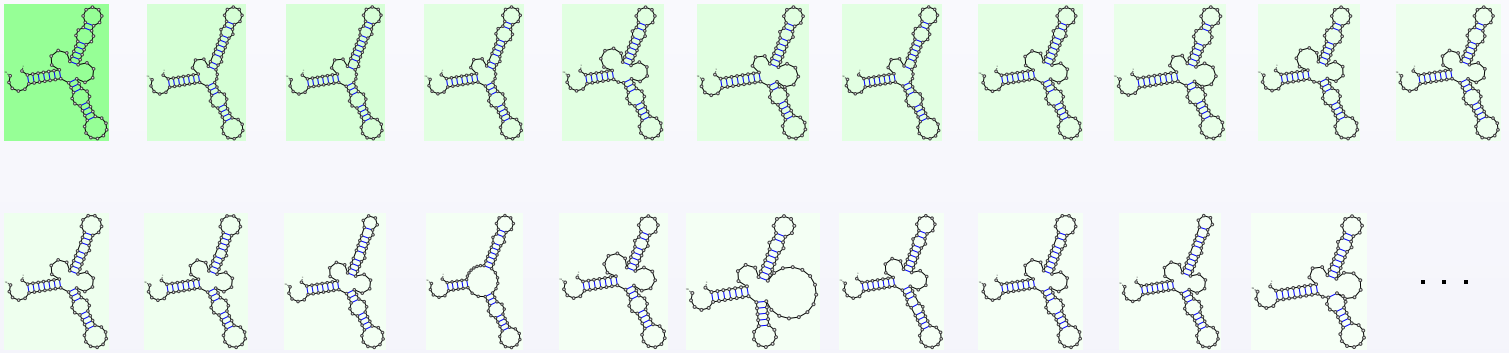
Génération de 10000 échantillons $\Rightarrow$ 1727 Struct. sec. ...



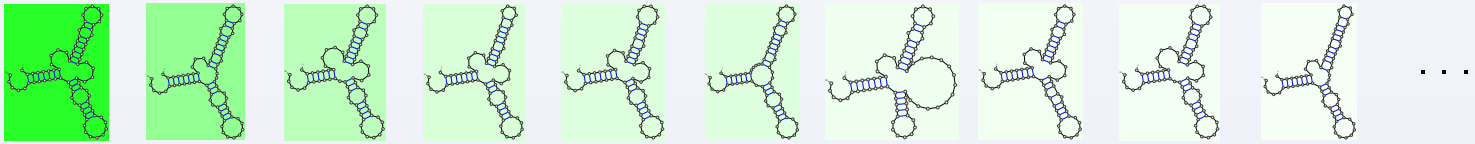
...406 shapes π'...

Motivation

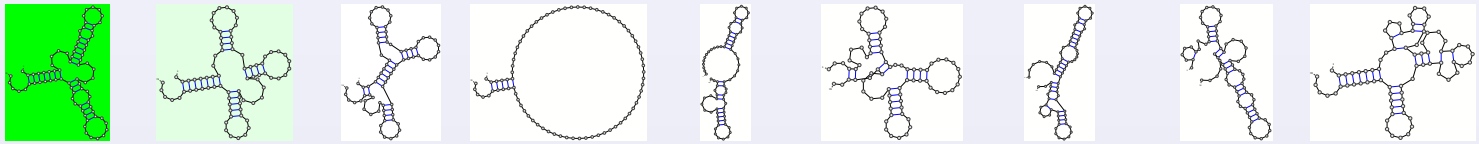
Les shapes permettent une exploration hiérarchique de l'ensemble de Boltzmann.



Génération de 10000 échantillons $\Rightarrow$ 1727 Struct. sec. ...



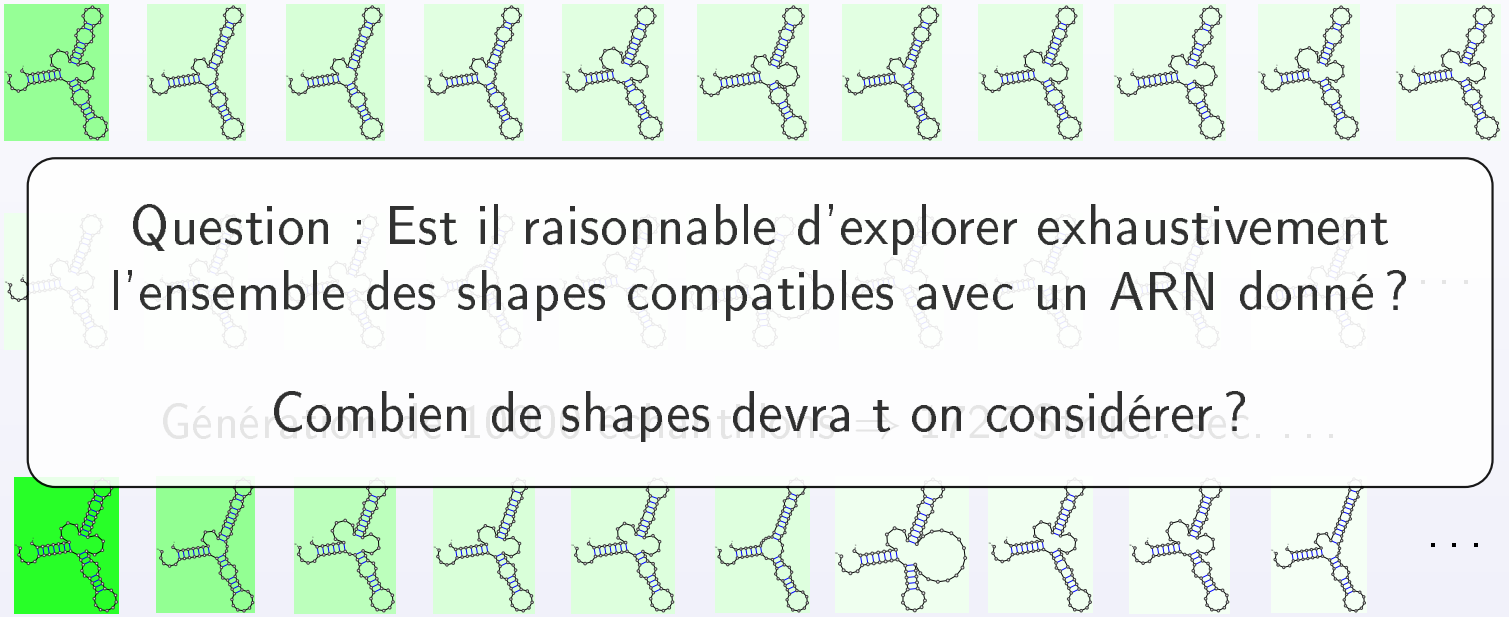
...406 shapes π'...



...mais seulement 9 shapes π !

Motivation

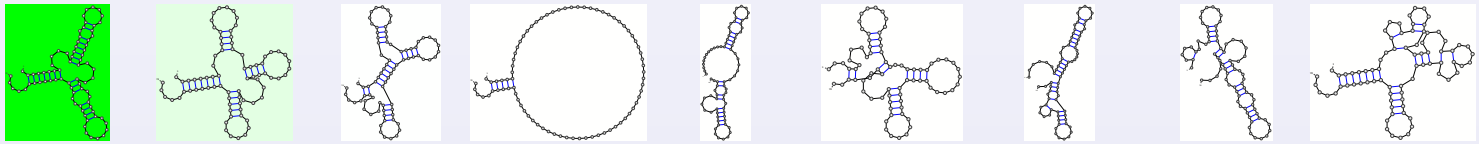
Les shapes permettent une exploration hiérarchique de l'ensemble de Boltzmann.



Question : Est il raisonnable d'explorer exhaustivement l'ensemble des shapes compatibles avec un ARN donné ?

Combien de shapes devra t on considérer ?

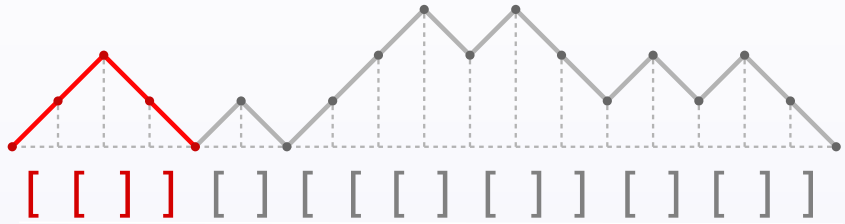
. . . 406 shapes π' . . .



. . . mais seulement 9 shapes π !

π-shapes

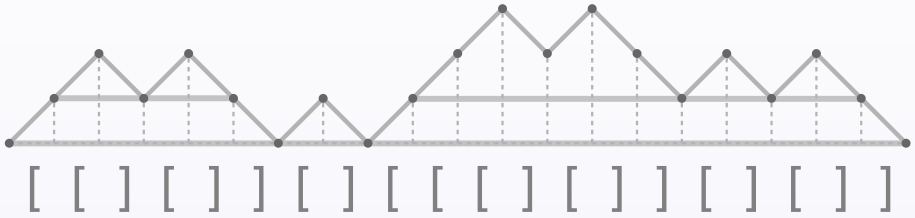
Objectif : Compter les shapes π à $2n$ parenthèses.



1 Shapes π = Mots bien parenthésés évitant le motif `[[...]]`.

π-shapes

Objectif : Compter les shapes π à $2n$ parenthèses.

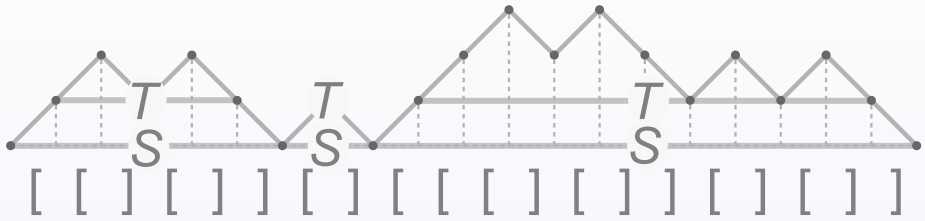


1. Shapes π = Mots bien parenthésés évitant le motif `[[...]]`.

2. $S \rightarrow [\, S_{/\{[\ldots]\}} \,]\, S \mid [\, S_{/\{[\ldots]\}} \,]$

π-shapes

Objectif : Compter les shapes π à $2n$ parenthèses.

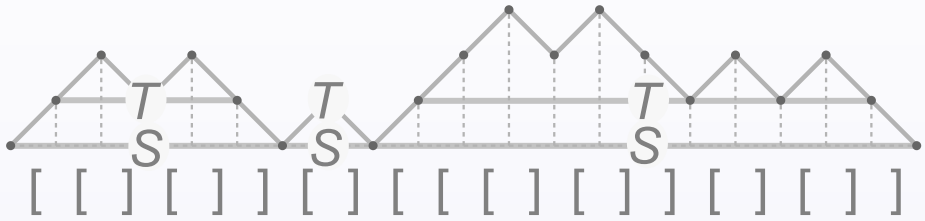


1. Shapes π = Mots bien parenthésés évitant le motif `[[...]]`.

2. $S \rightarrow [\,\mathbf{T}\,]\,S \mid [\,\mathbf{T}\,] \qquad \mathbf{T} \rightarrow [\,\mathbf{T}\,]\,S \mid \varepsilon$

π-shapes

Objectif : Compter les shapes π à $2n$ parenthèses.



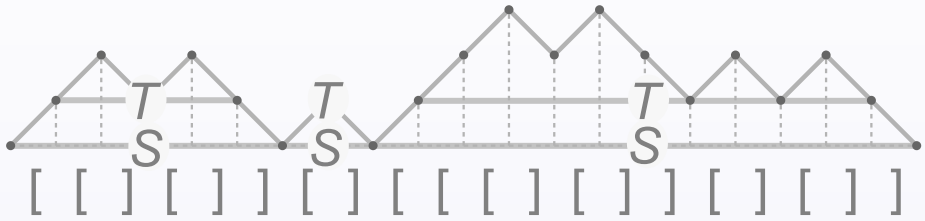
1. Shapes π = Mots bien parenthésés évitant le motif `[[...]]`.

2. $S \rightarrow [\,\mathbf{T}\,]\,S \mid [\,\mathbf{T}\,] \qquad \mathbf{T} \rightarrow [\,\mathbf{T}\,]\,S \mid \varepsilon$

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

π-shapes

Objectif : Compter les shapes π à $2n$ parenthèses.



1. Shapes π = Mots bien parenthésés évitant le motif `[[...]]`.

2. $S \rightarrow [\,\mathbf{T}\,]\,S \mid [\,\mathbf{T}\,] \qquad \mathbf{T} \rightarrow [\,\mathbf{T}\,]\,S \mid \varepsilon$

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

4. $$s_{2n} \sim \frac{\sqrt{3}}{2\sqrt{\pi}} \cdot \frac{3^n}{n\sqrt{\pi}}(1 + \mathcal{O}(1/n)) \quad \text{et} \quad s_{2n+1} = 0$$

Remarque : Cela vous fait il penser à quelque chose ? ? ?

Raffinement du modèle

Limitations

Nombre de shapes π de taille n

$\neq$

Nombre de shapes π compatibles avec un RNA de taille n

Raisons principales :

1. Shapes de taille $\leq n$ devraient être aussi considérées
2. Création de motif *terminal* [ ] implique au moins $\theta + 2$ bases.

2 $S \rightarrow [\ T\ ]\ S \mid [\ T\ ] \qquad T \rightarrow [\ T\ ]\ S \mid \varepsilon$

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

4 $$\text{Pour } n \text{ pair} : s_{2n} \sim \frac{3\sqrt{3}}{4\sqrt{\pi}} \cdot \frac{3^n}{n\sqrt{\pi}}(1 + \mathcal{O}(1/n)) \approx 0.48 \cdot \frac{3^n}{n\sqrt{n}}$$

Raffinement du modèle

Limitations

Nombre de shapes π de taille n

$\neq$

Nombre de shapes π compatibles avec un RNA de taille n

Raisons principales :

1. Shapes de taille $\leq n$ devraient être aussi considérées
2. Création de motif *terminal* [ ] implique au moins $\theta + 2$ bases.

2 $$S \rightarrow [\ T\ ]\ S \mid [\ T\ ] \qquad T \rightarrow [\ T\ ]\ S \mid \bullet^{\theta}$$
$$R \rightarrow \square\ \mathrm{S} \mid \varepsilon$$

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

4 $$r_{2n} \sim \frac{3\sqrt{3}}{4\sqrt{\pi}} \cdot \frac{3^n}{n\sqrt{n}}(1 + \mathcal{O}(1/n)) \Rightarrow r_n \approx 2.07 \cdot \frac{1.73^n}{n\sqrt{n}}$$

Raffinement du modèle

Limitations

Nombre de shapes π de taille n

$\neq$

Nombre de shapes π compatibles avec un RNA de taille n

Raisons principales :

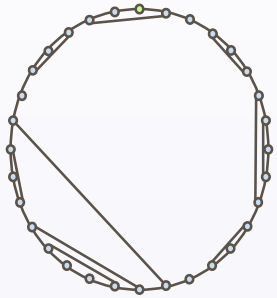
1. Shapes de taille $\leq n$ devraient être aussi considérées
2. Création de motif *terminal* [ ] implique au moins $\theta + 2$ bases.

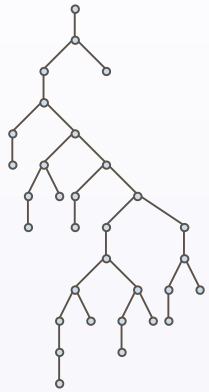
2 $$S \rightarrow [\ T\ ]\ S \mid [\ T\ ] \qquad T \rightarrow [\ T\ ]\ S \mid \bullet^{\theta}$$
$$R \rightarrow \square\ \mathrm{S} \mid \varepsilon$$

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

4 $$\theta = 3 \Rightarrow r_n \approx 2.44 \frac{1.32^n}{n\sqrt{n}}$$

Appartée : Une bijection amusante





Théorème

$$\#\ \textit{shapes}\ \pi\ \textit{de taille}\ 2n+2 = \#\ \textit{mots de Motzkin de taille}\ n$$

Preuve

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

$$S(z) = 1 + z^2 M(z^2) \quad \Rightarrow \quad s_{2n+2} = m_n$$

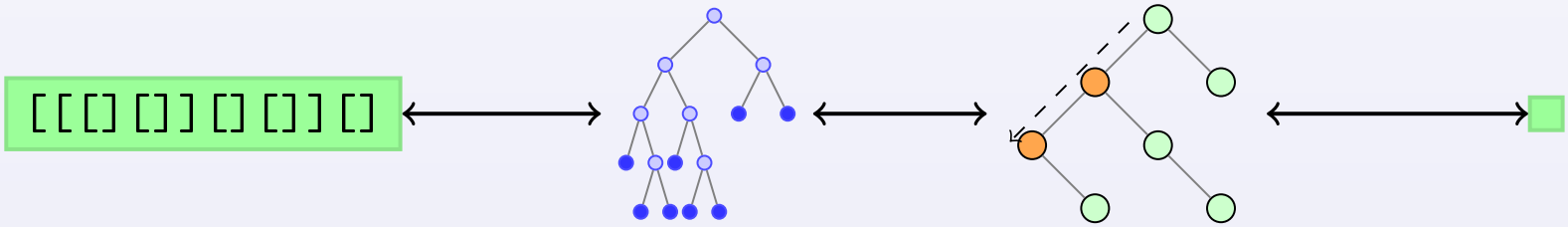
Ces deux classes sont en bijection.
Comment ? Comment se projettent les statistiques ?

Explicit bijection

Soit $\psi, \phi : \{ [,] \}^* \to \{ (,) , \bullet \}$ tels que

$$\begin{aligned} \psi((\ A\ )\ B) &= \begin{cases} \phi(A) & \text{Si } B = \varepsilon \\ \phi(A) \bullet \psi(B) & \text{Sinon} \end{cases} \\ \phi((\ A\ )\ B) &= \phi(A)[\psi(B)] \\ \phi(\varepsilon) &= \varepsilon. \end{aligned}$$

Alors ψ est bijective entre s_{2n+2} et m_n.



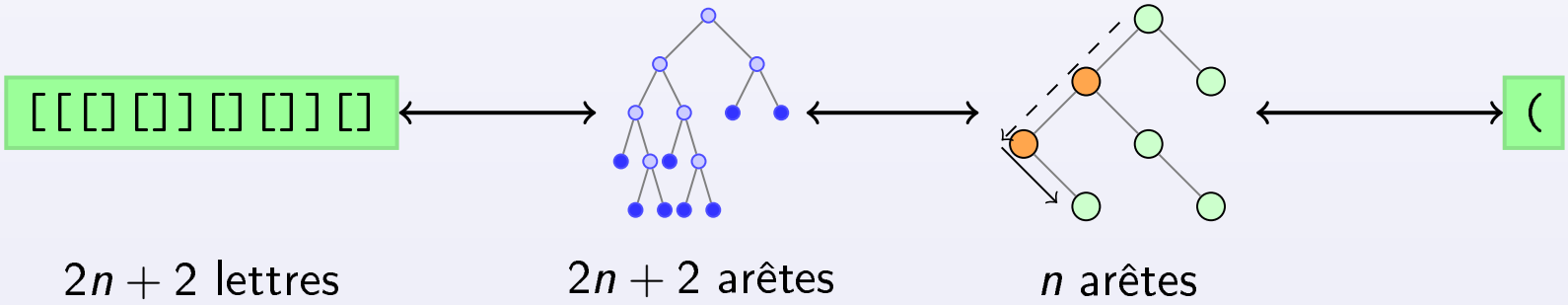
$2n+2$ lettres $\quad$ $2n+2$ arêtes $\quad$ n arêtes

Explicit bijection

Soit $\psi, \phi : \{ [,] \}^* \to \{ (,) , \bullet \}$ tels que

$$\begin{aligned} \psi((A)B) &= \begin{cases} \phi(A) & \text{Si } B = \varepsilon \\ \phi(A) \bullet \psi(B) & \text{Sinon} \end{cases} \\ \phi((A)B) &= \phi(A)[\psi(B)] \\ \phi(\varepsilon) &= \varepsilon. \end{aligned}$$

Alors ψ est bijective entre s_{2n+2} et m_n.



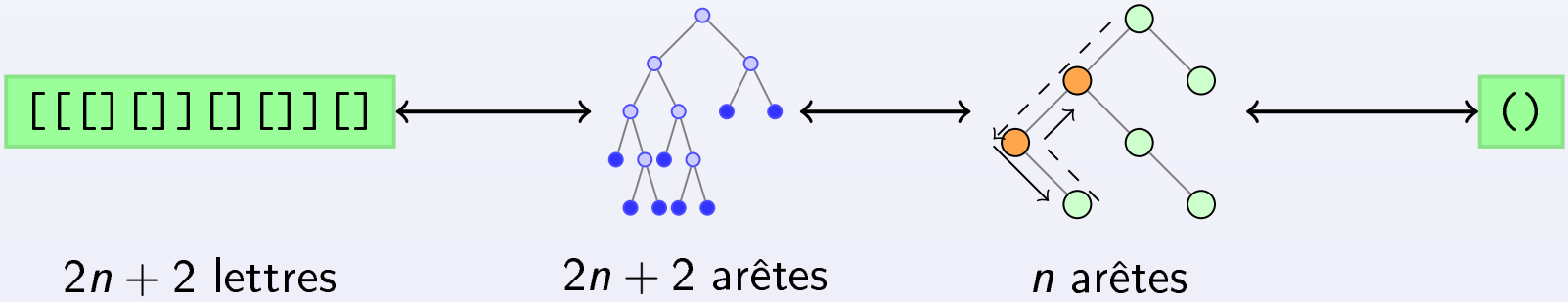
$2n+2$ lettres $\quad$ $2n+2$ arêtes $\quad$ n arêtes

Explicit bijection

Soit $\psi, \phi : \{ [,] \}^* \to \{ (,) , \bullet \}$ tels que

$$\begin{aligned} \psi((\ A\ )\ B) &= \begin{cases} \phi(A) & \text{Si } B = \varepsilon \\ \phi(A) \bullet \psi(B) & \text{Sinon} \end{cases} \\ \phi((\ A\ )\ B) &= \phi(A)[\psi(B)] \\ \phi(\varepsilon) &= \varepsilon. \end{aligned}$$

Alors ψ est bijective entre s_{2n+2} et m_n.

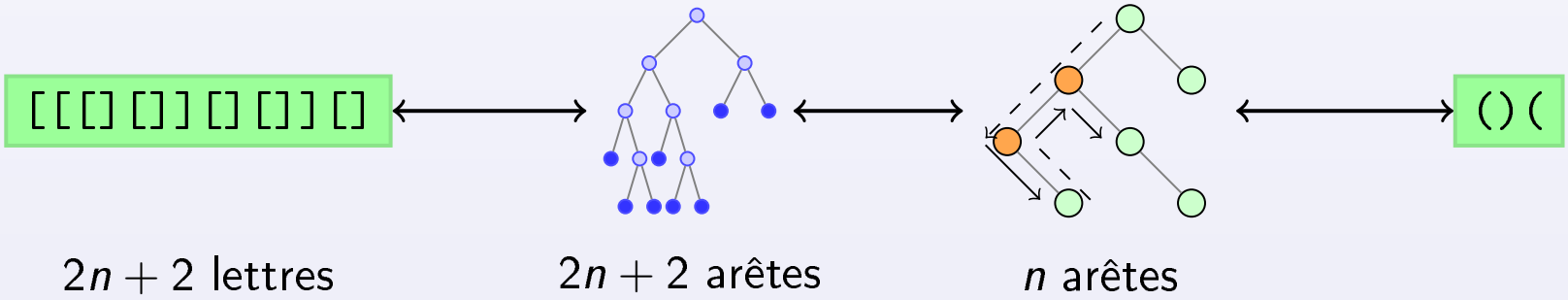


Explicit bijection

Soit $\psi, \phi : \{ [\ ,\ ] \}^* \to \{ (\ ,\ ) , \bullet \}$ tels que

$$\begin{aligned} \psi((\ A\ )\ B) &= \begin{cases} \phi(A) & \text{Si } B = \varepsilon \\ \phi(A) \bullet \psi(B) & \text{Sinon} \end{cases} \\ \phi((\ A\ )\ B) &= \phi(A)[\psi(B)] \\ \phi(\varepsilon) &= \varepsilon. \end{aligned}$$

Alors ψ est bijective entre s_{2n+2} et m_n.



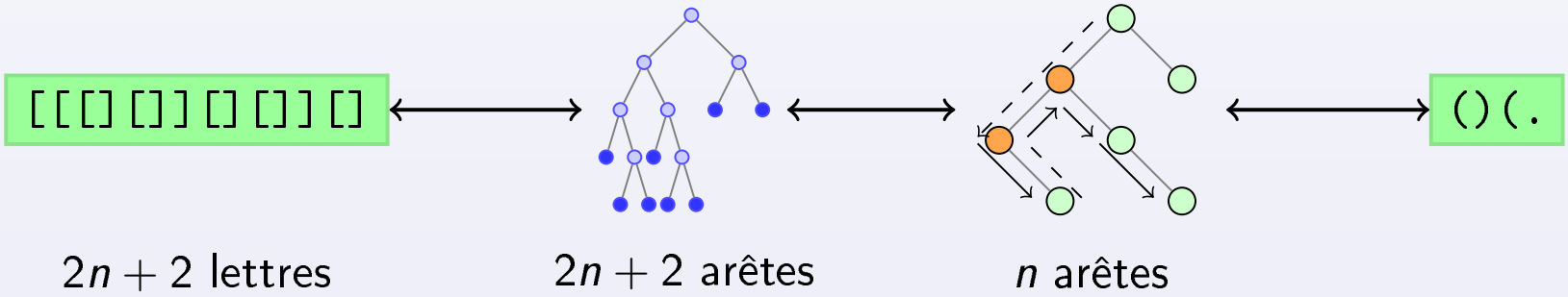
$2n+2$ lettres $\quad$ $2n+2$ arêtes $\quad$ n arêtes

Explicit bijection

Soit $\psi, \phi : \{ [,] \}^* \to \{ (,) , \bullet \}$ tels que

$$\begin{aligned} \psi((\ A\ )\ B) &= \begin{cases} \phi(A) & \text{Si } B = \varepsilon \\ \phi(A) \bullet \psi(B) & \text{Sinon} \end{cases} \\ \phi((\ A\ )\ B) &= \phi(A)[\psi(B)] \\ \phi(\varepsilon) &= \varepsilon. \end{aligned}$$

Alors ψ est bijective entre s_{2n+2} et m_n.

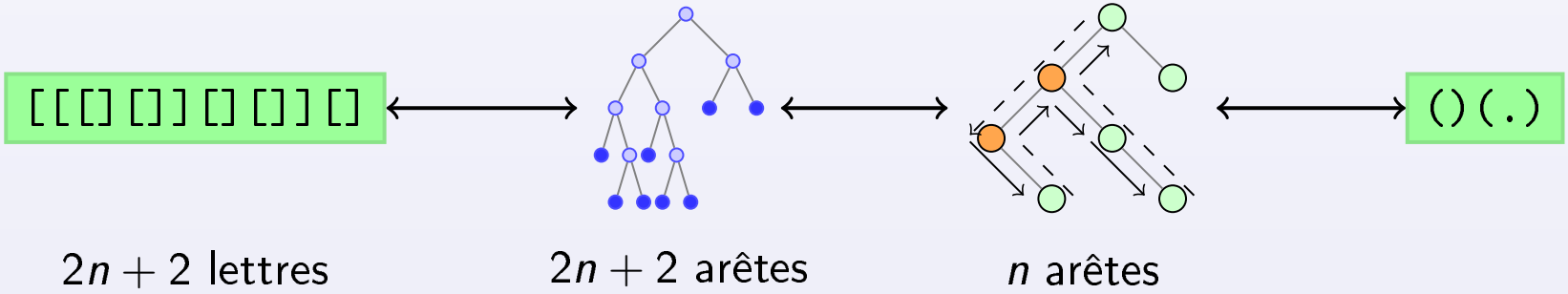


Explicit bijection

Soit $\psi, \phi : \{ [\ ,\ ] \}^* \to \{ (\ ,\ ) , \bullet \}$ tels que

$$\begin{aligned} \psi((\ A\ )\ B) &= \begin{cases} \phi(A) & \text{Si } B = \varepsilon \\ \phi(A) \bullet \psi(B) & \text{Sinon} \end{cases} \\ \phi((\ A\ )\ B) &= \phi(A)[\psi(B)] \\ \phi(\varepsilon) &= \varepsilon. \end{aligned}$$

Alors ψ est bijective entre s_{2n+2} et m_n.



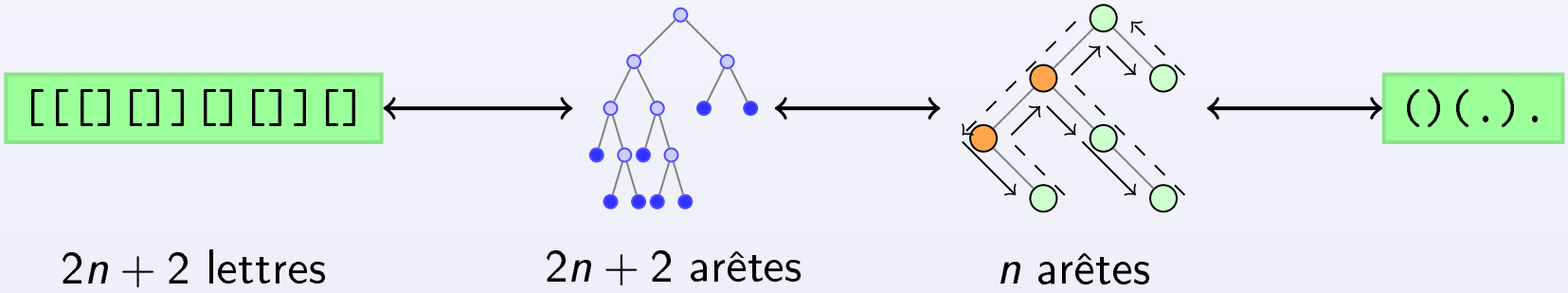
$2n+2$ lettres $\quad$ $2n+2$ arêtes $\quad$ n arêtes

Explicit bijection

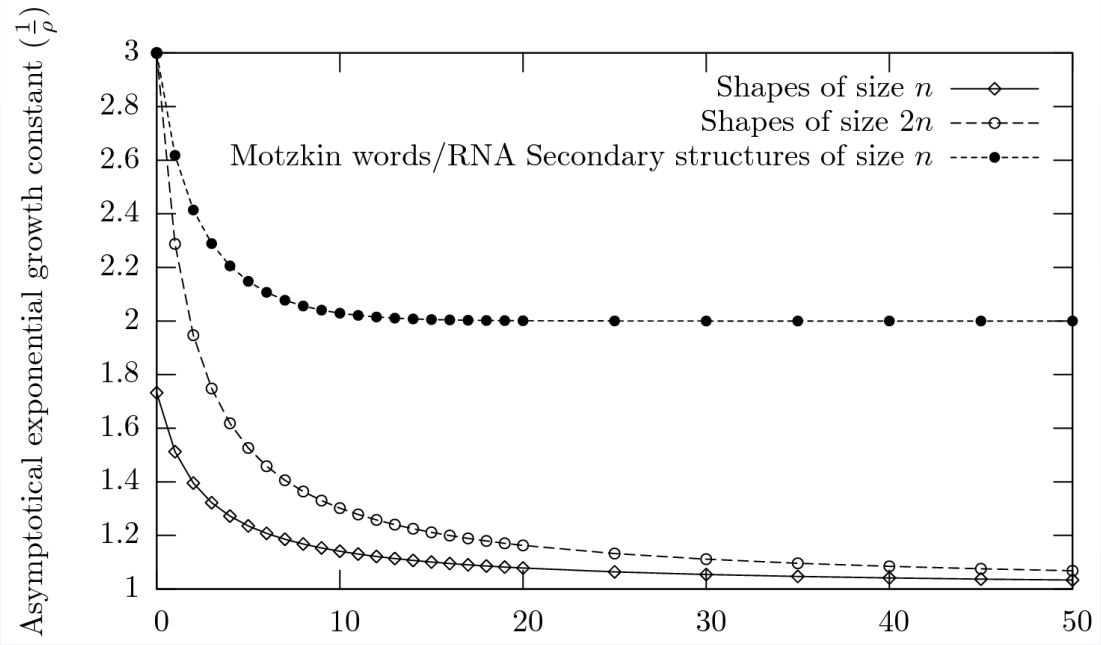
Soit $\psi, \phi : \{ [,] \}^* \to \{ (,) , \bullet \}$ tels que

$$\begin{aligned} \psi((\ A\ )\ B) &= \begin{cases} \phi(A) & \text{Si } B = \varepsilon \\ \phi(A) \bullet \psi(B) & \text{Sinon} \end{cases} \\ \phi((\ A\ )\ B) &= \phi(A)[\psi(B)] \\ \phi(\varepsilon) &= \varepsilon. \end{aligned}$$

Alors ψ est bijective entre s_{2n+2} et m_n.



Limite de la bijection



Impact de θ très différent sur les shapes et les mots de Motzkin.

Théorème

Nombres moyens de motifs terminaux dans les mots de Motzkin words et shapes π grandissent en $m_n^t \sim \frac{n}{6} + \mathcal{O}(1)$ and $s_{2n+2}^t \sim \frac{2n}{3} + \mathcal{O}(1)$.

Preuve utilise séries (temporairement) bivariées.

Shapes π'

Objectif : Compter les shapes π' compatibles avec ARN de taille n.

1. Shapes π' = Mots bien parenthésés évitant `[[...]]` et $\bullet\bullet$

2. $$R \to \square\, R \mid S \qquad S \to U\,[\,T\,]\,S \mid U \qquad U \to \diamond \mid \varepsilon$$
$$T \to U\,[\,T\,]\,U\,[\,T\,]\,S \mid \diamond[\,T\,] \mid [\,T\,]\diamond \mid \diamond[\,T\,]\diamond \mid \bullet^{\theta}$$

3. $$\theta = 3,\ R(z) = \frac{1+2z^2+2z^3+z^4-z^5-z^6-\sqrt{1-4z^3-2z^4-2z^5+2z^6-7z^8-z^{10}+2z^{11}+z^{12}}}{2z^2(1-z^2)}$$

4. $$r_n \sim 1.27\,\frac{1.81^n}{n\sqrt{n}}$$

Shapes π'

Objectif : Compter les shapes π' compatibles avec ARN de taille n.

1. Shapes π' = Mots bien parenthésés évitant $[[\ldots]]$ et $\bullet\bullet$

2. $$R \to \square\, R \mid S \qquad S \to U\,[\,T\,]\,S \mid U \qquad U \to \diamond \mid \varepsilon$$
$$T \to U\,[\,T\,]\,U\,[\,T\,]\,S \mid \diamond[\,T\,] \mid [\,T\,]\diamond \mid \diamond[\,T\,]\diamond \mid \bullet^{\theta}$$

3. $$\theta = 3,\ R(z) = \frac{1+2z^2+2z^3+z^4-z^5-z^6-\sqrt{1-4z^3-2z^4-2z^5+2z^6-7z^8-z^{10}+2z^{11}+z^{12}}}{2z^2(1-z^2)}$$

4. $$r_n \sim 1.27\,\frac{1.81^n}{n\sqrt{n}}$$

Shapes π'

Objectif : Compter les shapes π' compatibles avec ARN de taille n.

1. Shapes π' = Mots bien parenthésés évitant $[[\ldots]]$ et $\bullet\bullet$

2. $$R \to \square\, R \mid S \qquad S \to U\,[\,T\,]\,S \mid U \qquad U \to \diamond \mid \varepsilon$$
$$T \to U\,[\,T\,]\,U\,[\,T\,]\,S \mid \diamond[\,T\,] \mid [\,T\,]\diamond \mid \diamond[\,T\,]\diamond \mid \bullet^{\theta}$$

3. $$\theta = 3,\ R(z) = \frac{1+2z^2+2z^3+z^4-z^5-z^6-\sqrt{1-4z^3-2z^4-2z^5+2z^6-7z^8-z^{10}+2z^{11}+z^{12}}}{2z^2(1-z^2)}$$

4. $$r_n \sim 1.27\frac{1.81^n}{n\sqrt{n}}$$

Shapes π'

Objectif : Compter les shapes π' compatibles avec ARN de taille n.

1. Shapes π' = Mots bien parenthésés évitant $[[\ldots]]$ et $\bullet\bullet$

2. $$R \to \square\, R \mid S \qquad S \to U\,[\,T\,]\,S \mid U \qquad U \to \diamond \mid \varepsilon$$
$$T \to U\,[\,T\,]\,U\,[\,T\,]\,S \mid \diamond[\,T\,] \mid [\,T\,]\diamond \mid \diamond[\,T\,]\diamond \mid \bullet^{\theta}$$

3. $$\theta = 3,\ R(z) = \frac{1+2z^2+2z^3+z^4-z^5-z^6-\sqrt{1-4z^3-2z^4-2z^5+2z^6-7z^8-z^{10}+2z^{11}+z^{12}}}{2z^2(1-z^2)}$$

4. $$r_n \sim 1.27\,\frac{1.81^n}{n\sqrt{n}}$$

Résumé

Modèle	Asymptotique
Sec. str. – Taille n (Combinatoire)	$1.1 \cdot \frac{2.6^n}{n\sqrt{n}}$
Sec. str – Empirique séquence taille n	$0.04 \cdot \frac{1.4^n}{n\sqrt{n}}$
Shapes π – Taille n (Combinatoire)	$1.38 \cdot \frac{1.73^n}{n\sqrt{n}}$
Shapes π – Compatible avec ARN de taille n	$2.44 \cdot \frac{1.32^n}{n\sqrt{n}}$
Shapes π – Empirique séquence taille n	$0.21 \cdot \frac{1.1^n}{n\sqrt{n}}$
Shapes π' – Taille n (Combinatoire)	$0.99 \cdot \frac{2.41^n}{n\sqrt{n}}$
Shapes π' – Compatible avec ARN de taille n	$1.28 \cdot \frac{1.81^n}{n\sqrt{n}}$

Conclusion

Échantillonnage statistique

- Bioalgorithmiciens ont redécouvert la méthode récursive (pondérée).
- Analyse en moyenne et optimisation (Boustrophedon).
- Évaluer sans échantillonnage $\Rightarrow$ Modif. automatisée de la spec. ?
- Boltzmann dans une distribution de Boltzmann ?

Shapes ARN (Collab. : W. A. Lorenz et P. Clote – Boston College)

- Beaucoup moins de shapes que de struct. secondaires !
  $\Rightarrow$ Énumération exhaustive réalisable ($n < 200$).
- Bijection entre les mots de Motzkin et les shapes π
- Modèle homopolymère surestime nombre de shapes
  Incorporer un modèle probabiliste pour possibilité d'appariement.
  Mais Bernoulli n'est pas suffisant . . .

En fournissant des outils pour décrire l'espace des conformations, la combinatoire peut (et doit) aider la bioinfo.
$\Rightarrow$ Incorporer des contraintes et/ou aspects probabilistes à nos modèles.

Conclusion

Échantillonnage statistique

- Bioalgorithmiciens ont redécouvert la méthode récursive (pondérée).
- Analyse en moyenne et optimisation (Boustrophedon).
- Évaluer sans échantillonnage $\Rightarrow$ Modif. automatisée de la spec. ?
- Boltzmann dans une distribution de Boltzmann ?

Shapes ARN (Collab. : W. A. Lorenz et P. Clote – Boston College)

- Beaucoup moins de shapes que de struct. secondaires !
  $\Rightarrow$ Énumération exhaustive réalisable ($n < 200$).
- Bijection entre les mots de Motzkin et les shapes π
- Modèle homopolymère surestime nombre de shapes
  Incorporer un modèle probabiliste pour possibilité d'appariement.
  Mais Bernoulli n'est pas suffisant ...

En fournissant des outils pour décrire l'espace des conformations, la combinatoire peut (et doit) aider la bioinfo.
$\Rightarrow$ Incorporer des contraintes et/ou aspects probabilistes à nos modèles.

Conclusion

Échantillonnage statistique

- Bioalgorithmiciens ont redécouvert la méthode récursive (pondérée).
- Analyse en moyenne et optimisation (Boustrophedon).
- Évaluer sans échantillonnage $\Rightarrow$ Modif. automatisée de la spec. ?
- Boltzmann dans une distribution de Boltzmann ?

Shapes ARN (Collab. : W. A. Lorenz et P. Clote – Boston College)

- Beaucoup moins de shapes que de struct. secondaires !
  $\Rightarrow$ Énumération exhaustive réalisable ($n < 200$).
- Bijection entre les mots de Motzkin et les shapes π
- Modèle homopolymère surestime nombre de shapes
  Incorporer un modèle probabiliste pour possibilité d'appariement.
  Mais Bernoulli n'est pas suffisant . . .

En fournissant des outils pour décrire l'espace des conformations, la combinatoire peut (et doit) aider la bioinfo.
$\Rightarrow$ Incorporer des contraintes et/ou aspects probabilistes à nos modèles.

References I

A. Condon, B. Davy, B. Rastegari, S. Zhao, and F. Tarrant.

Classifying RNA pseudoknotted structures.
Theoretical Computer Science, 320(1) :35–50, 2004.

Y. Ding and E. Lawrence.

A statistical sampling algorithm for RNA secondary structure prediction.
Nucleic Acids Research, 31(24) :7280–7301, 2003.

P. Flajolet, P. Zimmermann, and B. Van Cutsem.

Calculus for the random generation of labelled combinatorial structures.
Theoretical Computer Science, 132 :1–35, 1994.

R. Giegerich, B. Voss, and M. Rehmsmeier.

Abstract shapes of RNA.
Nucleic Acids Res., 32(16) :4843–4851, 2004.

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.

N. Leontis and E. Westhof.

Geometric nomenclature and classification of RNA base pairs.
RNA, 7 :499–512, 2001.

J.S. McCaskill.

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

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.

References II

M. S. Waterman.

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

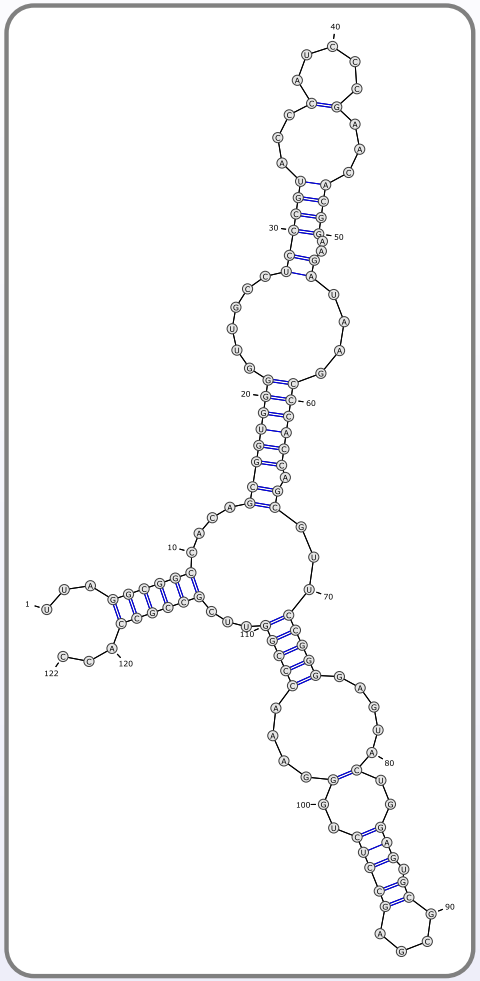
M. Zuker and D. Sankoff.

Rna secondary structures and their prediction.
Bull Math Bio, 46 :591–621, 1984.

Modèle de Turner

Basée sur décomposition non-ambiguë en boucles de la structure 2^{aire} :

- Boucles internes
- Renflements
- Boucles terminales
- Boucles multiples
- Empilements

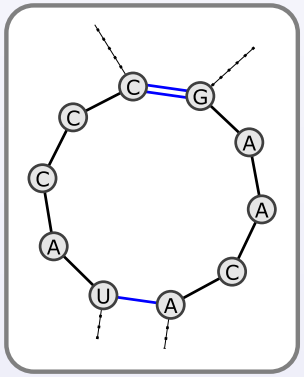


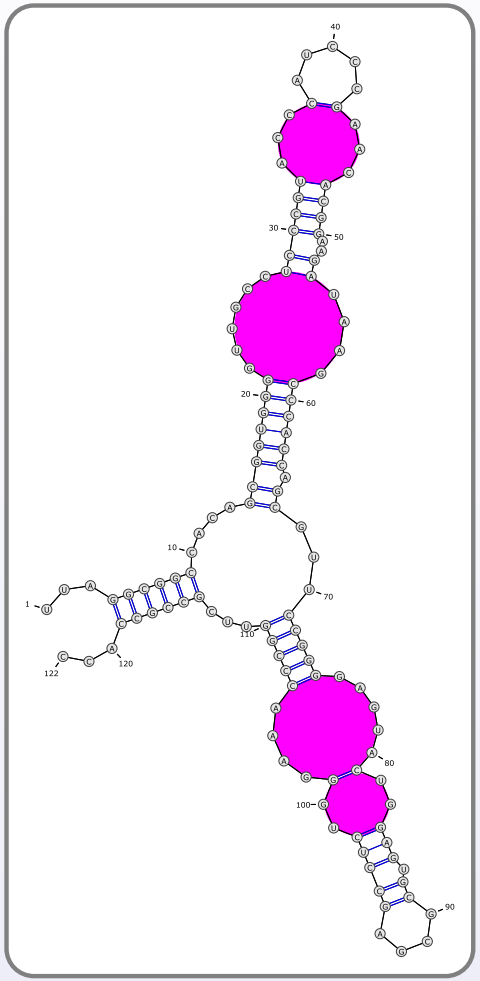
Énergies libres des boucles déterminées expérimentalement
+ Interpolation pour les grandes boucles

Modèle de Turner

Basée sur décomposition non-ambiguë en boucles de la structure 2^{aire} :

- Boucles internes
- Renflements
- Boucles terminales
- Boucles multiples
- Empilements



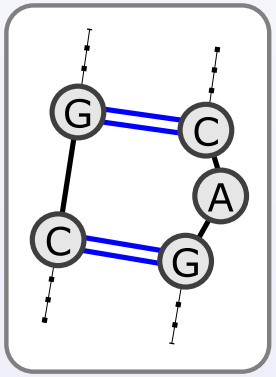


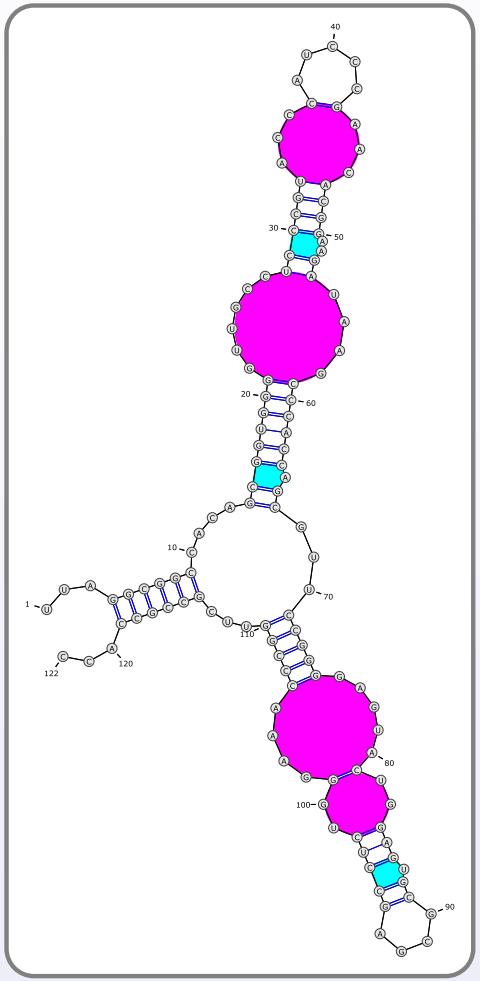
Énergies libres des boucles déterminées expérimentalement
+ Interpolation pour les grandes boucles

Modèle de Turner

Basée sur décomposition non-ambiguë en boucles de la structure 2^{aire} :

- Boucles internes
- Renflements
- Boucles terminales
- Boucles multiples
- Empilements



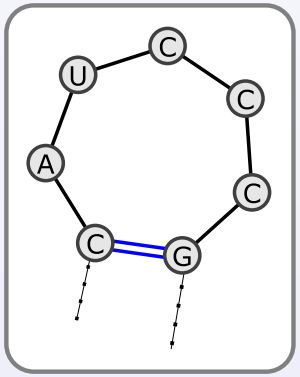


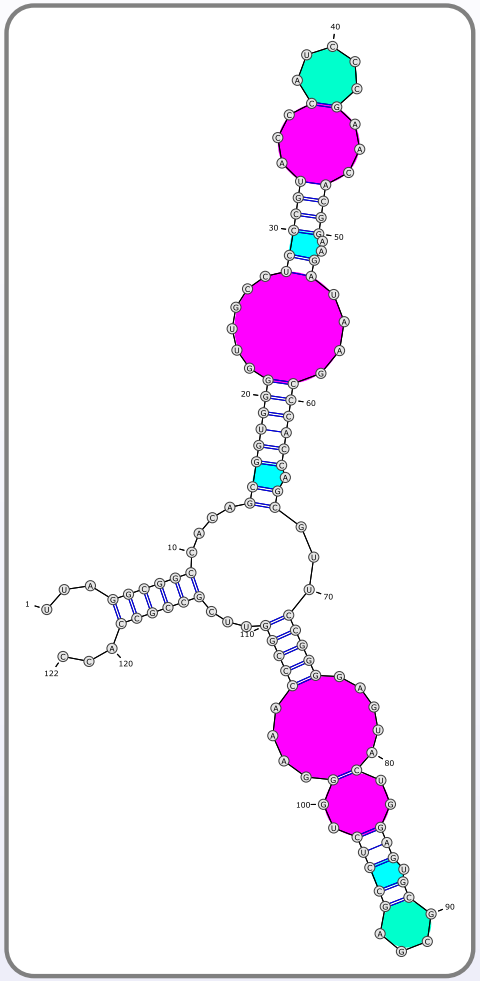
Énergies libres des boucles déterminées expérimentalement
+ Interpolation pour les grandes boucles

Modèle de Turner

Basée sur décomposition non-ambiguë en boucles de la structure 2^{aire} :

- Boucles internes
- Renflements
- Boucles terminales
- Boucles multiples
- Empilements



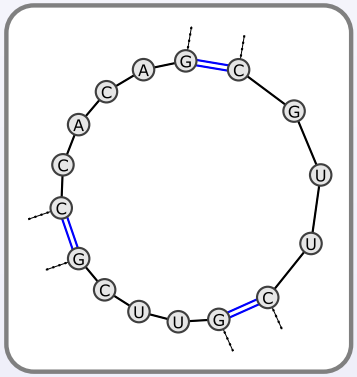


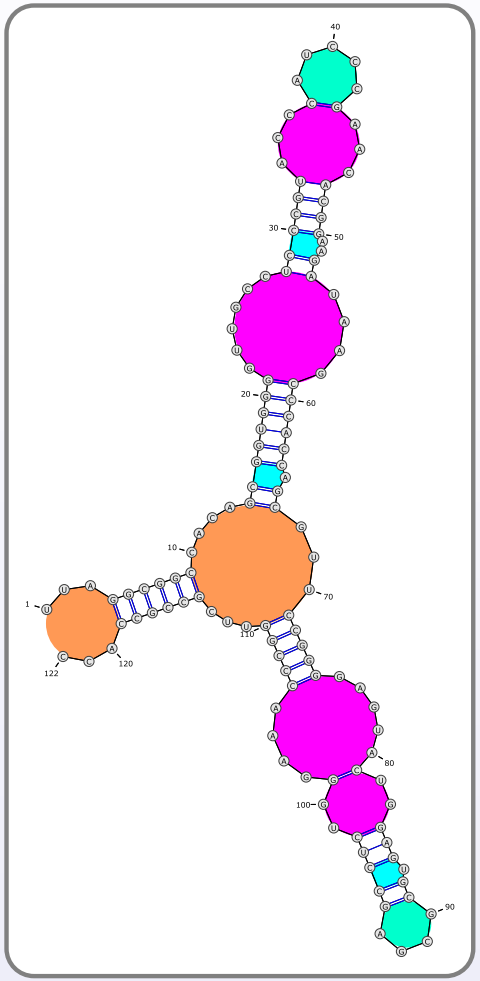
Énergies libres des boucles déterminées expérimentalement
+ Interpolation pour les grandes boucles

Modèle de Turner

Basée sur décomposition non-ambiguë en boucles de la structure 2^{aire} :

- Boucles internes
- Renflements
- Boucles terminales
- Boucles multiples
- Empilements



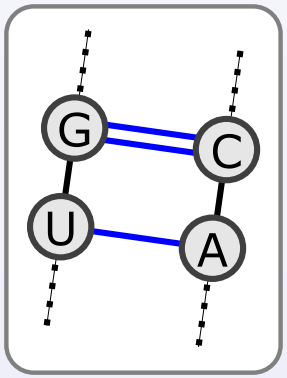


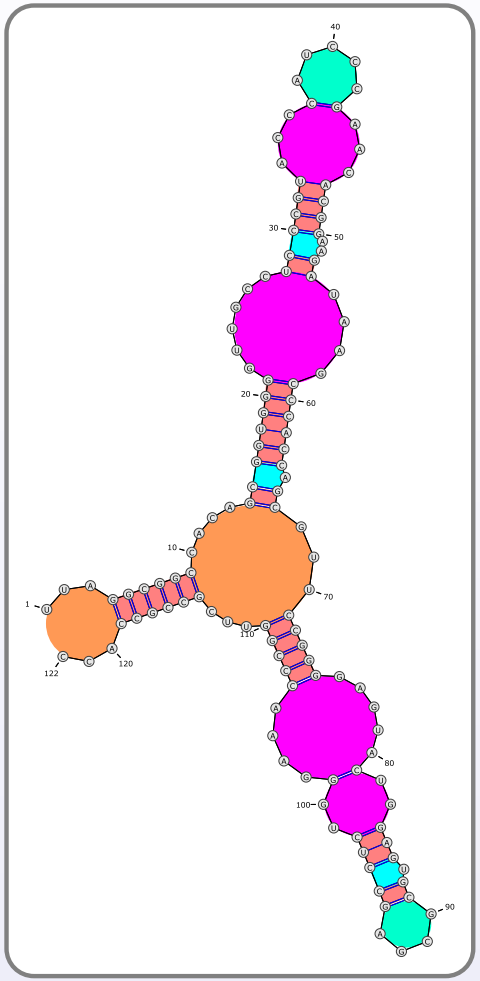
Énergies libres des boucles déterminées expérimentalement
+ Interpolation pour les grandes boucles

Modèle de Turner

Basée sur décomposition non-ambiguë en boucles de la structure 2^{aire} :

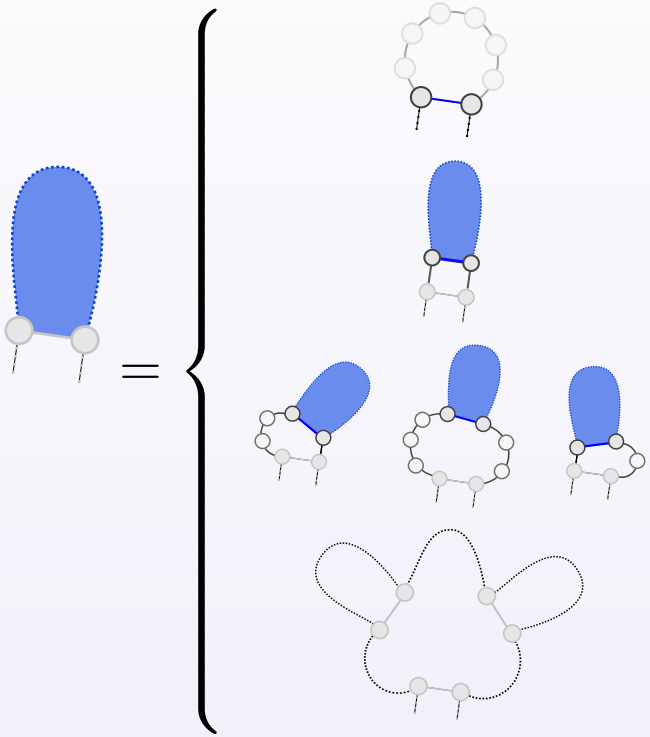
- Boucles internes
- Renflements
- Boucles terminales
- Boucles multiples
- Empilements



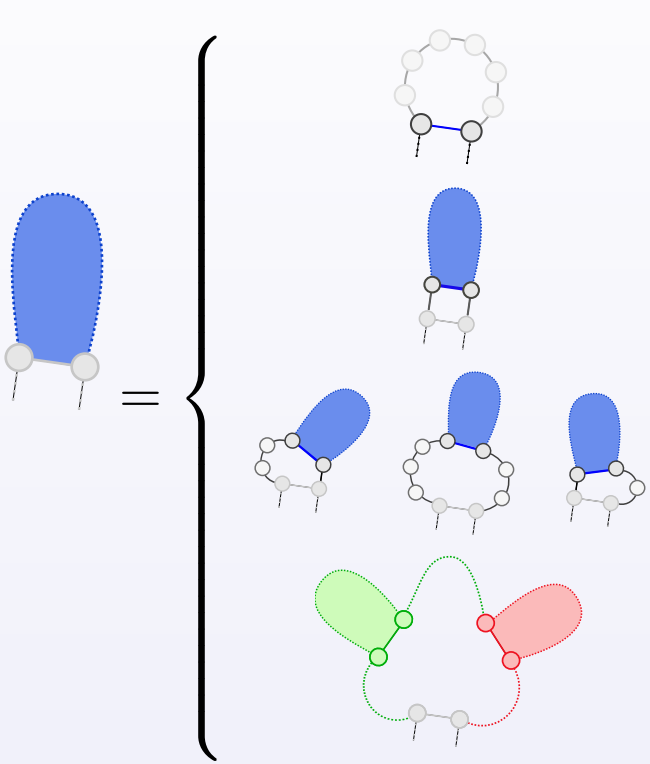


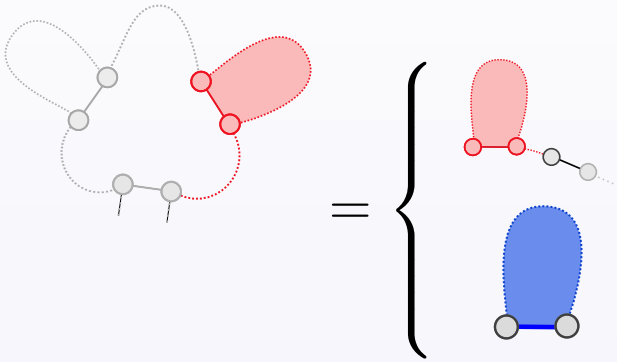
Énergies libres des boucles déterminées expérimentalement
+ Interpolation pour les grandes boucles

MFE DP equations

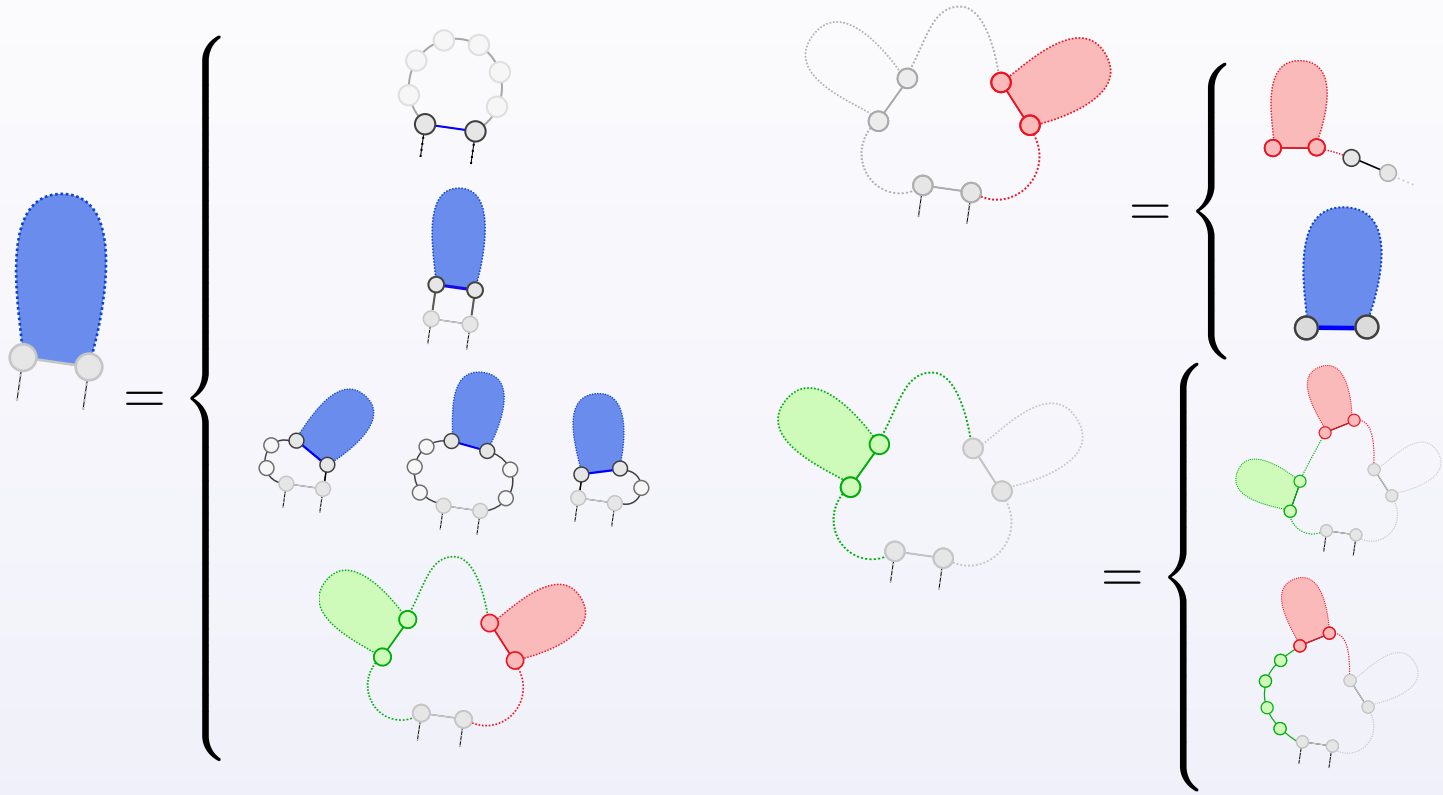


MFE DP equations



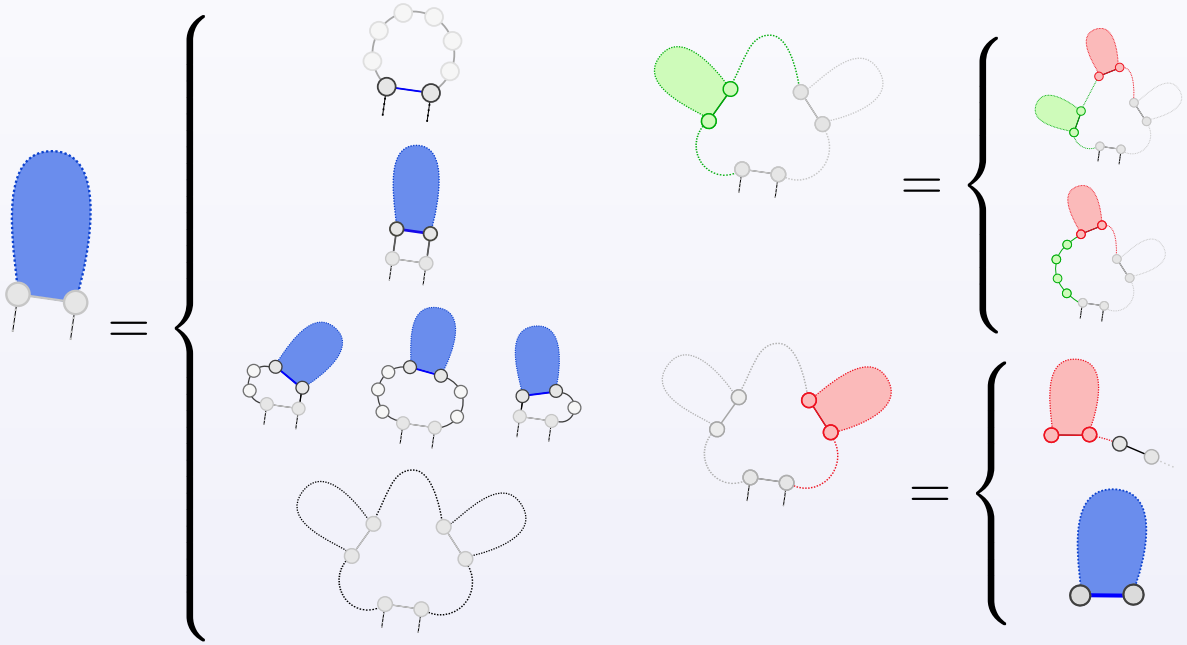


MFE DP equations



Unambiguity : Not too hard to check !
Completeness $\Rightarrow$ Enumerative combinatorics.

Validation



Validation

Generating function $\mathcal{T}(z) = \sum_{n \geq 0} t_n z^n$
With t_n = #Secondary structures of size n

$$\mathcal{A}(z) = \begin{cases} \mathcal{S}(z) \\ z^2\mathcal{A}(z) \\ z\mathcal{S}(z)z^2\mathcal{A}(z) + z^2\mathcal{A}(z)\mathcal{S}(z)z \\ +z\mathcal{S}(z)z^2\mathcal{A}(z)\mathcal{S}(z)z \\ \mathcal{B}(z)\mathcal{C}(z) \end{cases} \qquad \mathcal{B}(z) = \begin{cases} \mathcal{B}(z)\mathcal{C}(z) \\ \mathcal{S}(z)\mathcal{B}(z) \end{cases} \qquad \mathcal{C}(z) = \begin{cases} \mathcal{C}(z)z \\ z^2\mathcal{A}(z) \end{cases}$$

$$\mathcal{S}(z) = 1 + z\mathcal{S}(z)$$

Validation

$$\mathcal{A}(z) = \begin{cases} \mathcal{S}(z) \\ z^2\mathcal{A}(z) \\ z\mathcal{S}(z)z^2\mathcal{A}(z) + z^2\mathcal{A}(z)\mathcal{S}(z)z \\ +z\mathcal{S}(z)z^2\mathcal{A}(z)\mathcal{S}(z)z \\ \mathcal{B}(z)\mathcal{C}(z) \end{cases} \qquad \mathcal{B}(z) = \begin{cases} \mathcal{B}(z)\mathcal{C}(z) \\ \mathcal{S}(z)\mathcal{B}(z) \end{cases} \qquad \mathcal{C}(z) = \begin{cases} \mathcal{C}(z)z \\ z^2\mathcal{A}(z) \end{cases}$$

$$\mathcal{S}(z) = 1 + z\mathcal{S}(z)$$

Reminder : Waterman counted Sec. Str. [Wat78] and found the gen. fun.

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

Here we have

$$\begin{aligned} \Rightarrow \mathcal{A}(z) &= \frac{1 - z - z^2 - \sqrt{1 - 2z - z^2 - 2z^3 + z^4}}{2z^2} \\ &= \mathcal{W}(z) - 1 \quad \text{(Woops, we forgot the } \textit{empty} \text{ RNA)} \end{aligned}$$