Échantillonnage non-redondant de l'ensemble de Boltzmann

Danièle Gardi[‡] Andy Lorenz[†] Yann Ponty[⋆]

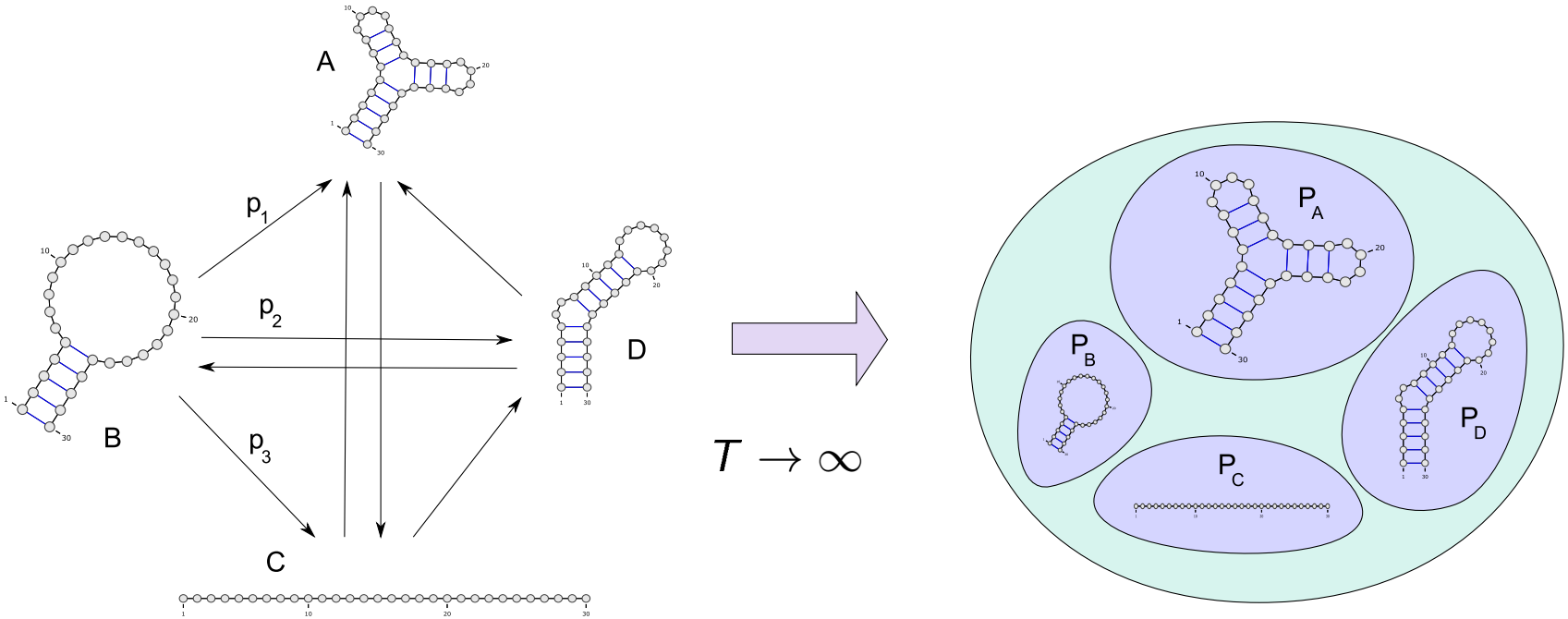
‡ Université Versailles St Quentin – France

† Boston College – Boston – USA

⋆ Polytechnique/CNRS/INRIA AMIB – France

26 Février 2010

Fondements du repliement thermodynamique



Convergence vers une distribution stationnaire de probabilité, l'équilibre de Boltzmann, où la probabilité est exponentiellement faible sur l'énergie libre.

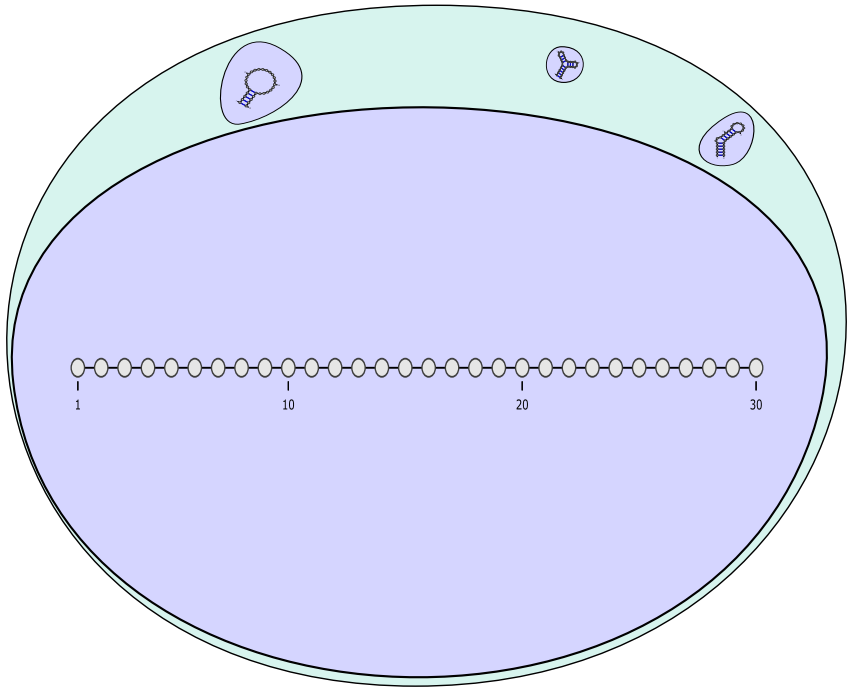
Problèmes soulevés :

Étant donné des modèles pour l'ensemble des conformations et l'énergie libre.

A. Déterminer la structure la plus probable (= Energie libre min.) à l'équilibre

B. Déterminer des propriétés moyennes de l'ensemble de Boltzmann

Hors de l'équilibre



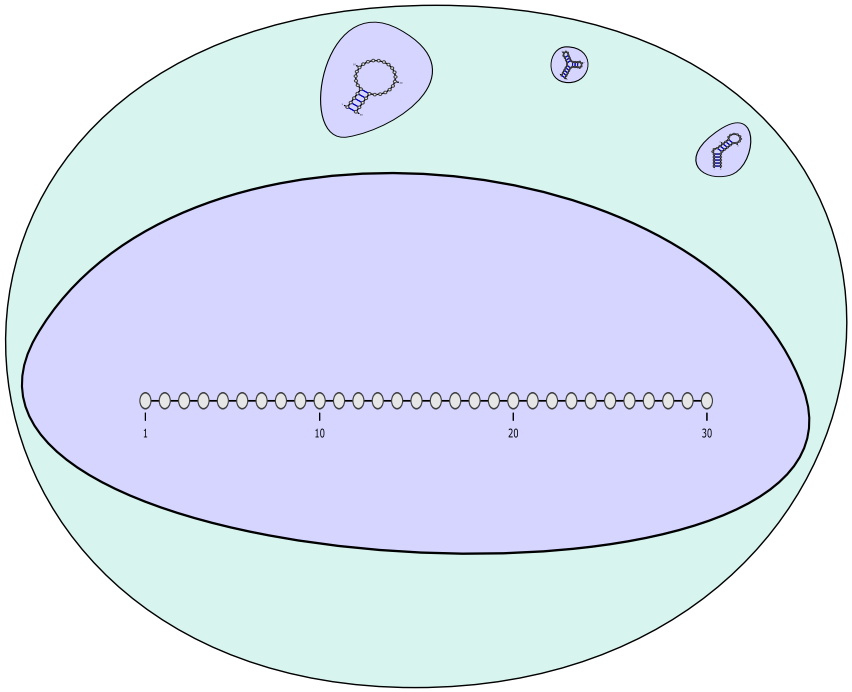
$T = 0$

Mais majorité des transcrits dégradés avant 7h (Org.: Souris [SSN+09]).

Équilibre de Boltzmann ⇐ repliements rapides.
Ne permet pas l'étudier des phénomènes cinétiques (Repliement co-transcriptionnel, états transients ou *riboswitches*).

A. Déterminer la structure la plus probable (= Energie libre min.) à l'équilibre
B. Déterminer des propriétés moyennes de l'ensemble de Boltzmann
C. Déterminer la structure la plus probable à temps T.
   (c.f. H. Isambert par simulation, NP-complet en déterministe [MTSC09])

Hors de l'équilibre



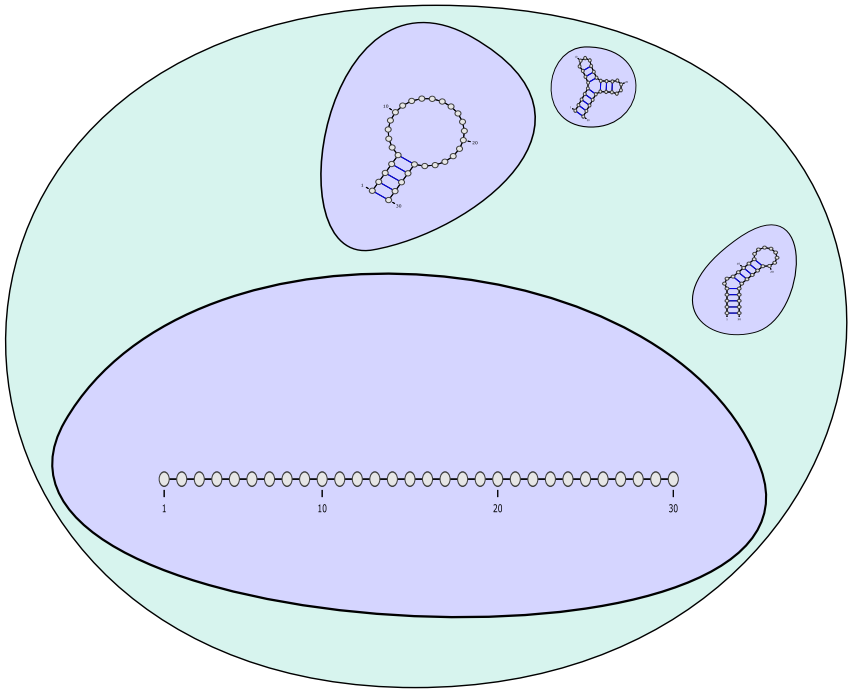
$T = 1h$

Mais majorité des transcrits dégradés avant 7h (Org.: Souris [SSN+09]).

Équilibre de Boltzmann ⇐ repliements rapides.
Ne permet pas l'étudier des phénomènes cinétiques (Repliement co-transcriptionnel, états transients ou *riboswitches*).

A. Déterminer la structure la plus probable (= Energie libre min.) à l'équilibre
B. Déterminer des propriétés moyennes de l'ensemble de Boltzmann
C. Déterminer la structure la plus probable à temps T.
   (c.f. H. Isambert par simulation, NP-complet en déterministe [MTSC09])

Hors de l'équilibre



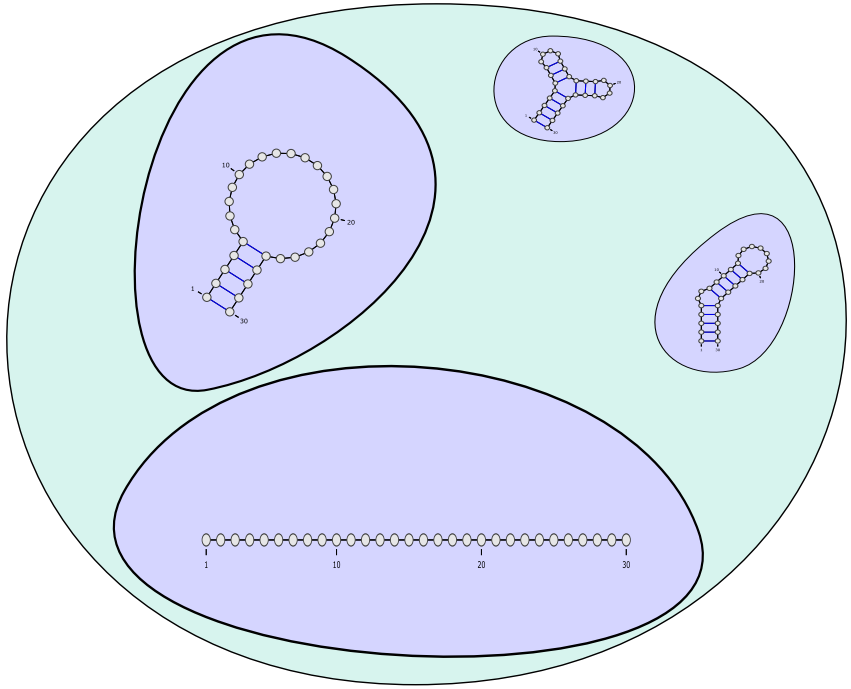
$T = 2h$

Mais majorité des transcrits dégradés avant 7h (Org.: Souris [SSN+09]).

Équilibre de Boltzmann $\Leftarrow$ repliements rapides.
Ne permet pas l'étudier des phénomènes cinétiques (Repliement co-transcriptionnel, états transients ou *riboswitches*).

A. Déterminer la structure la plus probable (= Energie libre min.) à l'équilibre
B. Déterminer des propriétés moyennes de l'ensemble de Boltzmann
C. Déterminer la structure la plus probable à temps T.
   (c.f. H. Isambert par simulation, NP-complet en déterministe [MTSC09])

Hors de l'équilibre



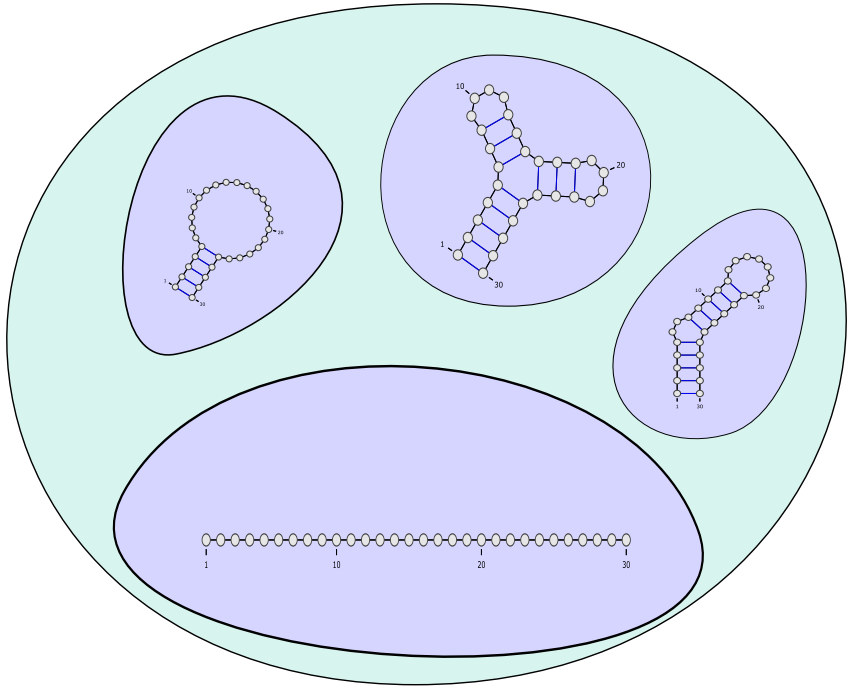
$T = 5h$

Mais majorité des transcrits dégradés avant 7h (Org.: Souris [SSN+09]).

Équilibre de Boltzmann $\Leftarrow$ repliements rapides.
Ne permet pas l'étudier des phénomènes cinétiques (Repliement co-transcriptionnel, états transients ou *riboswitches*).

A. Déterminer la structure la plus probable (= Energie libre min.) à l'équilibre
B. Déterminer des propriétés moyennes de l'ensemble de Boltzmann
C. Déterminer la structure la plus probable à temps T.
   (c.f. H. Isambert par simulation, NP-complet en déterministe [MTSC09])

Hors de l'équilibre



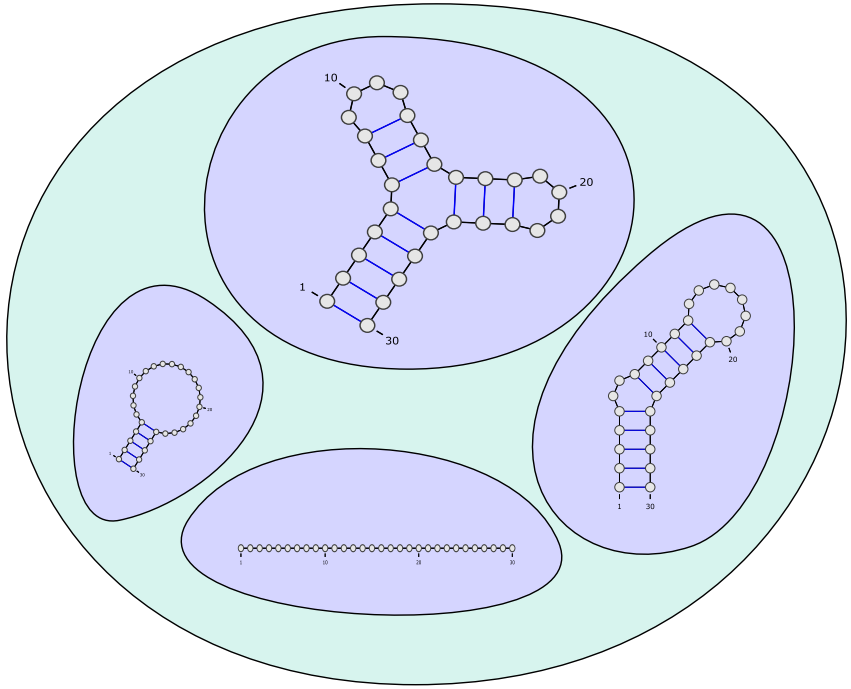
$T = 10h$

Mais majorité des transcrits dégradés avant 7h (Org.: Souris [SSN+09]).

Équilibre de Boltzmann $\Leftarrow$ repliements rapides.
Ne permet pas l'étudier des phénomènes cinétiques (Repliement co-transcriptionnel, états transients ou *riboswitches*).

A. Déterminer la structure la plus probable (= Energie libre min.) à l'équilibre
B. Déterminer des propriétés moyennes de l'ensemble de Boltzmann
C. Déterminer la structure la plus probable à temps T.
   (c.f. H. Isambert par simulation, NP-complet en déterministe [MTSC09])

Hors de l'équilibre



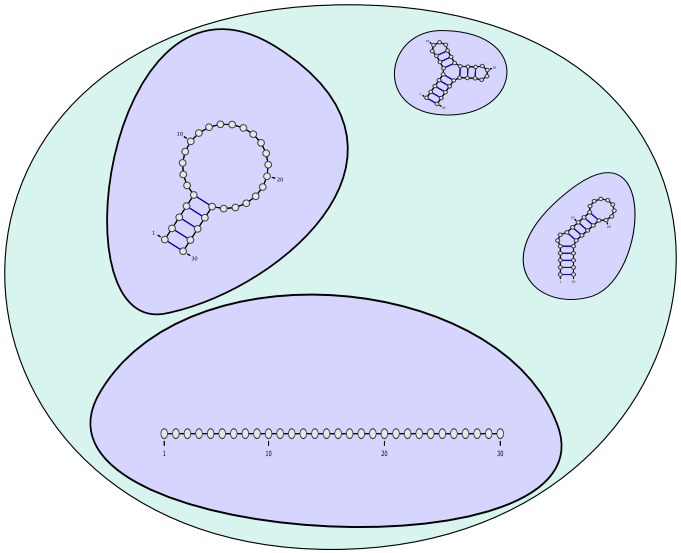
$$T \to \infty$$

Mais majorité des transcrits dégradés avant 7h (Org.: Souris [SSN+09]).

Équilibre de Boltzmann ⇐ repliements rapides.
Ne permet pas l'étudier des phénomènes cinétiques (Repliement co-transcriptionnel, états transients ou *riboswitches*).

A. Déterminer la structure la plus probable (= Energie libre min.) à l'équilibre
B. Déterminer des propriétés moyennes de l'ensemble de Boltzmann
C. Déterminer la structure la plus probable à temps T.
   (c.f. H. Isambert par simulation, NP-complet en déterministe [MTSC09])

Hors de l'équilibre



$T = 10h$

Mais majorité des transcrits dégradés avant 7h (Org.: Souris [SSN+09]).

Équilibre de Boltzmann $\Leftarrow$ repliements rapides.
Ne permet pas l'étudier des phénomènes cinétiques (Repliement co-transcriptionnel, états transients ou *riboswitches*).

A. Déterminer la structure la plus probable (= Energie libre min.) à l'équilibre
B. Déterminer des propriétés moyennes de l'ensemble de Boltzmann
C. Déterminer la structure la plus probable à temps T.
   (c.f. H. Isambert par simulation, NP-complet en déterministe [MTSC09])

Approche thermodynamique :

Modèle d'énergie

\+ Espace de conformations

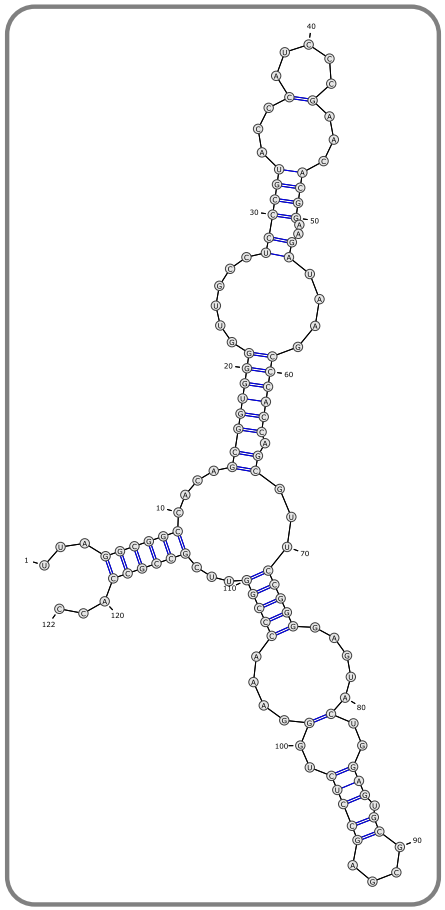
Une vision combinatoire aide à :

- Énumérer l'ensemble des conformations
- Extraire des propriétés moyennes de l'ensemble de Boltzmann
- Analyser les algorithmes liés à la prédiction . . .
- . . . et les améliorer.

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 Δ G des boucles dépendent des bases, assymétrie, bases *libres* (dangle) ...

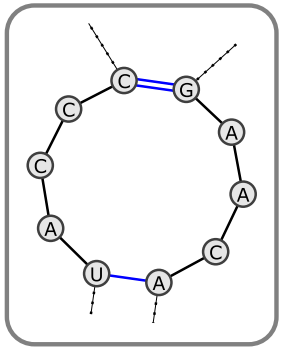
Déterminées expérimentalement
+ Interpolation pour grandes boucles.

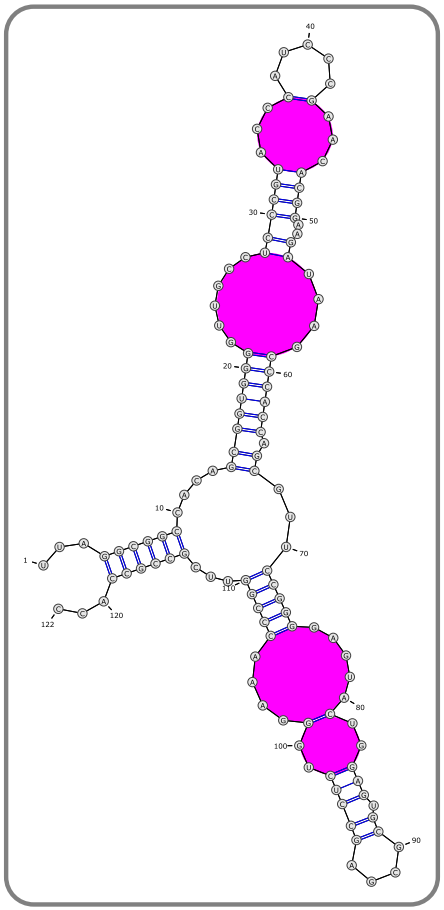
Meilleure résultats grâce à la prise en compte de l'empilement.

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 Δ G des boucles dépendent des bases, assymétrie, bases *libres* (dangle) ...

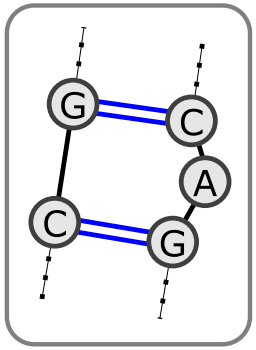
Déterminées expérimentalement
+ Interpolation pour grandes boucles.

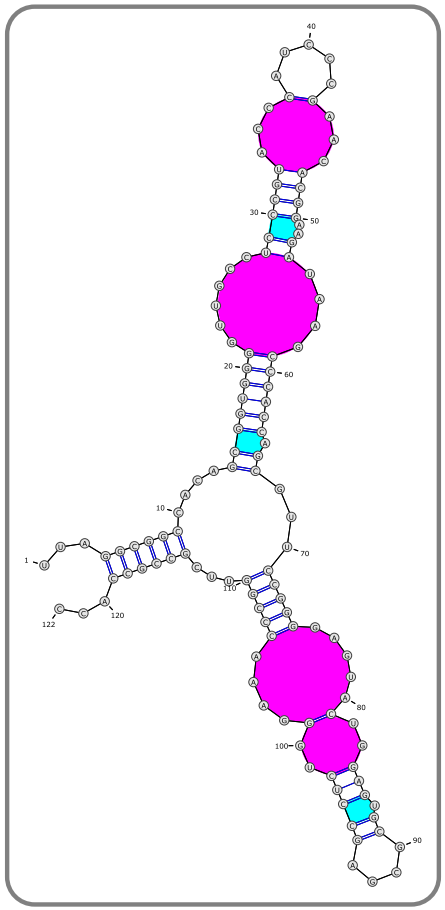
Meilleure résultats grâce à la prise en compte de l'empilement.

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 Δ G des boucles dépendent des bases, assymétrie, bases *libres* (dangle) ...

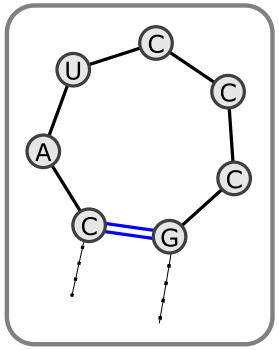
Déterminées expérimentalement
+ Interpolation pour grandes boucles.

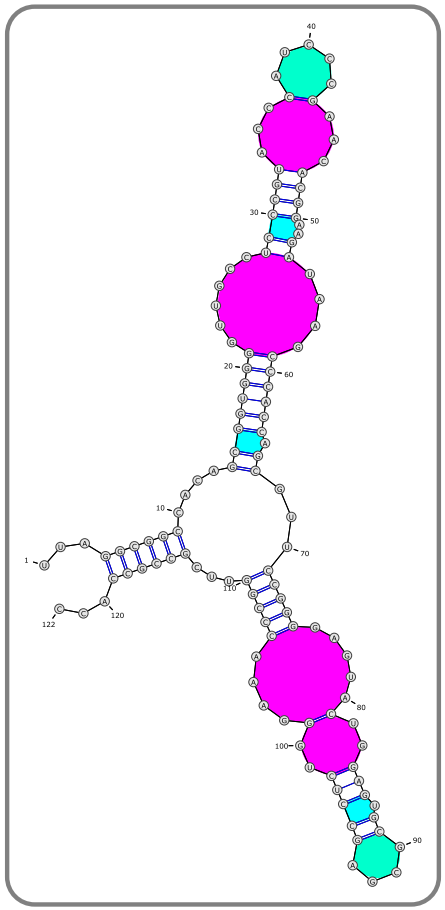
Meilleure résultats grâce à la prise en compte de l'empilement.

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 Δ G des boucles dépendent des bases, assymétrie, bases *libres* (dangle) ...

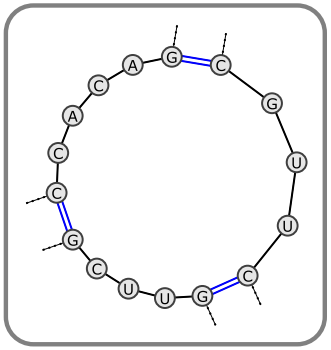
Déterminées expérimentalement
+ Interpolation pour grandes boucles.

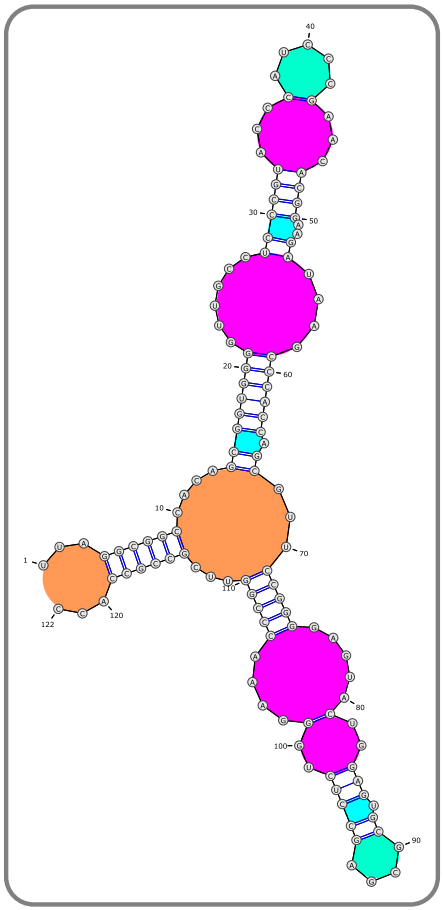
Meilleure résultats grâce à la prise en compte de l'empilement.

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 Δ G des boucles dépendent des bases, assymétrie, bases *libres* (dangle) ...

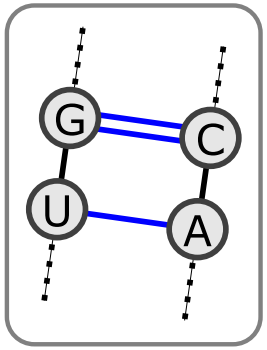
Déterminées expérimentalement
+ Interpolation pour grandes boucles.

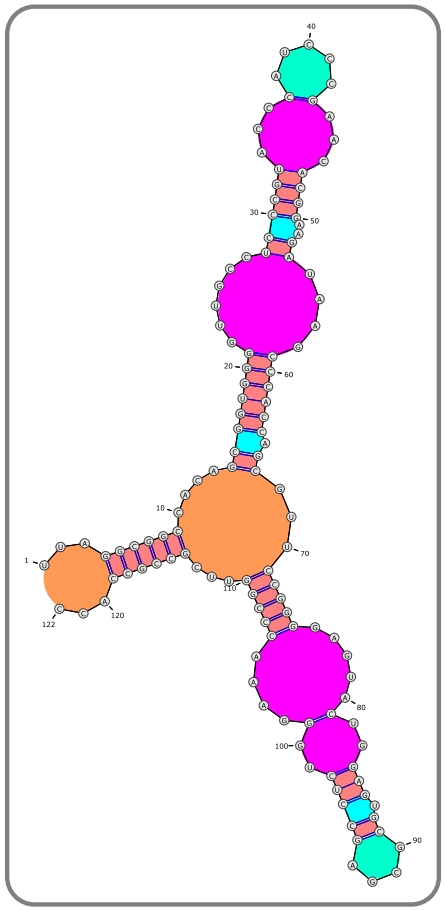
Meilleure résultats grâce à la prise en compte de l'empilement.

Modèle de Turner

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

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



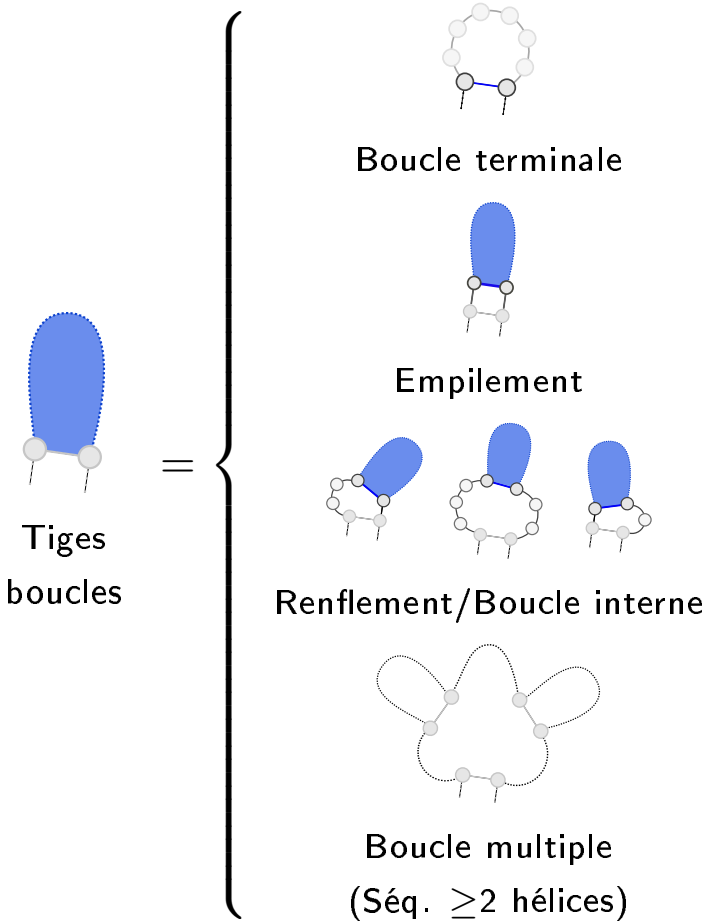


Énergies libres Δ G des boucles dépendent des bases, assymétrie, bases *libres* (dangle) ...

Déterminées expérimentalement
+ Interpolation pour grandes boucles.

Meilleure résultats grâce à la prise en compte de l'empilement.

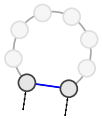
MFE DP equations



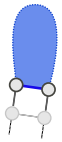
MFE DP equations

Tiges boucles

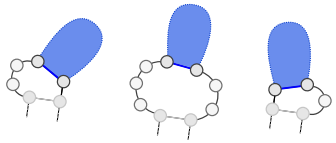
$=$



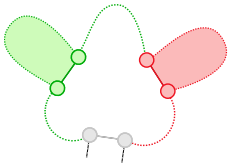
Boucle terminale



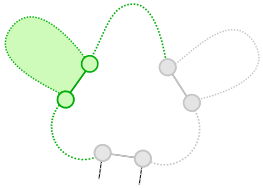
Empilement



Renflement/Boucle interne

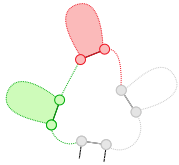


Boucle multiple
(Séq. $\geq$2 hélices)



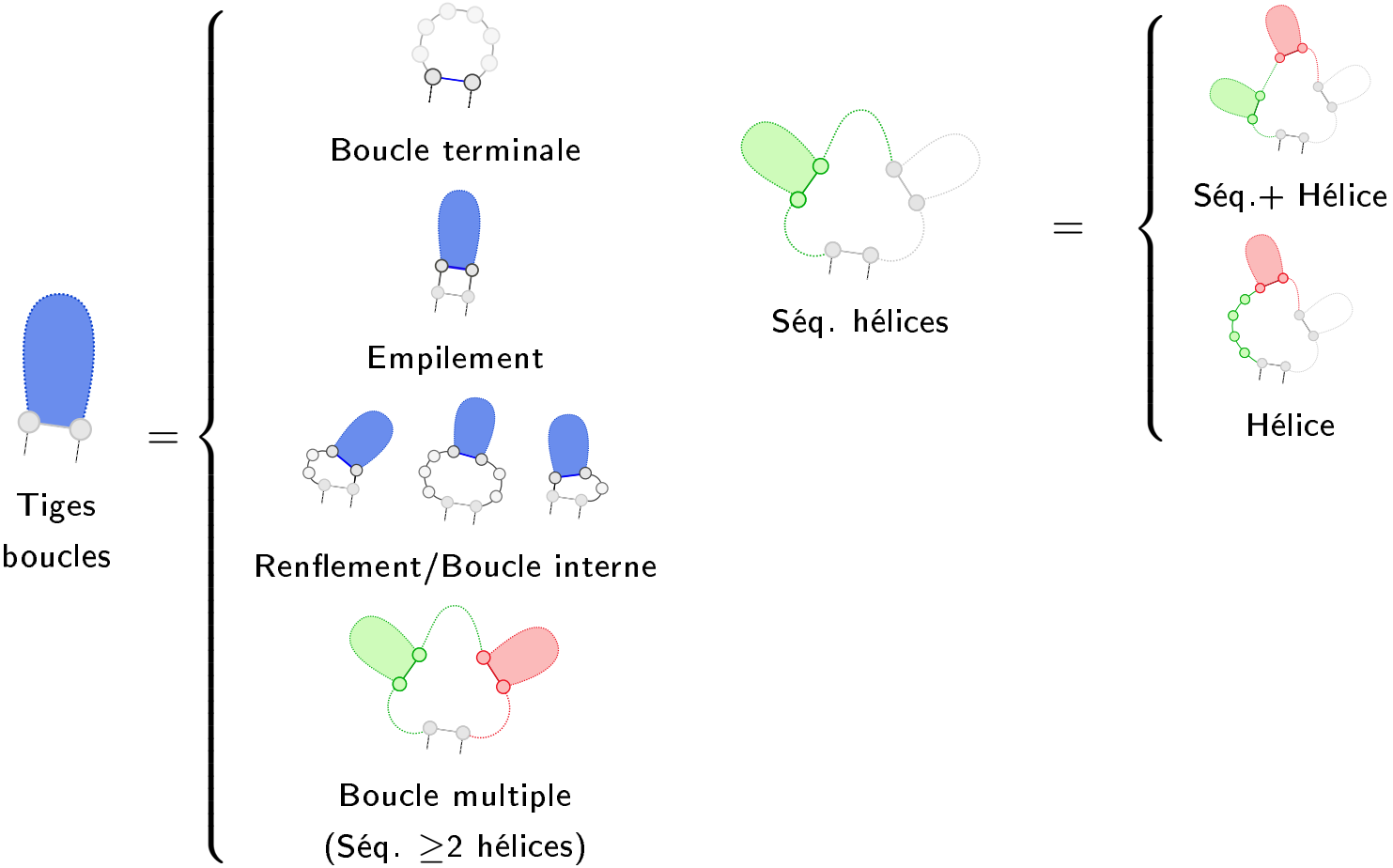
Séq. hélices

$=$

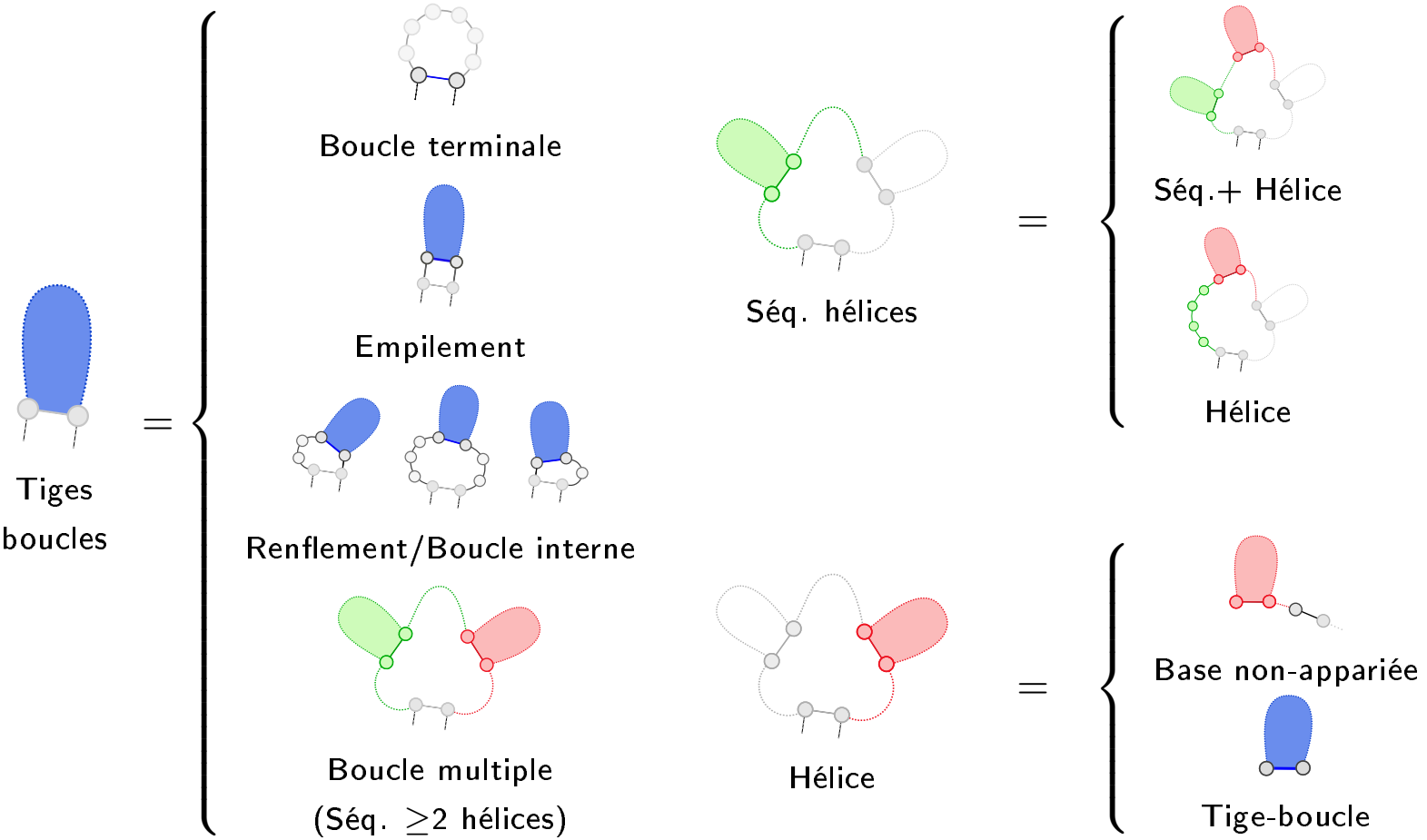


Séq.+ Hélice

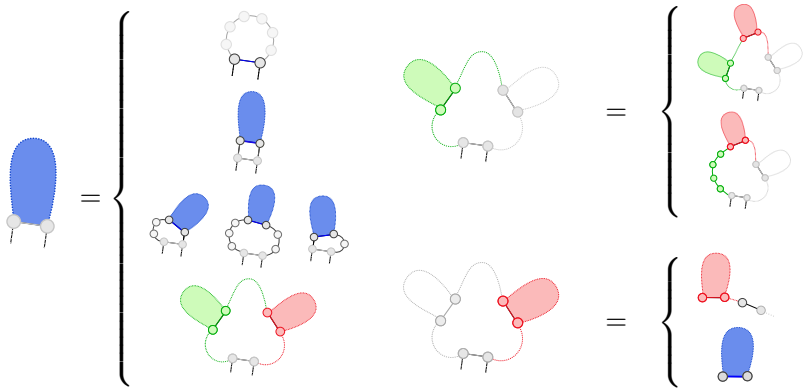
MFE DP equations



MFE DP equations



- $E_H(i,j)$: Energie de boucle terminale *fermée* par une paire (i,j)
- $E_{BI}(i,j)$: Energie de renflement ou boucle interne *fermée* par une paire (i,j)
- $E_S(i,j)$: Energie d'empilement $(i,j)/(i+1,j-1)$
- a,c,b: Pénalité de boucle multiple, hélice et non-appariées dans multiboucle.



Calcul des matrices

$$
\begin{aligned}
\mathcal{M}'_{i,j} &= \min \begin{cases} E_H(i,j) \\ E_S(i,j) + \mathcal{M}'_{i+1,j-1} \\ \mathrm{Min}_{i',j'}(E_{BI}(i,i',j',j) + \mathcal{M}'_{i',j'}) \\ a + c + \mathrm{Min}_k(\mathcal{M}_{i+1,k-1} + \mathcal{M}^1{}_{k,j-1}) \end{cases} \\
\mathcal{M}_{i,j} &= \mathrm{Min}_k \left\{ \min\left(\mathcal{M}_{i,k-1}, b(k-1)\right) + \mathcal{M}^1{}_{k,j} \right\} \\
\mathcal{M}^1{}_{i,j} &= \mathrm{Min}_k \left\{ b + \mathcal{M}^1{}_{i,j-1}, c + \mathcal{M}'_{i,j} \right\}
\end{aligned}
$$

Fonction de partition/Probabilité de Boltzmann

- Soit ω une séquence d'ARN
- et $\mathcal{S}_\omega$ l'ensemble des structures secondaires compatibles avec ω,

$$\textbf{Fonction de partition} \quad \mathcal{Z}_\omega = \sum_{S \in \mathcal{S}_\omega} e^{\frac{-E_{S,\omega}}{RT}}$$

où T est la température en Kelvin et R la constante des gaz parfaits.

$$\textbf{Probabilité de Boltzmann} \quad P_{S,\omega} = \frac{e^{\frac{-E_{S,\omega}}{RT}}}{\mathcal{Z}_\omega}$$

$P_{S,\omega}$ est la probabilité d'observer ω dans une conformation S.

$\Rightarrow$ Offre une vision moins statique du repliement

$\Rightarrow$ Fournit un modèle pour calculer des statistiques (BPs, motifs ...)

$\Rightarrow$ Unifie la génération de sous-optimaux et la minimisation ($RT \to \infty$)

$\Rightarrow$ Incorporation triviale dans des équations de programmation dynamique non-ambiguës

Fonction de partition/Probabilité de Boltzmann

- Soit ω une séquence d'ARN
- et $\mathcal{S}_\omega$ l'ensemble des structures secondaires compatibles avec ω,

$$\textbf{Fonction de partition} \quad \mathcal{Z}_\omega = \sum_{S \in \mathcal{S}_\omega} e^{\frac{-E_{S,\omega}}{RT}}$$

où T est la température en Kelvin et R la constante des gaz parfaits.

$$\textbf{Probabilité de Boltzmann} \quad P_{S,\omega} = \frac{e^{\frac{-E_{S,\omega}}{RT}}}{\mathcal{Z}_\omega}$$

$P_{S,\omega}$ est la probabilité d'observer ω dans une conformation S.

$\Rightarrow$ Offre une vision moins statique du repliement

$\Rightarrow$ Fournit un modèle pour calculer des statistiques (BPs, motifs ...)

$\Rightarrow$ Unifie la génération de sous-optimaux et la minimisation ($RT \to \infty$)

$\Rightarrow$ Incorporation triviale dans des équations de programmation dynamique non-ambiguës

Fonction de partition/Probabilité de Boltzmann

- Soit ω une séquence d'ARN
- et $\mathcal{S}_\omega$ l'ensemble des structures secondaires compatibles avec ω,

$$\textbf{Fonction de partition} \quad \mathcal{Z}_\omega = \sum_{S \in \mathcal{S}_\omega} e^{\frac{-E_{S,\omega}}{RT}}$$

où T est la température en Kelvin et R la constante des gaz parfaits.

$$\textbf{Probabilité de Boltzmann} \quad P_{S,\omega} = \frac{e^{\frac{-E_{S,\omega}}{RT}}}{\mathcal{Z}_\omega}$$

$P_{S,\omega}$ est la probabilité d'observer ω dans une conformation S.

$\Rightarrow$ Offre une vision moins statique du repliement

$\Rightarrow$ Fournit un modèle pour calculer des statistiques (BPs, motifs ...)

$\Rightarrow$ Unifie la génération de sous-optimaux et la minimisation ($RT \to \infty$)

$\Rightarrow$ Incorporation triviale dans des équations de programmation dynamique non-ambiguës

Fonction de partition/Probabilité de Boltzmann

- Soit ω une séquence d'ARN
- et $\mathcal{S}_\omega$ l'ensemble des structures secondaires compatibles avec ω,

$$\textbf{Fonction de partition} \quad \mathcal{Z}_\omega = \sum_{S \in \mathcal{S}_\omega} e^{\frac{-E_{S,\omega}}{RT}}$$

où T est la température en Kelvin et R la constante des gaz parfaits.

$$\textbf{Probabilité de Boltzmann} \quad P_{S,\omega} = \frac{e^{\frac{-E_{S,\omega}}{RT}}}{\mathcal{Z}_\omega}$$

$P_{S,\omega}$ est la probabilité d'observer ω dans une conformation S.

$\Rightarrow$ Offre une vision moins statique du repliement

$\Rightarrow$ Fournit un modèle pour calculer des statistiques (BPs, motifs ...)

$\Rightarrow$ Unifie la génération de sous-optimaux et la minimisation ($RT \to \infty$)

$\Rightarrow$ Incorporation triviale dans des équations de programmation dynamique non-ambiguës

Fonction de partition/Probabilité de Boltzmann

- Soit ω une séquence d'ARN
- et $\mathcal{S}_\omega$ l'ensemble des structures secondaires compatibles avec ω,

$$\textbf{Fonction de partition} \quad \mathcal{Z}_\omega = \sum_{S \in \mathcal{S}_\omega} e^{\frac{-E_{S,\omega}}{RT}}$$

où T est la température en Kelvin et R la constante des gaz parfaits.

$$\textbf{Probabilité de Boltzmann} \quad P_{S,\omega} = \frac{e^{\frac{-E_{S,\omega}}{RT}}}{\mathcal{Z}_\omega}$$

$P_{S,\omega}$ est la probabilité d'observer ω dans une conformation S.

$\Rightarrow$ Offre une vision moins statique du repliement

$\Rightarrow$ Fournit un modèle pour calculer des statistiques (BPs, motifs ...)

$\Rightarrow$ Unifie la génération de sous-optimaux et la minimisation ($RT \to \infty$)

$\Rightarrow$ Incorporation triviale dans des équations de programmation dynamique non-ambiguës

De la minimisation à la fonction de partition [McC90]:

- Contribution énergétique $E \rightarrow$ Facteur de Boltzmann $e^{\frac{-E}{RT}}$
- Contribution énergétique passent à l'exposant :
  $\Rightarrow$ Sommes $(+) \rightarrow$ Produits $(\times)$
- Sommer au lieu de minimiser : Min $\rightarrow$ Sommes $(\sum)$

$$
\begin{aligned}
\mathcal{M}'(i,j) &= \text{Min}\left\{\begin{array}{c} E_{\boldsymbol{H}}(i,j) \\ E_{\boldsymbol{S}}(i,j) + \mathcal{M}'(i+1,j-1) \\ \text{Min}(E_{\boldsymbol{BI}}(i,i',j',j) + \mathcal{M}'(i',j')) \\ a + c + \text{Min}\left(\mathcal{M}'(i+1,k-1) + \mathcal{M}^{\mathbf{1}}(k,j-1)\right) \end{array}\right\} \\
\mathcal{M}(i,j) &= \text{Min}\left\{\text{Min}\left(\mathcal{M}(i,k-1), b(k-1)\right) + \mathcal{M}^{\mathbf{1}}(k,j)\right\} \\
\mathcal{M}^{\mathbf{1}}(i,j) &= \text{Min}\left\{b + \mathcal{M}^{\mathbf{1}}(i,j-1), c + \mathcal{M}'(i,j)\right\}
\end{aligned}
$$

Message #2

Partant d'une décomposition **non-ambiguë** de l'espace des conformations, le calcul de la partition function est direct via un changement d'algèbre $(\text{Min}, +) \rightarrow (+, \times)$.

Calcul des probabilités d'appariement immédiat à partir de $\mathcal{Z}'$ (`RNAFold -p`).

De la minimisation à la fonction de partition [McC90]:

- Contribution énergétique $E \rightarrow$ Facteur de Boltzmann $e^{\frac{-E}{RT}}$
- Contribution énergétique passent à l'exposant :
  $\Rightarrow$ Sommes $(+) \rightarrow$ Produits $(\times)$
- Sommer au lieu de minimiser : Min $\rightarrow$ Sommes $(\sum)$

$$\mathcal{M}'(i,j) = \text{Min}\left\{\begin{array}{c} e^{\frac{-E_H(i,j)}{RT}} \\ e^{\frac{-E_S(i,j)}{RT}} + \mathcal{M}'(i+1,j-1) \\ \text{Min}\left(e^{\frac{-E_{BI}(i,i',j',j)}{RT}} + \mathcal{M}'(i',j')\right) \\ e^{\frac{-(a+c)}{RT}} + \text{Min}\left(\mathcal{M}'(i+1,k-1) + \mathcal{M}^1(k,j-1)\right) \end{array}\right\}$$

$$\mathcal{M}(i,j) = \text{Min}\left\{\text{Min}\left(\mathcal{M}(i,k-1), e^{\frac{-b(k-1)}{RT}}\right) + \mathcal{M}^1(k,j)\right\}$$

$$\mathcal{M}^1(i,j) = \text{Min}\left\{e^{\frac{-b}{RT}} + \mathcal{M}^1(i,j-1), e^{\frac{-c}{RT}} + \mathcal{M}'(i,j)\right\}$$

Message #2

Partant d'une décomposition **non-ambiguë** de l'espace des conformations, le calcul de la partition function est direct via un changement d'algèbre $(\text{Min}, +) \rightarrow (+, \times)$.

Calcul des probabilités d'appariement immédiat à partir de $\mathcal{Z}'$ (`RNAFold -p`).

De la minimisation à la fonction de partition [McC90]:

- Contribution énergétique $E \rightarrow$ Facteur de Boltzmann $e^{\frac{-E}{RT}}$
- Contribution énergétique passent à l'exposant :
  $\Rightarrow$ Sommes $(+) \rightarrow$ Produits $(\times)$
- Sommer au lieu de minimiser : Min $\rightarrow$ Sommes $(\sum)$

$$\mathcal{M}'(i,j) = \text{Min}\left\{\begin{array}{c} e^{\frac{-E_H(i,j)}{RT}} \\ e^{\frac{-E_S(i,j)}{RT}}\mathcal{M}'(i+1,j-1) \\ \text{Min}\left(e^{\frac{-E_{BI}(i,i',j',j)}{RT}}\mathcal{M}'(i',j')\right) \\ e^{\frac{-(a+c)}{RT}}\text{Min}\left(\mathcal{M}'(i+1,k-1)\mathcal{M}^1(k,j-1)\right) \end{array}\right\}$$

$$\mathcal{M}(i,j) = \text{Min}\left\{\text{Min}\left(\mathcal{M}(i,k-1), e^{\frac{-b(k-1)}{RT}}\right)\mathcal{M}^1(k,j)\right\}$$

$$\mathcal{M}^1(i,j) = \text{Min}\left\{e^{\frac{-b}{RT}}\mathcal{M}^1(i,j-1), e^{\frac{-c}{RT}}\mathcal{M}'(i,j)\right\}$$

Message #2

Partant d'une décomposition **non-ambiguë** de l'espace des conformations, le calcul de la partition function est direct via un changement d'algèbre $(\text{Min}, +) \rightarrow (+, \times)$.

Calcul des probabilités d'appariement immédiat à partir de $\mathcal{Z}'$ (`RNAFold -p`).

De la minimisation à la fonction de partition [McC90]:

- Contribution énergétique $E \rightarrow$ Facteur de Boltzmann $e^{\frac{-E}{RT}}$
- Contribution énergétique passent à l'exposant :
  $\Rightarrow$ Sommes $(+) \rightarrow$ Produits $(\times)$
- Sommer au lieu de minimiser : Min $\rightarrow$ Sommes $(\sum)$

$$
\begin{aligned}
\mathcal{Z}'(i,j) &= \sum \left\{ \begin{array}{c} e^{\frac{-E_H(i,j)}{RT}} + e^{\frac{-E_S(i,j)}{RT}} \mathcal{Z}'(i+1,j-1) \\ + \sum \left(e^{\frac{-E_{BI}(i,i',j',j)}{RT}} \mathcal{Z}'(i',j') \right) \\ + e^{\frac{-(a+c)}{RT}} \sum \left(\mathcal{Z}'(i+1,k-1) \mathcal{Z}^1(k,j-1) \right) \end{array} \right\} \\
\mathcal{Z}(i,j) &= \sum \left(\mathcal{Z}(i,k-1) + e^{\frac{-b(k-1)}{RT}} \right) \mathcal{Z}^1(k,j) \\
\mathcal{Z}^1(i,j) &= e^{\frac{-b}{RT}} \mathcal{Z}^1(i,j-1) + e^{\frac{-c}{RT}} \mathcal{Z}'(i,j)
\end{aligned}
$$

Message #2

Partant d'une décomposition **non-ambiguë** de l'espace des conformations, le calcul de la partition function est direct via un changement d'algèbre $(\text{Min}, +) \rightarrow (+, \times)$.

Calcul des probabilités d'appariement immédiat à partir de $\mathcal{Z}'$ (`RNAFold -p`).

De la minimisation à la fonction de partition [McC90]:

- Contribution énergétique $E \rightarrow$ Facteur de Boltzmann $e^{\frac{-E}{RT}}$
- Contribution énergétique passent à l'exposant :
  $\Rightarrow$ Sommes $(+) \rightarrow$ Produits $(\times)$
- Sommer au lieu de minimiser : Min $\rightarrow$ Sommes $(\sum)$

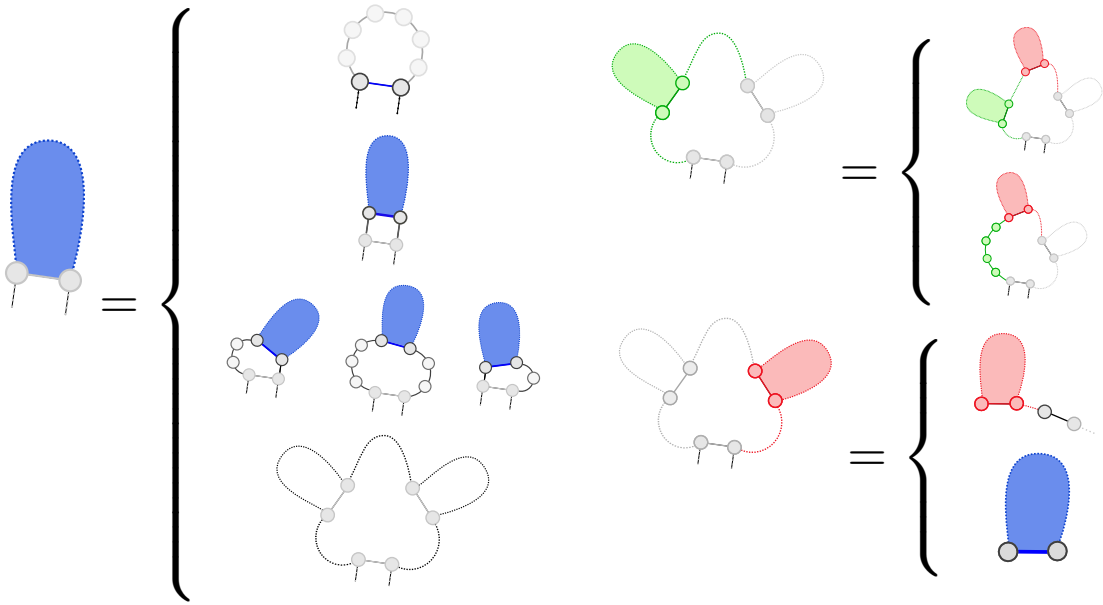
$$
\begin{aligned}
\mathcal{Z}'(i,j) &= \sum \left\{ \begin{array}{c} e^{\frac{-E_H(i,j)}{RT}} + e^{\frac{-E_S(i,j)}{RT}} \mathcal{Z}'(i+1,j-1) \\ + \sum \left(e^{\frac{-E_{BI}(i,i',j',j)}{RT}} \mathcal{Z}'(i',j') \right) \\ + e^{\frac{-(a+c)}{RT}} \sum \left(\mathcal{Z}'(i+1,k-1) \mathcal{Z}^1(k,j-1) \right) \end{array} \right\} \\
\mathcal{Z}(i,j) &= \sum \left(\mathcal{Z}(i,k-1) + e^{\frac{-b(k-1)}{RT}} \right) \mathcal{Z}^1(k,j) \\
\mathcal{Z}^1(i,j) &= e^{\frac{-b}{RT}} \mathcal{Z}^1(i,j-1) + e^{\frac{-c}{RT}} \mathcal{Z}'(i,j)
\end{aligned}
$$

Message #2

Partant d'une décomposition **non-ambiguë** de l'espace des conformations, le calcul de la partition function est direct via un changement d'algèbre $(\text{Min}, +) \rightarrow (+, \times)$.

Calcul des probabilités d'appariement immédiat à partir de $\mathcal{Z}'$ (`RNAFold -p`).

Problème : Est ce qu'une décomposition est non-ambiguë et complète ?
Non-ambiguité ok, mais complétude dur dur ...
⇒ Approche combinatoire ?



Problème : Est ce qu'une décomposition est non-ambiguë et complète ?
Non-ambiguité ok, mais complétude dur dur ...
$\Rightarrow$ Approche combinatoire ?

Série génératrice $\mathcal{T}(z) = \sum_{n \geq 0} t_n z^n$
Avec t_n = #Structures secondaires de taille n.

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

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

Problème : Est ce qu'une décomposition est non-ambiguë et complète ?

Non-ambiguité ok, mais complétude dur dur ...

$\Rightarrow$ Approche combinatoire ?

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

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

Rappel : Waterman a énuméré les str. sec. [Wat78] et trouvé la série génératrice

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

Ici, on obtient

$$\begin{aligned} \Rightarrow \mathcal{A}(z) &= \frac{1 - z - z^2 - \sqrt{1 - 2z - z^2 - 2z^3 + z^4}}{2z^2} \\ &= \mathcal{W}(z) - 1 \quad \text{(Arghh, on oublie la struct. sec. de taille 0 !)} \end{aligned}$$

Réécriture de l'algorithme `SFold` [DL03]:

1. Générer un nombre aléatoire dans $[0, \mathcal{Z}'(i,j))$
2. Retirer de r les contributions individuelles à $\mathcal{Z}'(i,j)$, jusqu'à $r < 0$
3. Réitérer récursivement sur les matrices des sous-intervalles

Idée : On choisit un cas ($\Leftrightarrow$ sous-ensemble de struct. sec.) proportionnellement à son facteur de Boltzmann total ($\Leftrightarrow$ fonction de partition).

$$\mathcal{Z}'(i,j) = \sum \left\{ \begin{array}{cc} e^{\frac{-E_H(i,j)}{RT}} + e^{\frac{-E_S(i,j)}{RT}} \mathcal{Z}'(i+1,j-1) & \text{A} \\ \sum \left(e^{\frac{-E_{BI}(i,i',j',j)}{RT}} \mathcal{Z}'(i',j') \right) & \text{B} \\ e^{\frac{-(a+c)}{RT}} \sum \left(\mathcal{Z}'(i+1,k-1) \mathcal{Z}^1(k,j-1) \right) & \text{C} \end{array} \right\}$$

Réécriture de l'algorithme `SFold` [DL03]:

1. Générer un nombre aléatoire dans $[0, \mathcal{Z}'(i,j))$
2. Retirer de r les contributions individuelles à $\mathcal{Z}'(i,j)$, jusqu'à $r < 0$
3. Réitérer récursivement sur les matrices des sous-intervalles

Idée : On choisit un cas ($\Leftrightarrow$ sous-ensemble de struct. sec.) proportionnellement à son facteur de Boltzmann total ($\Leftrightarrow$ fonction de partition).

$$\mathcal{Z}'(i,j) = \sum \; ??? \begin{cases} \rightarrow e^{\frac{-E_H(i,j)}{RT}} + e^{\frac{-E_S(i,j)}{RT}} \mathcal{Z}'(i+1,j-1) & \text{A} \\ \rightarrow \sum \left(e^{\frac{-E_{BI}(i,i',j',j)}{RT}} \mathcal{Z}'(i',j') \right) & \text{B} \\ \rightarrow e^{\frac{-(a+c)}{RT}} \sum \left(\mathcal{Z}'(i+1,k-1) \mathcal{Z}^1(k,j-1) \right) & \text{C} \end{cases}$$

Réécriture de l'algorithme `SFold` [DL03]:

1. Générer un nombre aléatoire dans $[0, \mathcal{Z}'(i,j))$
2. Retirer de r les contributions individuelles à $\mathcal{Z}'(i,j)$, jusqu'à $r < 0$
3. Réitérer récursivement sur les matrices des sous-intervalles

Idée : On choisit un cas ($\Leftrightarrow$ sous-ensemble de struct. sec.) proportionnellement à son facteur de Boltzmann total ($\Leftrightarrow$ fonction de partition).

$$\mathcal{Z}'(i,j) = \sum \left\{ \begin{array}{cc} e^{\frac{-E_H(i,j)}{RT}} + e^{\frac{-E_S(i,j)}{RT}} \mathcal{Z}'(i+1,j-1) & \text{A} \\ \sum \left(e^{\frac{-E_{BI}(i,i',j',j)}{RT}} \mathcal{Z}'(i',j') \right) & \text{B} \\ e^{\frac{-(a+c)}{RT}} \sum \left(\mathcal{Z}'(i+1,k-1) \mathcal{Z}^1(k,j-1) \right) & \text{C} \end{array} \right\}$$

r

$$A_1 | A_2 | B_i | B_{i+1} | \ldots | B_{j-1} | B_j | C_i | C_{i+1} | \ldots | C_{j-1} | C_j$$

Après $\Theta(n)$ opérations, réitération sur intervalle de longueur $n-k$.
$\Rightarrow$ Complexité au pire pour k échantillons en $\mathcal{O}(n^2 k)$

Remarque : Peut être vu comme une instance pondérée d'un problème de génération aléatoire [DRT00].

Réécriture de l'algorithme `SFold` [DL03]:

1. Générer un nombre aléatoire dans $[0, \mathcal{Z}'(i,j))$
2. Retirer de r les contributions individuelles à $\mathcal{Z}'(i,j)$, jusqu'à $r < 0$
3. Réitérer récursivement sur les matrices des sous-intervalles

Idée : On choisit un cas ($\Leftrightarrow$ sous-ensemble de struct. sec.) proportionnellement à son facteur de Boltzmann total ($\Leftrightarrow$ fonction de partition).

$$\mathcal{Z}'(i,j) = \sum \left\{ \begin{array}{cc} e^{\frac{-E_H(i,j)}{RT}} + e^{\frac{-E_S(i,j)}{RT}} \mathcal{Z}'(i+1,j-1) & \text{A} \\ \sum \left(e^{\frac{-E_{BI}(i,i',j',j)}{RT}} \mathcal{Z}'(i',j') \right) & \text{B} \\ e^{\frac{-(a+c)}{RT}} \sum \left(\mathcal{Z}'(i+1,k-1) \mathcal{Z}^1(k,j-1) \right) & \text{C} \end{array} \right\}$$

r

$A_1 | A_2 | B_i | B_{i+1} | \ldots | B_{j-1} | B_j | C_i | C_{i+1} | \ldots | C_{j-1} | C_j$

Après $\Theta(n)$ opérations, réitération sur intervalle de longueur $n-k$.
$\Rightarrow$ Complexité au pire pour k échantillons en $\mathcal{O}(n^2 k)$

Remarque : Peut être vu comme une instance pondérée d'un problème de génération aléatoire [DRT00].

Réécriture de l'algorithme `SFold` [DL03]:

1. Générer un nombre aléatoire dans $[0, \mathcal{Z}'(i,j))$
2. Retirer de r les contributions individuelles à $\mathcal{Z}'(i,j)$, jusqu'à $r < 0$
3. Réitérer récursivement sur les matrices des sous-intervalles

Idée : On choisit un cas ($\Leftrightarrow$ sous-ensemble de struct. sec.) proportionnellement à son facteur de Boltzmann total ($\Leftrightarrow$ fonction de partition).

$$\mathcal{Z}'(i,j) = \sum \left\{ \begin{array}{cc} e^{\frac{-E_H(i,j)}{RT}} + e^{\frac{-E_S(i,j)}{RT}} \mathcal{Z}'(i+1,j-1) & \text{A} \\ \sum \left(e^{\frac{-E_{BI}(i,i',j',j)}{RT}} \mathcal{Z}'(i',j') \right) & \text{B} \\ e^{\frac{-(a+c)}{RT}} \sum \left(\mathcal{Z}'(i+1,k-1) \mathcal{Z}^1(k,j-1) \right) & \text{C} \end{array} \right\}$$

r

$A_1 | A_2 | B_i | B_{i+1} | \ldots | B_{j-1} | B_j | C_i | C_{i+1} | \ldots | C_{j-1} | C_j$

Après $\Theta(n)$ opérations, réitération sur intervalle de longueur $n-k$.
$\Rightarrow$ Complexité au pire pour k échantillons en $\mathcal{O}(n^2 k)$

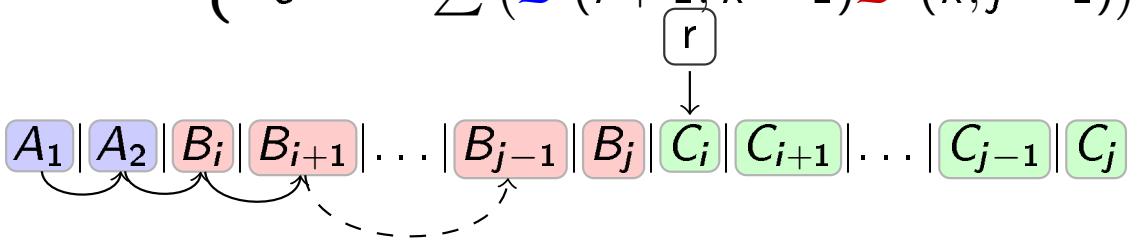
Remarque : Peut être vu comme une instance pondérée d'un problème de génération aléatoire [DRT00].

Réécriture de l'algorithme `SFold` [DL03]:

1. Générer un nombre aléatoire dans $[0, \mathcal{Z}'(i,j))$
2. Retirer de r les contributions individuelles à $\mathcal{Z}'(i,j)$, jusqu'à $r < 0$
3. Réitérer récursivement sur les matrices des sous-intervalles

Idée : On choisit un cas ($\Leftrightarrow$ sous-ensemble de struct. sec.) proportionnellement à son facteur de Boltzmann total ($\Leftrightarrow$ fonction de partition).

$$\mathcal{Z}'(i,j) = \sum \left\{ \begin{array}{cc} e^{\frac{-E_H(i,j)}{RT}} + e^{\frac{-E_S(i,j)}{RT}} \mathcal{Z}'(i+1,j-1) & \text{A} \\ \sum \left(e^{\frac{-E_{BI}(i,i',j',j)}{RT}} \mathcal{Z}'(i',j') \right) & \text{B} \\ e^{\frac{-(a+c)}{RT}} \sum \left(\mathcal{Z}'(i+1,k-1) \mathcal{Z}^1(k,j-1) \right) & \text{C} \end{array} \right\}$$



Après $\Theta(n)$ opérations, réitération sur intervalle de longueur $n-k$.
$\Rightarrow$ Complexité au pire pour k échantillons en $\mathcal{O}(n^2 k)$

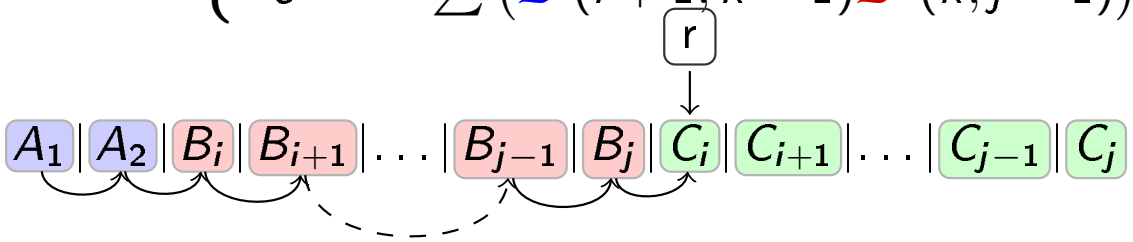
Remarque : Peut être vu comme une instance pondérée d'un problème de génération aléatoire [DRT00].

Réécriture de l'algorithme `SFold` [DL03]:

1. Générer un nombre aléatoire dans $[0, \mathcal{Z}'(i,j))$
2. Retirer de r les contributions individuelles à $\mathcal{Z}'(i,j)$, jusqu'à $r < 0$
3. Réitérer récursivement sur les matrices des sous-intervalles

Idée : On choisit un cas ($\Leftrightarrow$ sous-ensemble de struct. sec.) proportionnellement à son facteur de Boltzmann total ($\Leftrightarrow$ fonction de partition).

$$\mathcal{Z}'(i,j) = \sum \left\{ \begin{array}{cc} e^{\frac{-E_H(i,j)}{RT}} + e^{\frac{-E_S(i,j)}{RT}} \mathcal{Z}'(i+1,j-1) & \text{A} \\ \sum \left(e^{\frac{-E_{BI}(i,i',j',j)}{RT}} \mathcal{Z}'(i',j') \right) & \text{B} \\ e^{\frac{-(a+c)}{RT}} \sum \left(\mathcal{Z}'(i+1,k-1) \mathcal{Z}^1(k,j-1) \right) & \text{C} \end{array} \right\}$$



Après $\Theta(n)$ opérations, réitération sur intervalle de longueur $n-k$.
$\Rightarrow$ Complexité au pire pour k échantillons en $\mathcal{O}(n^2 k)$

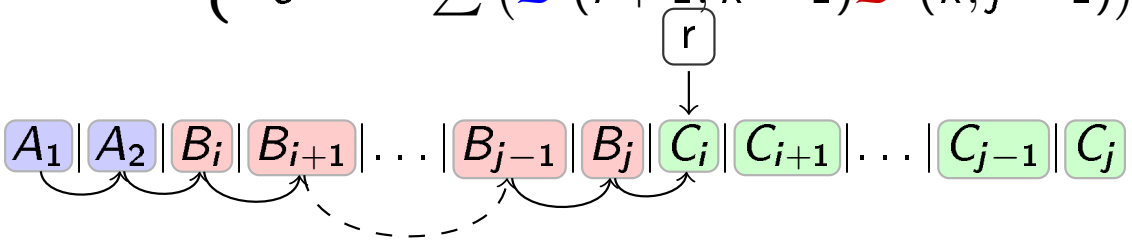
Remarque : Peut être vu comme une instance pondérée d'un problème de génération aléatoire [DRT00].

Réécriture de l'algorithme `SFold` [DL03]:

1. Générer un nombre aléatoire dans $[0, \mathcal{Z}'(i,j))$
2. Retirer de r les contributions individuelles à $\mathcal{Z}'(i,j)$, jusqu'à $r < 0$
3. Réitérer récursivement sur les matrices des sous-intervalles

Idée : On choisit un cas ($\Leftrightarrow$ sous-ensemble de struct. sec.) proportionnellement à son facteur de Boltzmann total ($\Leftrightarrow$ fonction de partition).

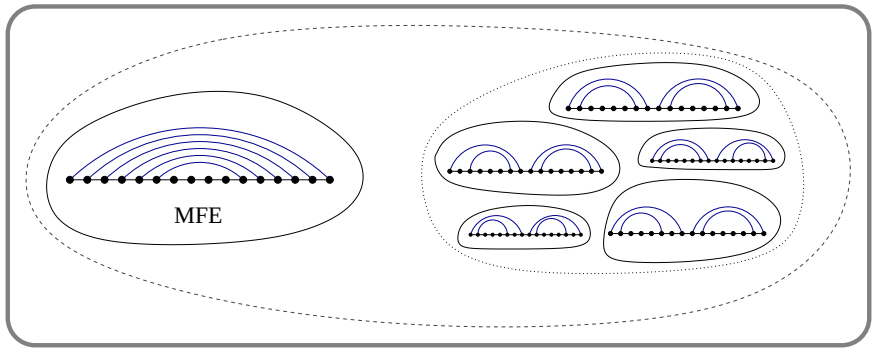
$$\mathcal{Z}'(i,j) = \sum \left\{ \begin{array}{cc} e^{\frac{-E_H(i,j)}{RT}} + e^{\frac{-E_S(i,j)}{RT}} \mathcal{Z}'(i+1,j-1) & \text{A} \\ \sum \left(e^{\frac{-E_{BI}(i,i',j',j)}{RT}} \mathcal{Z}'(i',j') \right) & \text{B} \\ e^{\frac{-(a+c)}{RT}} \sum \left(\mathcal{Z}'(i+1,k-1) \mathcal{Z}^1(k,j-1) \right) & \text{C} \end{array} \right\}$$



Après $\Theta(n)$ opérations, réitération sur intervalle de longueur $n-k$.
$\Rightarrow$ Complexité au pire pour k échantillons en $\mathcal{O}(n^2 k)$

Remarque : Peut être vu comme une instance pondérée d'un problème de génération aléatoire [DRT00].

Hypothèse : La MFE (Probabilité maximale) peut être *écrasée* par un ensemble de sous-optimaux structurellement similaires.
⇒ Conformation fonctionnelle plus probablement dans les sous-optimales.



Expérience : [DCL05]

- Échantillonner des structures selon leur probabilité de Boltzmann
- Effectuer un clustering basé sur la distance deux à deux
- Construire des structures consensus pour :
  - Moyenne de l'ensemble (*Ensemble centroid*)
  - Cluster le plus lourd (*Largest cluster centroid*)
  - Meilleur cluster (*Best cluster centroid*)

TABLE 3. Sensitivity of MFE structure and sensitivity improvement by centroids over MFE structure

RNA type	Number of sequences	Sensitivity[a] of MFE structure	Average percentage of improvement in sensitivity with respect to MFE structure[b]		
			Ensemble centroid	Largest cluster centroid	Best cluster centroid
SSU (16S) rRNA	10	49.80 ± 14.62	–4.14 ± 9.31	–2.75 ± 8.18	9.10 ± 16.51
LSU (23S) rRNA	10	35.35 ± 13.26	0.75 ± 15.41	0.42 ± 15.33	5.46 ± 12.43
5S rRNA	10	55.93 ± 24.52	2.41 ± 37.38	15.15 ± 61.20	41.81 ± 84.82
Group I intron	9	45.48 ± 19.97	6.46 ± 21.72	4.60 ± 24.19	29.06 ± 55.63
Group II intron	2	44.48 ± 6.74	0.54 ± 9.46	0.33 ± 8.08	–2.09 ± 4.66
RNase P	10	48.47 ± 18.52	–5.60 ± 13.95	–13.37 ± 33.04	4.48 ± 20.35
SRP RNA	10	76.20 ± 13.20	–1.93 ± 9.48	–0.76 ± 14.97	4.00 ± 5.73
tmRNA	10	36.16 ± 19.06	31.50 ± 81.06	24.06 ± 78.30	42.64 ± 85.71
tRNA	10	64.16 ± 17.55	–0.25 ± 3.39	9.76 ± 31.90	42.83 ± 38.90
Total	81	51.34 ± 21.29	3.54 ± 33.73	4.53 ± 39.81	21.74 ± 50.24

[a]Sensitivity = (number of base pairs in common between structure determined by comparative sequence analysis and predicted structure)/(number of base pairs in structure determined by comparative sequence analysis) × 100%.
[b]Sensitivity improvement by a centroid with respect to the MFE structure = [(sensitivity of centroid)/(sensitivity of MFE structure) – 1] × 100%.

TABLE 4. Positive predictive values (PPV) for MFE structure and PPV improvement by centroids over MFE structure

RNA type	Number of sequences	PPV[a] of MFE structure	Average percentage of improvement in PPV with respect to MFE structure[b]		
			Ensemble centroid	Largest cluster centroid	Best cluster centroid
SSU (16S) rRNA	10	48.10 ± 14.37	14.08 ± 10.67	9.54 ± 11.17	28.31 ± 22.97
LSU (23S) rRNA	10	33.68 ± 13.80	43.86 ± 52.10	36.84 ± 47.51	46.37 ± 43.28
5S rRNA	10	59.78 ± 25.76	26.52 ± 52.76	21.60 ± 57.60	51.42 ± 88.62
Group I intron	9	37.95 ± 20.81	36.37 ± 44.36	20.04 ± 32.99	56.50 ± 57.24
Group II intron	2	29.31 ± 26.14	33.33 ± 4.37	31.55 ± 1.26	32.84 ± 3.09
RNase P	10	42.89 ± 16.49	14.16 ± 18.84	–3.86 ± 37.00	27.99 ± 27.41
SRP RNA	10	77.59 ± 15.83	1.69 ± 13.23	–0.48 ± 15.64	9.63 ± 12.99
tmRNA	10	30.28 ± 19.83	76.78 ± 116.97	39.57 ± 91.33	95.55 ± 122.81
tRNA	10	55.15 ± 20.34	26.49 ± 36.50	15.29 ± 30.62	60.07 ± 44.22
Total	81	47.84 ± 23.28	30.00 ± 55.19	17.64 ± 46.87	46.51 ± 64.02

[a]Positive predictive value (PPV) = (number of common base pairs between structure determined by comparative sequence analysis and predicted structure)/(number of base pairs in predicted structure) × 100%.
[b]PPV improvement by a centroid with respect to the MFE structure = [(PPV of centroid)/(PPV of MFE structure) – 1] × 100%.

Conclusion [DCL05] : Prendre en compte les propriétés de l'ensemble de Boltzmann améliore la prédiction *ab initio*.

TABLE 3. Sensitivity of MFE structure and sensitivity improvement by centroids over MFE structure

RNA type	Number of sequences	Sensitivity[a] of MFE structure	Average percentage of improvement in sensitivity with respect to MFE structure[b]		
			Ensemble centroid	Largest cluster centroid	Best cluster centroid
SSU (16S) rRNA	10	49.80 ± 14.62	–4.14 ± 9.31	–2.75 ± 8.18	9.10 ± 16.51
LSU (23S) rRNA	10	35.35 ± 13.26	0.75 ± 15.41	0.42 ± 15.33	5.46 ± 12.43
5S rRNA	10	55.93 ± 24.52	2.41 ± 37.38	15.15 ± 61.20	41.81 ± 84.82
Group I intron	9	45.48 ± 19.97	6.46 ± 21.72	4.60 ± 24.19	29.06 ± 55.63
Group II intron	2	44.48 ± 6.74	0.54 ± 9.46	0.33 ± 8.08	–2.09 ± 4.66
RNase P	10	48.47 ± 18.52	–5.60 ± 13.95	–13.37 ± 33.04	4.48 ± 20.35
SRP RNA	10	76.20 ± 13.20	–1.93 ± 9.48	–0.76 ± 14.97	4.00 ± 5.73
tmRNA	10	36.16 ± 19.06	31.50 ± 81.06	24.06 ± 78.30	42.64 ± 85.71
tRNA	10	64.16 ± 17.55	–0.25 ± 3.39	9.76 ± 31.90	42.83 ± 38.90
Total	81	51.34 ± 21.29	3.54 ± 33.73	4.53 ± 39.81	21.74 ± 50.24

[a]Sensitivity = (number of base pairs in common between structure determined by comparative sequence analysis and predicted structure)/(number of base pairs in structure determined by comparative sequence analysis) × 100%.
[b]Sensitivity improvement by a centroid with respect to the MFE structure = [(sensitivity of centroid)/(sensitivity of MFE structure) – 1] × 100%.

TABLE 4. Positive predictive values (PPV) for MFE structure and PPV improvement by centroids over MFE structure

RNA type	Number of sequences	PPV[a] of MFE structure	Average percentage of improvement in PPV with respect to MFE structure[b]		
			Ensemble centroid	Largest cluster centroid	Best cluster centroid
SSU (16S) rRNA	10	48.10 ± 14.37	14.08 ± 10.67	9.54 ± 11.17	28.31 ± 22.97
LSU (23S) rRNA	10	33.68 ± 13.80	43.86 ± 52.10	36.84 ± 47.51	46.37 ± 43.28
5S rRNA	10	59.78 ± 25.76	26.52 ± 52.76	21.60 ± 57.60	51.42 ± 88.62
Group I intron	9	37.95 ± 20.81	36.37 ± 44.36	20.04 ± 32.99	56.50 ± 57.24
Group II intron	2	29.31 ± 26.14	33.33 ± 4.37	31.55 ± 1.26	32.84 ± 3.09
RNase P	10	42.89 ± 16.49	14.16 ± 18.84	–3.86 ± 37.00	27.99 ± 27.41
SRP RNA	10	77.59 ± 15.83	1.69 ± 13.23	–0.48 ± 15.64	9.63 ± 12.99
tmRNA	10	30.28 ± 19.83	76.78 ± 116.97	39.57 ± 91.33	95.55 ± 122.81
tRNA	10	55.15 ± 20.34	26.49 ± 36.50	15.29 ± 30.62	60.07 ± 44.22
Total	81	47.84 ± 23.28	30.00 ± 55.19	17.64 ± 46.87	46.51 ± 64.02

[a]Positive predictive value (PPV) = (number of common base pairs between structure determined by comparative sequence analysis and predicted structure)/(number of base pairs in predicted structure) × 100%.
[b]PPV improvement by a centroid with respect to the MFE structure = [(PPV of centroid)/(PPV of MFE structure) – 1] × 100%.

Conclusion [DCL05] : Prendre en compte les propriétés de l'ensemble de Boltzmann améliore la prédiction *ab initio*.

TABLE 3. Sensitivity of MFE structure and sensitivity improvement by centroids over MFE structure

RNA type	Number of sequences	Sensitivity[a] of MFE structure	Average percentage of improvement in sensitivity with respect to MFE structure[b]		
			Ensemble centroid	Largest cluster centroid	Best cluster centroid
SSU (16S) rRNA	10	49.80 ± 14.62	–4.14 ± 9.31	–2.75 ± 8.18	9.10 ± 16.51
LSU (23S) rRNA	10	35.35 ± 13.26	0.75 ± 15.41	0.42 ± 15.33	5.46 ± 12.43
5S rRNA	10	55.93 ± 24.52	2.41 ± 37.38	15.15 ± 61.20	41.81 ± 84.82
Group I intron	9	45.48 ± 19.97	6.46 ± 21.72	4.60 ± 24.19	29.06 ± 55.63
Group II intron	2	44.48 ± 6.74	0.54 ± 9.46	0.33 ± 8.08	–2.09 ± 4.66
RNase P	10	48.47 ± 18.52	–5.60 ± 13.95	–13.37 ± 33.04	4.48 ± 20.35
SRP RNA	10	76.20 ± 13.20	–1.93 ± 9.48	–0.76 ± 14.97	4.00 ± 5.73
tmRNA	10	36.16 ± 19.06	31.50 ± 81.06	24.06 ± 78.30	42.64 ± 85.71
tRNA	10	64.16 ± 17.55	–0.25 ± 3.39	9.76 ± 31.90	42.83 ± 38.90
Total	81	51.34 ± 21.29	3.54 ± 33.73	4.53 ± 39.81	21.74 ± 50.24

[a]Sensitivity = (number of base pairs in common between structure determined by comparative sequence analysis and predicted structure)/(number of base pairs in structure determined by comparative sequence analysis) × 100%.
[b]Sensitivity improvement by a centroid with respect to the MFE structure = [(sensitivity of centroid)/(sensitivity of MFE structure) – 1] × 100%.

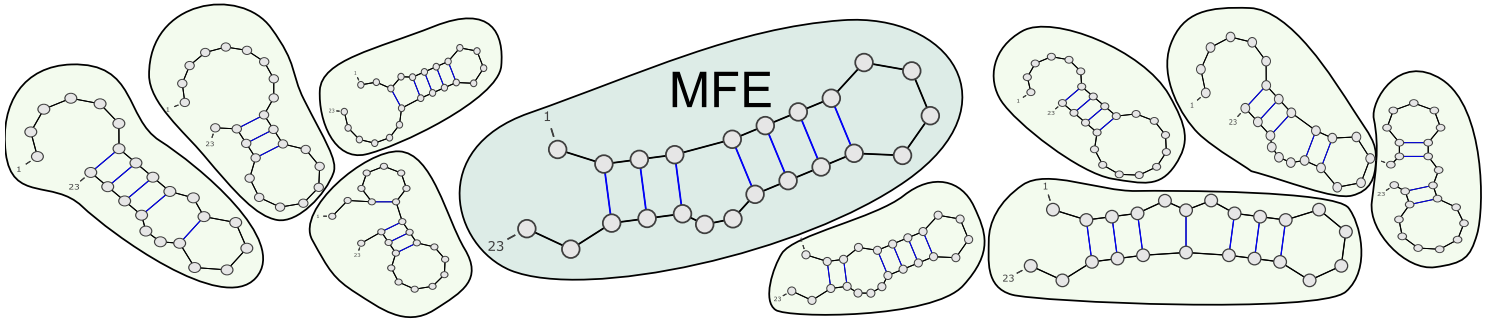
TABLE 4. Positive predictive values (PPV) for MFE structure and PPV improvement by centroids over MFE structure

RNA type	Number of sequences	PPV[a] of MFE structure	Average percentage of improvement in PPV with respect to MFE structure[b]		
			Ensemble centroid	Largest cluster centroid	Best cluster centroid
SSU (16S) rRNA	10	48.10 ± 14.37	14.08 ± 10.67	9.54 ± 11.17	28.31 ± 22.97
LSU (23S) rRNA	10	33.68 ± 13.80	43.86 ± 52.10	36.84 ± 47.51	46.37 ± 43.28
5S rRNA	10	59.78 ± 25.76	26.52 ± 52.76	21.60 ± 57.60	51.42 ± 88.62
Group I intron	9	37.95 ± 20.81	36.37 ± 44.36	20.04 ± 32.99	56.50 ± 57.24
Group II intron	2	29.31 ± 26.14	33.33 ± 4.37	31.55 ± 1.26	32.84 ± 3.09
RNase P	10	42.89 ± 16.49	14.16 ± 18.84	–3.86 ± 37.00	27.99 ± 27.41
SRP RNA	10	77.59 ± 15.83	1.69 ± 13.23	–0.48 ± 15.64	9.63 ± 12.99
tmRNA	10	30.28 ± 19.83	76.78 ± 116.97	39.57 ± 91.33	95.55 ± 122.81
tRNA	10	55.15 ± 20.34	26.49 ± 36.50	15.29 ± 30.62	60.07 ± 44.22
Total	81	47.84 ± 23.28	30.00 ± 55.19	17.64 ± 46.87	46.51 ± 64.02

[a]Positive predictive value (PPV) = (number of common base pairs between structure determined by comparative sequence analysis and predicted structure)/(number of base pairs in predicted structure) × 100%.
[b]PPV improvement by a centroid with respect to the MFE structure = [(PPV of centroid)/(PPV of MFE structure) – 1] × 100%.

Conclusion [DCL05] : Prendre en compte les propriétés de l'ensemble de Boltzmann améliore la prédiction *ab initio*.

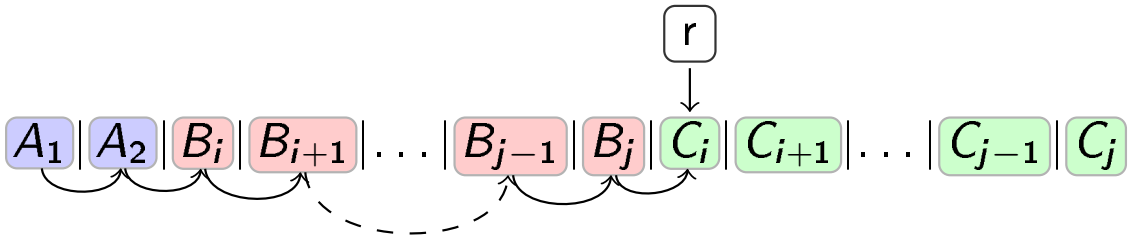
Conclusion [DCL05] : Prendre en compte les propriétés de l'ensemble de Boltzmann améliore la prédiction *ab initio*.



Malheureusement, l'échantillonnage des structures coûte cher ...
Comment l'optimiser ?

Comment améliorer l'échantillonnage ?

- Améliorer la complexité en temps:
  Complexité moyenne en temps en $\Theta(kn\sqrt{n})$ [Pon08]
  ($\Theta(n^2)$ lié à la réitération sur $n - \mathcal{O}(1)$ après $\Theta(n)$ opérations)
  - Intercaler les contributions des bulges (B) et des boucles multiples (C).
  - Boustrophedon [FZV94]
    Explorer les décompositions dissymétriques d'abord, puis les symétriques !



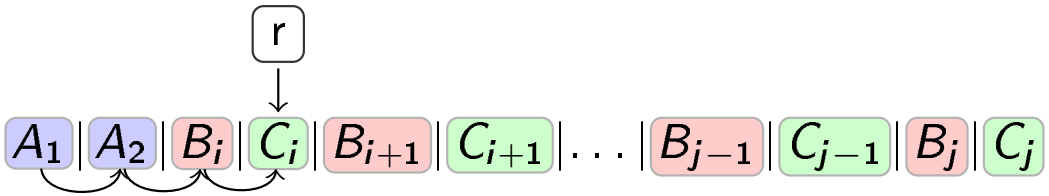
Message #3

Recherche Boustrophedon économise $\Theta(\frac{n}{\log n})/\Omega(\frac{\sqrt{n}}{\log n})$ au pire/en moyenne

Échantillonnage statistique efficace

Comment améliorer l'échantillonnage ?

- Améliorer la complexité en temps:
  Complexité moyenne en temps en $\Theta(kn\sqrt{n})$ [Pon08]
  ($\Theta(n^2)$ lié à la réitération sur $n - \mathcal{O}(1)$ après $\Theta(n)$ opérations)
  - Intercaler les contributions des bulges (B) et des boucles multiples (C).
  - Boustrophedon [FZV94]
    Explorer les décompositions dissymétriques d'abord, puis les symétriques !



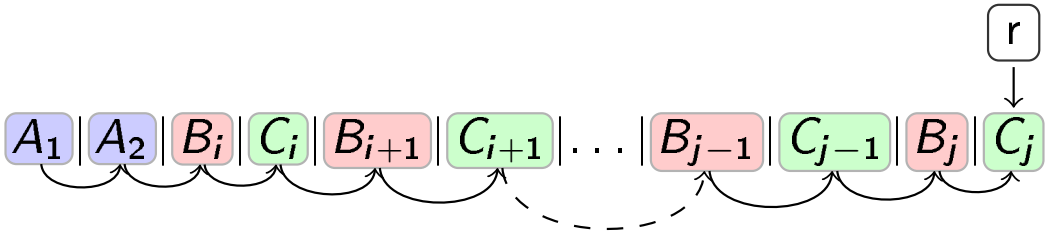
$\Rightarrow$ Quelques termes de B et C sont atteints en $\mathcal{O}(1)$ ops

Message #3

Recherche Boustrophedon économise $\Theta(\frac{n}{\log n})/\Omega(\frac{\sqrt{n}}{\log n})$ au pire/en moyenne

Comment améliorer l'échantillonnage ?

- Améliorer la complexité en temps:
  Complexité moyenne en temps en $\Theta(kn\sqrt{n})$ [Pon08]
  ($\Theta(n^2)$ lié à la réitération sur $n - \mathcal{O}(1)$ après $\Theta(n)$ opérations)
  - Intercaler les contributions des bulges (B) et des boucles multiples (C).
  - Boustrophedon [FZV94]
    Explorer les décompositions dissymétriques d'abord, puis les symétriques !



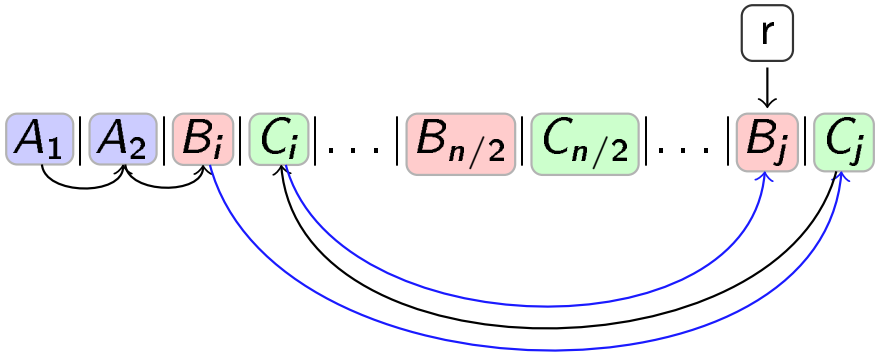
$\Rightarrow$ Quelques termes de B et C sont atteints en $\mathcal{O}(1)$ ops
Mais toujours $\Theta(n^2)$, car $\mathcal{Z}'(i,j) \to (\mathcal{Z}'(i+1,k-1), \mathcal{Z}^1(k,j-1))$

Message #3

Recherche Boustrophedon économise $\Theta(\frac{n}{\log n})/\Omega(\frac{\sqrt{n}}{\log n})$ au pire/en moyenne

Comment améliorer l'échantillonnage ?

- Améliorer la complexité en temps:
  Complexité moyenne en temps en $\Theta(kn\sqrt{n})$ [Pon08]
  ($\Theta(n^2)$ lié à la réitération sur $n - \mathcal{O}(1)$ après $\Theta(n)$ opérations)
  - Intercaler les contributions des bulges (B) et des boucles multiples (C).
  - Boustrophedon [FZV94]
    Explorer les décompositions dissymétriques d'abord, puis les symétriques !

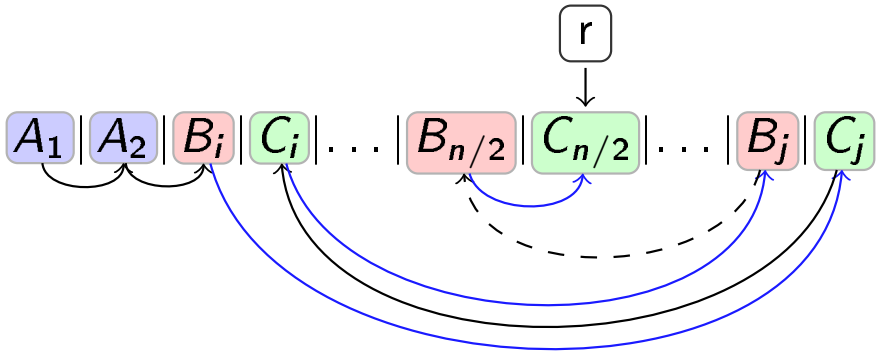


Message #3

Recherche Boustrophedon économise $\Theta(\frac{n}{\log n})/\Omega(\frac{\sqrt{n}}{\log n})$ au pire/en moyenne

Comment améliorer l'échantillonnage ?

- Améliorer la complexité en temps:
  Complexité moyenne en temps en $\Theta(kn\sqrt{n})$ [Pon08]
  ($\Theta(n^2)$ lié à la réitération sur $n - \mathcal{O}(1)$ après $\Theta(n)$ opérations)
  - Intercaler les contributions des bulges (B) et des boucles multiples (C).
  - Boustrophedon [FZV94] $\Rightarrow \Theta(n \log(n))$ au pire
    Explorer les décompositions dissymétriques d'abord, puis les symétriques !



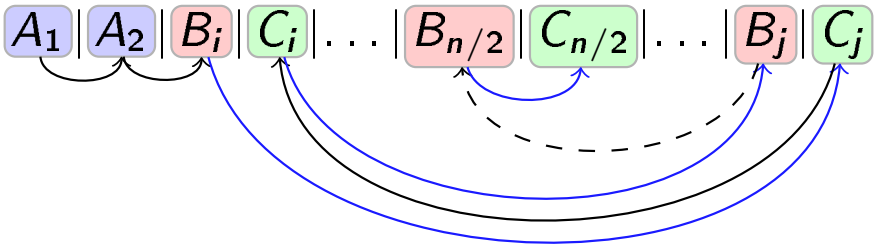
Cas au pire : Diviser exactement au milieu à chaque fois [GK81] $\Rightarrow \Theta(n \log(n))$

Message #3

Recherche Boustrophedon économise $\Theta(\frac{n}{\log n})/\Omega(\frac{\sqrt{n}}{\log n})$ au pire/en moyenne

Comment améliorer l'échantillonnage ?

- Améliorer la complexité en temps:
  Complexité moyenne en temps en $\Theta(kn\sqrt{n})$ [Pon08]
  ($\Theta(n^2)$ lié à la réitération sur $n - \mathcal{O}(1)$ après $\Theta(n)$ opérations)
  - Intercaler les contributions des bulges (B) et des boucles multiples (C).
  - Boustrophedon [FZV94]
    Explorer les décompositions dissymétriques d'abord, puis les symétriques !



Cas au pire : Diviser exactement au milieu à chaque fois [GK81] $\Rightarrow \Theta(n \log(n))$

Message #3

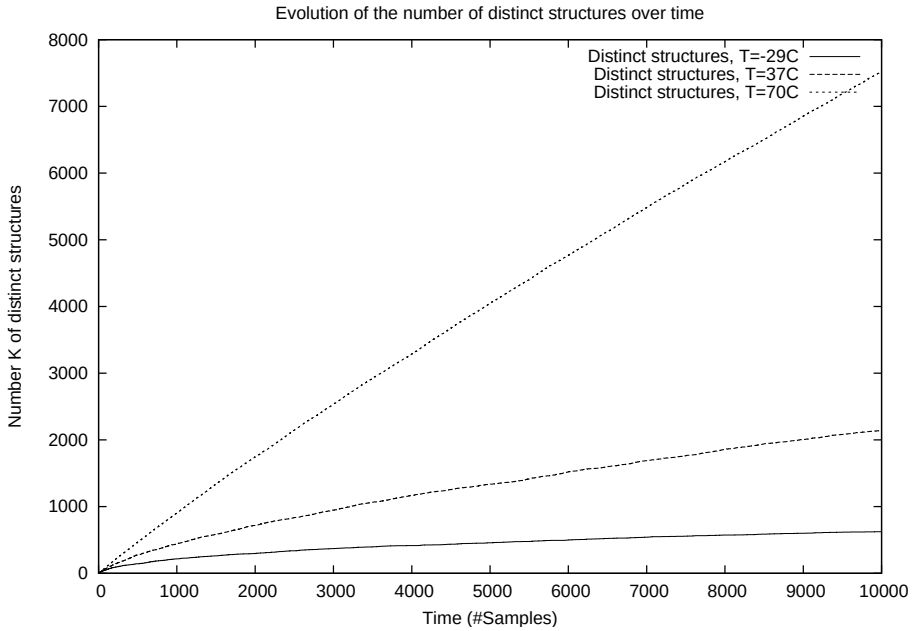
Recherche Boustrophedon économise $\Theta(\frac{n}{\log n})/\Omega(\frac{\sqrt{n}}{\log n})$ au pire/en moyenne

Comment améliorer l'échantillonnage ?

- Génération non-redondante (avec D. Gardy and A. Lorenz)

On calcule facilement la probabilité de Boltzmann d'une struct. engendrée.

$\Rightarrow$ Aucun intérêt à l'engendrer deux fois !!!



Problème : Combien de génération avant d'obtenir k échantillons **distincts** ?

Échantillonnage statistique efficace

Comment améliorer l'échantillonnage ?

- Génération non-redondante (avec D. Gardy and A. Lorenz)

Problème : Combien de génération avant d'obtenir k échantillons distincts ?

Collection complète (k = #structures): $E[C] \approx \mathcal{Z}'.n$
Bien plus que #struct. secondaires $\Rightarrow$ Nombre de collisions exponentiel.

Valeurs numériques (Homopolymère/Energie de Nussinov/$T = 37$):

$$E[C] \sim K \cdot 4.332^n/\sqrt{n} \quad \text{et} \quad \#\text{struct.} = S_n \sim K' \cdot 2.618^n/n\sqrt{n}$$

$\Rightarrow$ Chaque structure est engendrée $1.65^n \cdot n$ fois ($\neq \Theta(n)$ en uniforme)

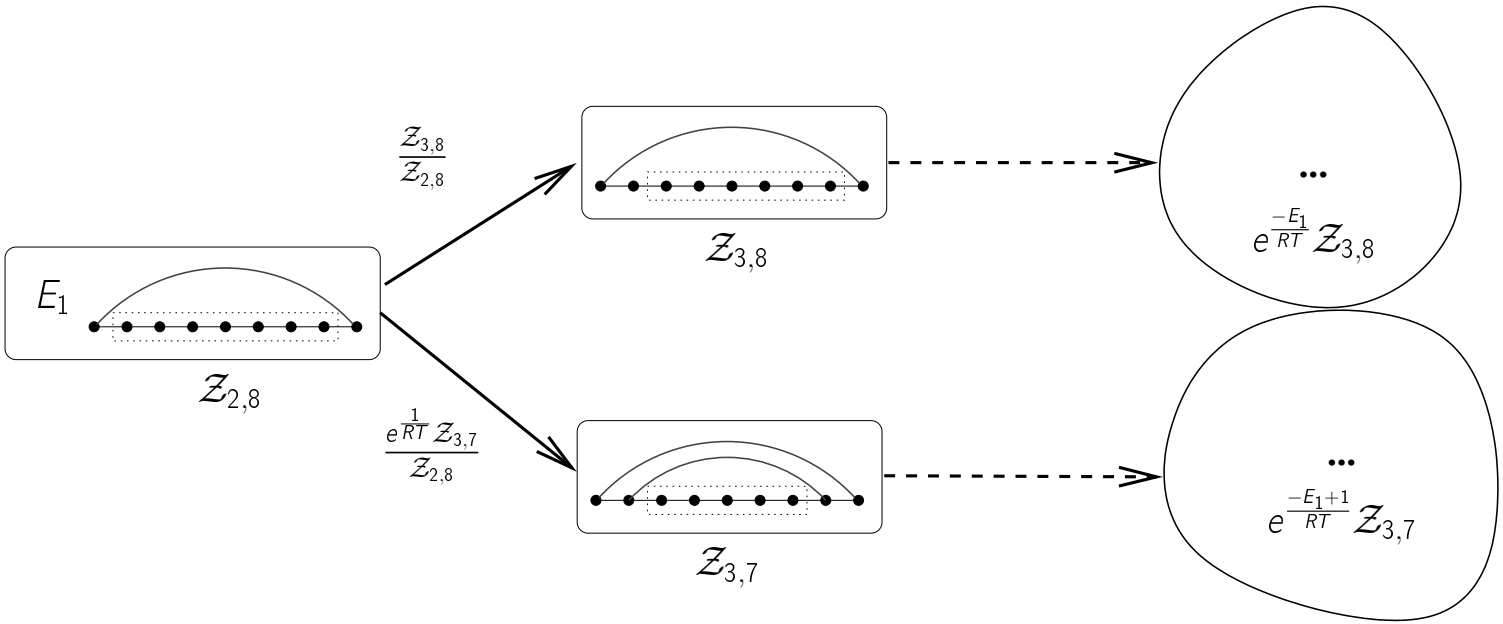
Message #4

Pour chaque ARN/T, il existe k tel que le temps d'échantillonnage de k structures distinctes est fortement dominé par le coût des collisions.

k dépend de la taille $\Rightarrow$ Nécessité de pousser notre analyse ...

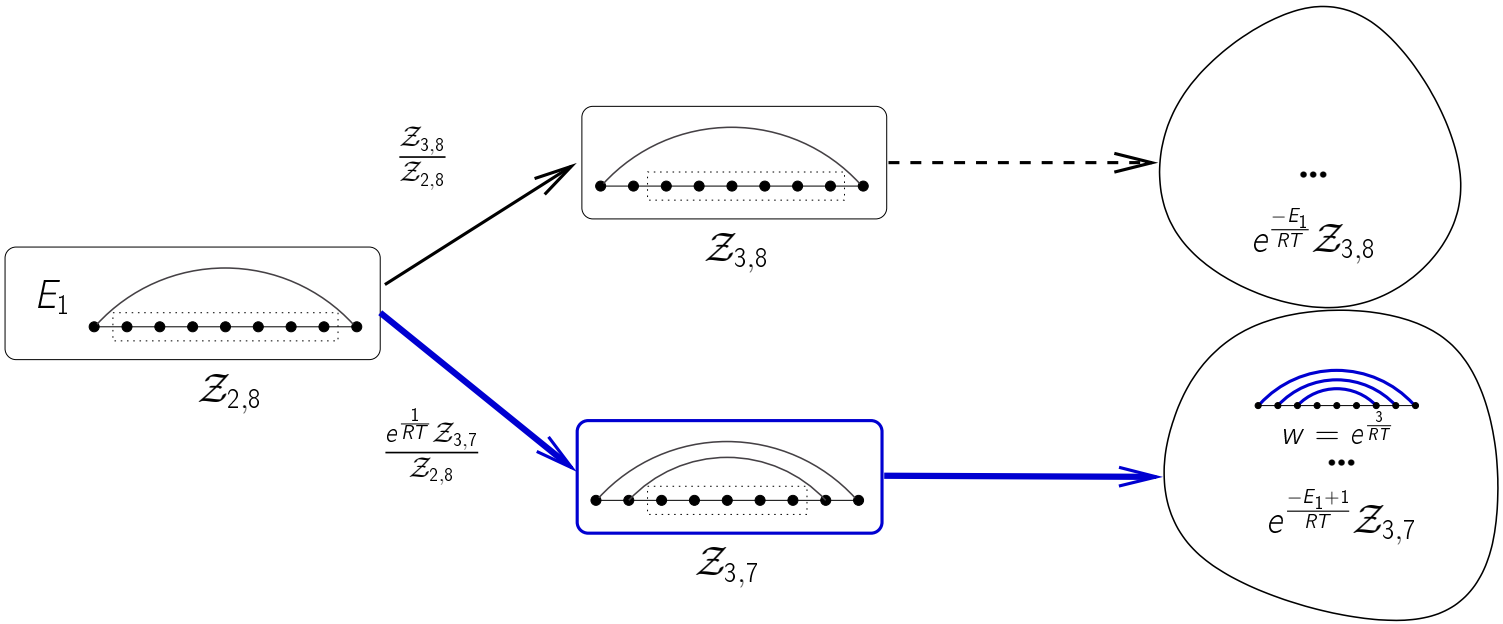
Comment améliorer l'échantillonnage ?

- Génération non-redondante (avec D. Gardy and A. Lorenz)
  - Construire un arbre préfixe des remontées contenant à chaque noeud les contributions $\sum_{S \in \mathcal{R}} e^{\frac{-E_S}{RT}}$ des structures déjà engendrées.
  - Pendant la remontée, modifier les termes en utilisant les termes précalculés [Pon08].



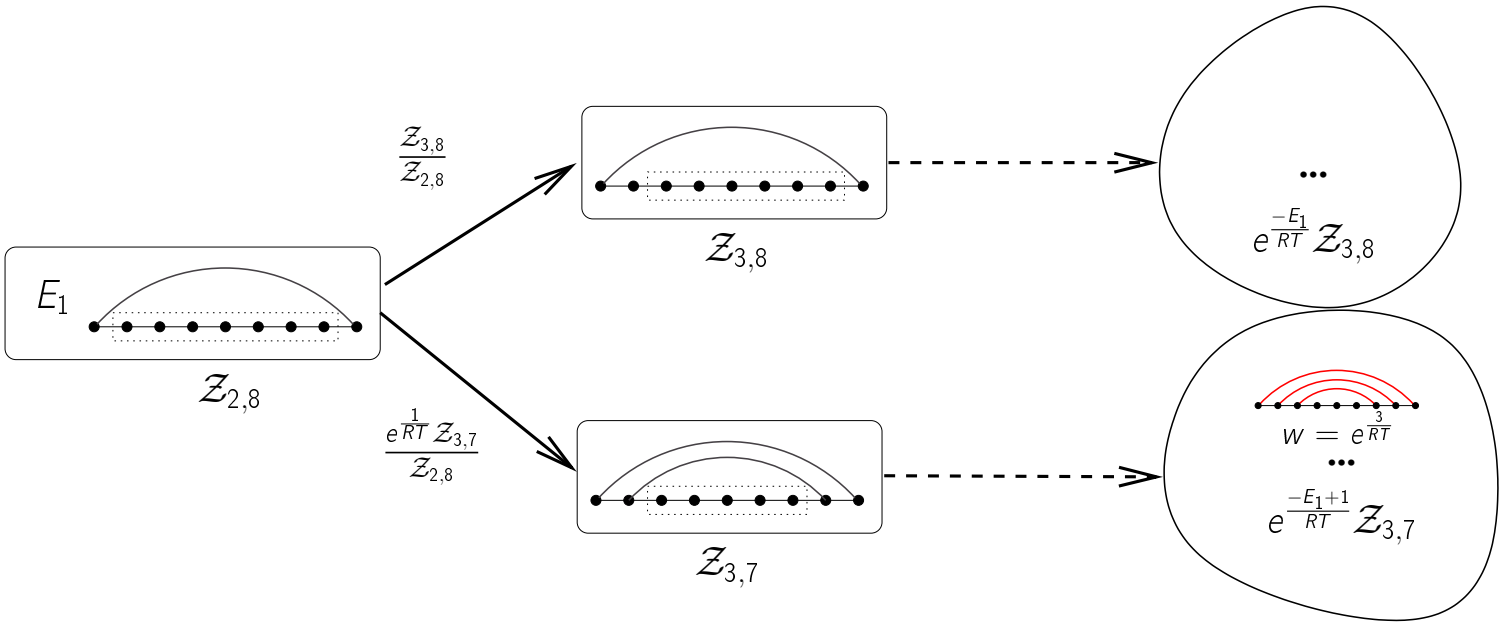
Comment améliorer l'échantillonnage ?

- Génération non-redondante (avec D. Gardy and A. Lorenz)
  - Construire un arbre préfixe des remontées contenant à chaque noeud les contributions $\sum_{S \in \mathcal{R}} e^{\frac{-E_S}{RT}}$ des structures déjà engendrées.
  - Pendant la remontée, modifier les termes en utilisant les termes précalculés [Pon08].



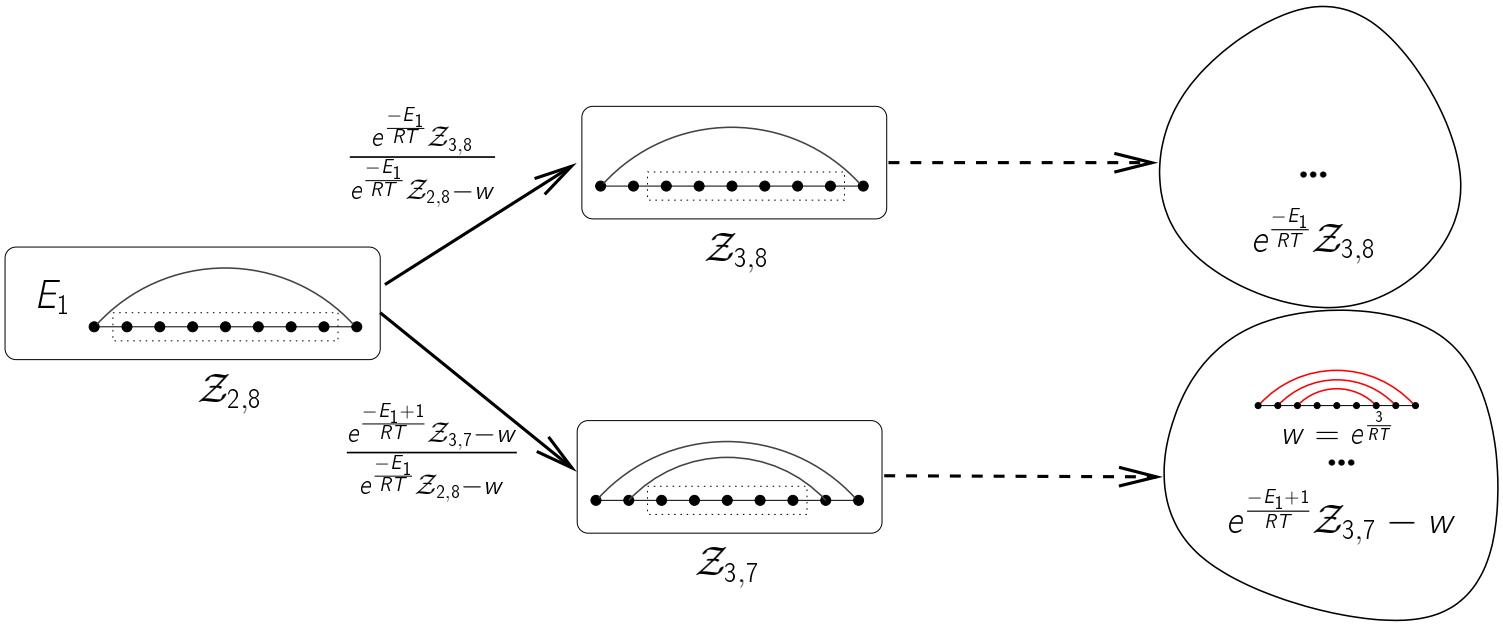
Comment améliorer l'échantillonnage ?

- Génération non-redondante (avec D. Gardy and A. Lorenz)
  - Construire un arbre préfixe des remontées contenant à chaque noeud les contributions $\sum_{S \in \mathcal{R}} e^{\frac{-E_S}{RT}}$ des structures déjà engendrées.
  - Pendant la remontée, modifier les termes en utilisant les termes précalculés [Pon08].



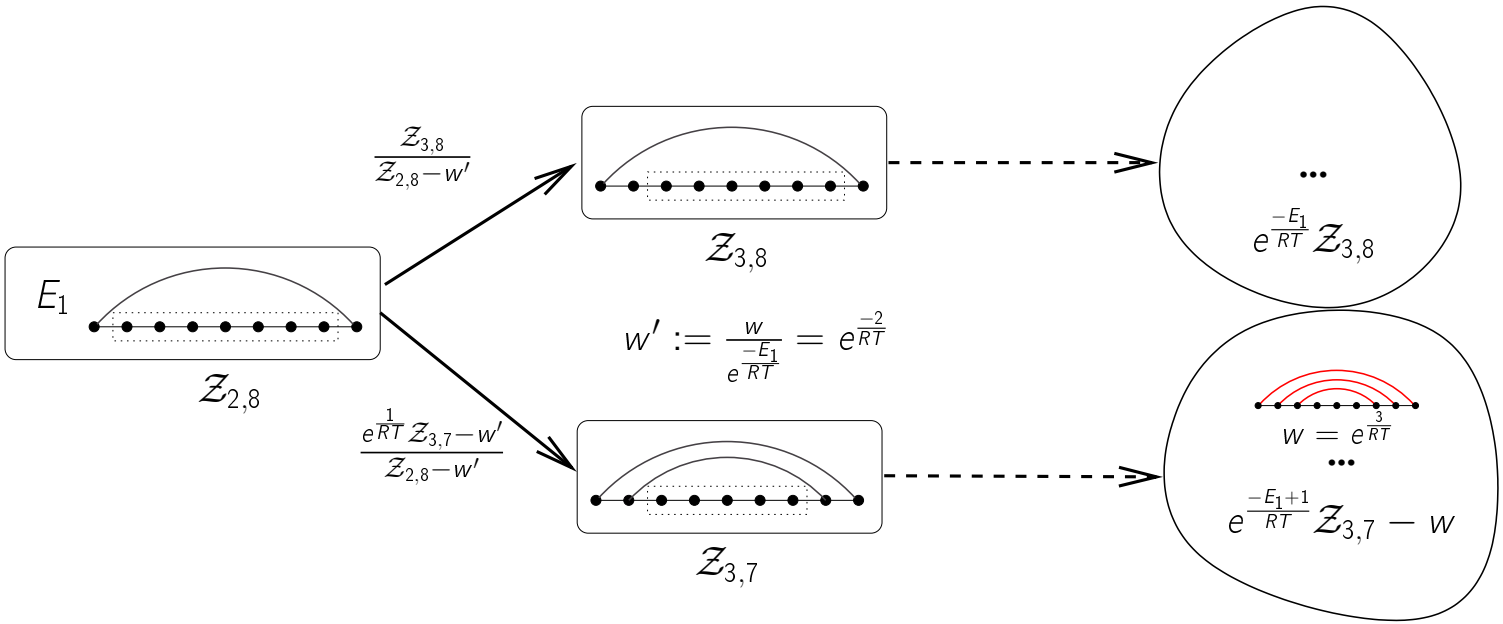
Comment améliorer l'échantillonnage ?

- Génération non-redondante (avec D. Gardy and A. Lorenz)
  - Construire un arbre préfixe des remontées contenant à chaque noeud les contributions $\sum_{S \in \mathcal{R}} e^{\frac{-E_S}{RT}}$ des structures déjà engendrées.
  - Pendant la remontée, modifier les termes en utilisant les termes précalculés [Pon08].



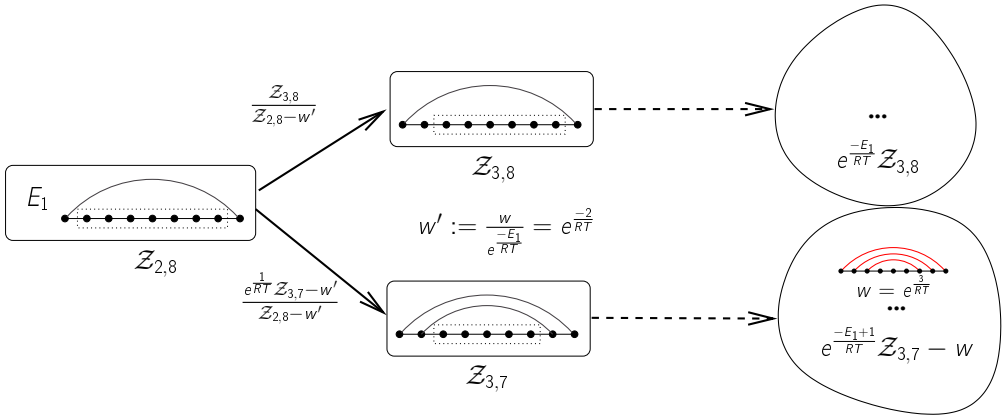
Comment améliorer l'échantillonnage ?

- Génération non-redondante (avec D. Gardy and A. Lorenz)
  - Construire un arbre préfixe des remontées contenant à chaque noeud les contributions $\sum_{S \in \mathcal{R}} e^{\frac{-E_S}{RT}}$ des structures déjà engendrées.
  - Pendant la remontée, modifier les termes en utilisant les termes précalculés [Pon08].



Comment améliorer l'échantillonnage ?

- Génération non-redondante (avec D. Gardy and A. Lorenz)
  - Construire un arbre préfixe des remontées contenant à chaque noeud les contributions $\sum_{S \in \mathcal{R}} e^{\frac{-E_S}{RT}}$ des structures déjà engendrées.
  - Pendant la remontée, modifier les termes en utilisant les termes précalculés [Pon08].



Message #5

En stockant les remontées et en biaisant les choix locaux, il est possible de réaliser un échantillonnage non-redondant en temps $\mathcal{O}(kn\log(n))$.

Conclusions

- Une vision combinatoire donne un cadre général pour valider/analyser/améliorer les algorithmes de programmation dynamique.
- Pendant la remontée stochastique, réordonner les comparaisons permet de gagner du temps !
- Échantillonnage ne gagne rien à être redondant
  $\Rightarrow$ Échantillonnage non redondant

Questions bioinformatiques :

- Revisiter les approches par centroides : Meilleures distances, reconstruction des centroides, détection ab initio du meilleur centroide.
- Origines des erreurs du repliement *ab initio* : Minimisation, pseudonoeuds, tertiaires ...

Questions algorithmiques :

- Quand est ce que le coût des collisions domine la complexité ?
- Existe t'il une alternative séquentielle à `RNASubopt`?

Merci aux organisatrices d'ARENA 2010 !!!

References I

Y. Ding, C. Y. Chan, and C. E. Lawrence.

RNA secondary structure prediction by centroids in a boltzmann weighted ensemble.
RNA, 11:1157–1166, 2005.

Y. Ding and E. Lawrence.

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

A. Denise, O. Roques, and M. Termier.

Random generation of words of context-free languagesaccording to the frequencies of letters.
In D. Gardy and A. Mokkadem, editors, *Mathematics and Computer Science: Algorithms, Trees, Combinatorics and probabilities*, Trends in Mathematics, pages 113–125. Birkhaüser, 2000.

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

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

D. H. Greene and D. E. Knuth.

Mathematics for the Analysis of Algorithms.
Birkhauser Boston, 1981.

J.S. McCaskill.

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

Ján Manuch, Chris Thachuk, Ladislav Stacho, and Anne Condon.

Np-completeness of the direct energy barrier problem without pseudoknots.
pages 106–115, 2009.

Y. 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.

References II

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.

M. S. Waterman.

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