

Cours M2 BIBS - Séance 2

Équilibre de Boltzmann et comparaison

Yann Ponty

Bioinformatics Team
École Polytechnique/CNRS/INRIA AMIB - France

14 Janvier 2012

- 1 Ensemble de Boltzmann
 - Ensemble de Boltzmann
 - Nussinov : Minimisation \Rightarrow Comptage
 - Calcul de la fonction de partition
 - Échantillonnage statistique

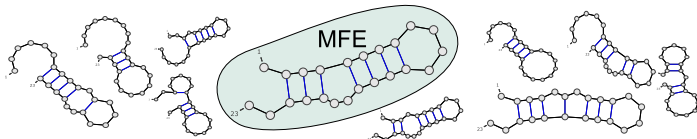
- 2 Extensions
 - Validité d'un schéma
 - Structures sous-optimales
 - Pseudo-noeuds

Ensemble canonique de Boltzmann

L'ARN *respire* \Rightarrow Il n'existe pas UNE unique conformation native.

Nouveau paradigme

Les conformations d'un ARN coexistent dans une distribution de Boltzmann.



Conséquence : La probabilité de la MFE peut être négligeable.
 \Rightarrow Comprendre les modes d'actions de l'ARN exige de prendre en considération l'ensemble des structures.

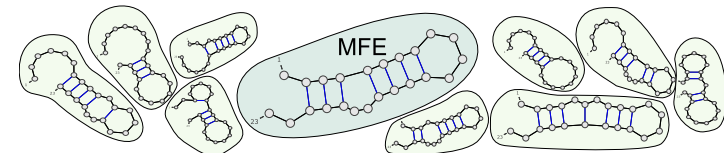
En particulier, des structures proches peuvent se *grouper* et devenir l'hypothèse la plus réaliste dans la recherche d'une conformation fonctionnelle.

Ensemble canonique de Boltzmann

L'ARN *respire* \Rightarrow Il n'existe pas UNE unique conformation native.

Nouveau paradigme

Les conformations d'un ARN coexistent dans une distribution de Boltzmann.



Conséquence : La probabilité de la MFE peut être négligeable.
 \Rightarrow Comprendre les modes d'actions de l'ARN exige de prendre en considération l'ensemble des structures.

En particulier, des structures proches peuvent se *grouper* et devenir l'hypothèse la plus réaliste dans la recherche d'une conformation fonctionnelle.

Une distribution de Boltzmann pondère chaque structure S pour un ARN ω par un **facteur de Boltzmann** $\mathcal{B}_{S,\omega} = e^{-\frac{E_{S,\omega}}{RT}}$ 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 ($1.986 \cdot 10^{-3} \text{ kCal.K}^{-1}.\text{mol}^{-1}$)

Distribution renormalisée sur S_ω par la **fonction de partition**

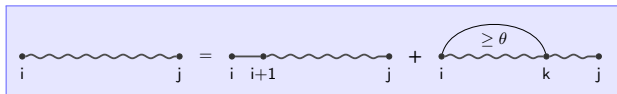
$$\mathcal{Z}_\omega = \sum_{S \in S_\omega} e^{-\frac{E_{S,\omega}}{RT}}$$

où S_ω est l'ensemble des conformations compatibles avec ω .

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

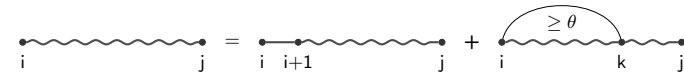
$$P_{S,\omega} = \frac{e^{-\frac{E_{S,\omega}}{RT}}}{\mathcal{Z}_\omega}$$

Fonction de partition = **Comptage pondéré** des structures compatibles



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

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



Récurrance sur l'**énergie minimale** d'un repliement :

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

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

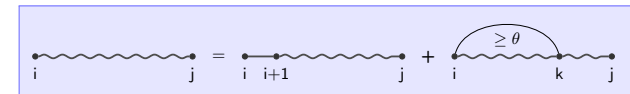
Récurrance de **comptage des structures compatibles** :

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

$$C_{i,j} = \sum \left\{ \begin{array}{l} C_{i+1,j} \\ \sum_{k=i+\theta+1}^j 1 \times C_{i+1,k-1} \times C_{k+1,j} \end{array} \right. \quad \begin{array}{l} (i \text{ non apparié}) \\ (i \text{ comp. avec } k) \end{array}$$

La décomposition est importante, le reste (MFE, comptage...) suit !

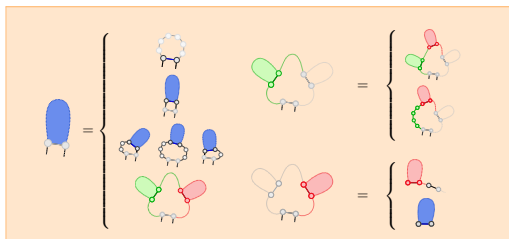
Fonction de partition = **Comptage pondéré** des structures compatibles



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

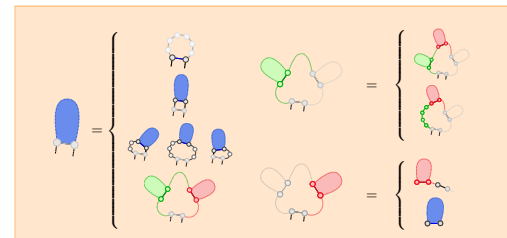
$$\mathcal{Z}_{i,j} = \sum \left\{ \begin{array}{l} \mathcal{Z}_{i+1,j} \\ \sum_{k=i+\theta+1}^j e^{-\frac{E_{bp}(i,k)}{RT}} \times \mathcal{Z}_{i+1,k-1} \times \mathcal{Z}_{k+1,j} \end{array} \right.$$

Fonction de partition = Comptage pondéré des structures compatibles



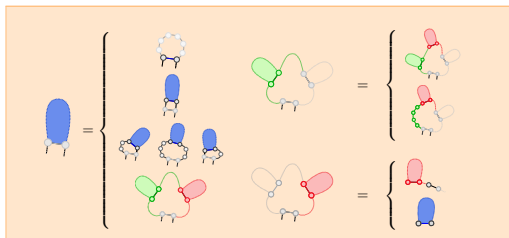
$$\begin{aligned}
 \mathcal{M}'_{i,j} &= \text{Min} \left\{ \begin{array}{l} E_H(i,j) \\ E_S(i,j) + \mathcal{M}'_{i+1,j-1} \\ \text{Min}(E_B(i,i',j',j) + \mathcal{M}'_{i',j'}) \\ a + c + \text{Min}(\mathcal{M}_{i+1,k-1} + \mathcal{M}^1_{k,j-1}) \end{array} \right\} \\
 \mathcal{M}_{i,j} &= \text{Min} \left\{ \text{Min}(\mathcal{M}_{i,k-1}, b(k-1)) + \mathcal{M}^1_{k,j} \right\} \\
 \mathcal{M}^1_{i,j} &= \text{Min} \left\{ b + \mathcal{M}^1_{i,j-1}, c + \mathcal{M}'_{i,j} \right\}
 \end{aligned}$$

Fonction de partition = Comptage pondéré des structures compatibles



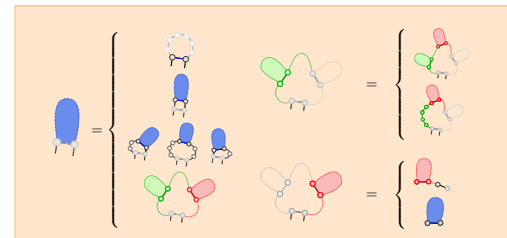
$$\begin{aligned}
 \mathcal{M}'_{i,j} &= \text{Min} \left\{ \begin{array}{l} e^{-\frac{E_H(i,j)}{RT}} \\ e^{-\frac{E_S(i,j)}{RT}} + \mathcal{M}'_{i+1,j-1} \\ \text{Min} \left(e^{-\frac{E_B(i,i',j',j)}{RT}} + \mathcal{M}'_{i',j'} \right) \\ e^{-\frac{(a+c)}{RT}} + \text{Min}(\mathcal{M}_{i+1,k-1} + \mathcal{M}^1_{k,j-1})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

Fonction de partition = Comptage pondéré des structures compatibles



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

Fonction de partition = Comptage pondéré des structures compatibles



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

Fonction de partition = Comptage pondéré des structures compatibles

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

$$Z_{i,j} = \sum \left\{ \begin{array}{l} Z_{i+1,j} \\ \sum_{k=i+\theta+1}^j e^{-\frac{E_{bp}(i,k)}{RT}} \times Z_{i+1,k-1} \times Z_{k+1,j} \end{array} \right.$$

Validité de la fonction de partition :

- Exhaustivité/non ambiguïté du schéma

Fonction de partition = Comptage pondéré des structures compatibles

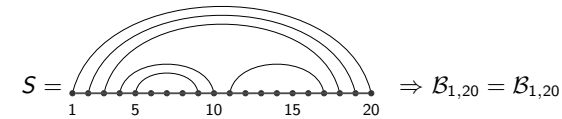
$$Z_{i,t} = 1, \quad \forall t \in [i, i + \theta]$$

$$Z_{i,j} = \sum \left\{ \begin{array}{l} Z_{i+1,j} \\ \sum_{k=i+\theta+1}^j e^{-\frac{E_{bp}(i,k)}{RT}} \times Z_{i+1,k-1} \times Z_{k+1,j} \end{array} \right.$$

Validité de la fonction de partition :

- Exhaustivité/non ambiguïté du schéma
 - Correction du facteur de Boltzmann
- Facteur d'un backtrack = Produit des facteurs de ses parties
Contributions énergétiques passent à l'exposant
- $$(e^{-a/RT} \cdot Z^1 \cdot Z' = e^{-a/RT} \cdot \sum_x e^{-E_x/RT} \cdot \sum_y e^{-E_y/RT}$$
- $$= \sum_{x,y} e^{-a/RT} \cdot e^{-E_x/RT} \cdot e^{-E_y/RT} = \sum_{x,y} e^{-(a+E_x+E_y)/RT})$$

Exemple :



Fonction de partition = Comptage pondéré des structures compatibles

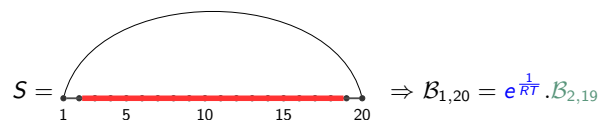
$$Z_{i,t} = 1, \quad \forall t \in [i, i + \theta]$$

$$Z_{i,j} = \sum \left\{ \begin{array}{l} Z_{i+1,j} \\ \sum_{k=i+\theta+1}^j e^{-\frac{E_{bp}(i,k)}{RT}} \times Z_{i+1,k-1} \times Z_{k+1,j} \end{array} \right.$$

Validité de la fonction de partition :

- Exhaustivité/non ambiguïté du schéma
 - Correction du facteur de Boltzmann
- Facteur d'un backtrack = Produit des facteurs de ses parties
Contributions énergétiques passent à l'exposant
- $$(e^{-a/RT} \cdot Z^1 \cdot Z' = e^{-a/RT} \cdot \sum_x e^{-E_x/RT} \cdot \sum_y e^{-E_y/RT}$$
- $$= \sum_{x,y} e^{-a/RT} \cdot e^{-E_x/RT} \cdot e^{-E_y/RT} = \sum_{x,y} e^{-(a+E_x+E_y)/RT})$$

Exemple :



Fonction de partition = Comptage pondéré des structures compatibles

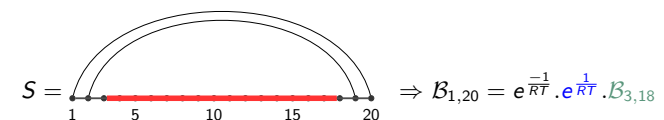
$$Z_{i,t} = 1, \quad \forall t \in [i, i + \theta]$$

$$Z_{i,j} = \sum \left\{ \begin{array}{l} Z_{i+1,j} \\ \sum_{k=i+\theta+1}^j e^{-\frac{E_{bp}(i,k)}{RT}} \times Z_{i+1,k-1} \times Z_{k+1,j} \end{array} \right.$$

Validité de la fonction de partition :

- Exhaustivité/non ambiguïté du schéma
 - Correction du facteur de Boltzmann
- Facteur d'un backtrack = Produit des facteurs de ses parties
Contributions énergétiques passent à l'exposant
- $$(e^{-a/RT} \cdot Z^1 \cdot Z' = e^{-a/RT} \cdot \sum_x e^{-E_x/RT} \cdot \sum_y e^{-E_y/RT}$$
- $$= \sum_{x,y} e^{-a/RT} \cdot e^{-E_x/RT} \cdot e^{-E_y/RT} = \sum_{x,y} e^{-(a+E_x+E_y)/RT})$$

Exemple :



Fonction de partition

Fonction de partition = Comptage pondéré des structures compatibles

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

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

Validité de la fonction de partition :

- Exhaustivité/non ambiguïté du schéma
- Correction du facteur de Boltzmann

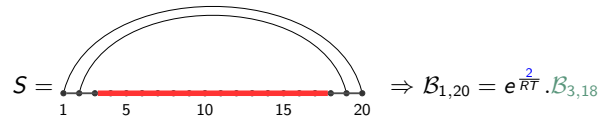
Facteur d'un backtrack = Produit des facteurs de ses parties

Contributions énergétiques passent à l'exposant

$$\left(e^{-a/RT} \cdot \mathcal{Z}^1 \cdot \mathcal{Z}^j = e^{-a/RT} \cdot \sum_x e^{-E_x/RT} \cdot \sum_y e^{-E_y/RT} \right)$$

$$= \sum_{x,y} e^{-a/RT} \cdot e^{-E_x/RT} \cdot e^{-E_y/RT} = \sum_{x,y} e^{-(a+E_x+E_y)/RT}$$

Exemple :



Fonction de partition

Fonction de partition = Comptage pondéré des structures compatibles

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

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

Validité de la fonction de partition :

- Exhaustivité/non ambiguïté du schéma
- Correction du facteur de Boltzmann

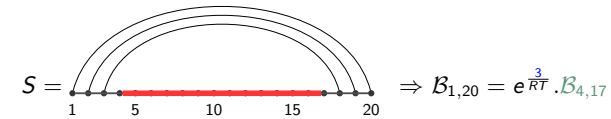
Facteur d'un backtrack = Produit des facteurs de ses parties

Contributions énergétiques passent à l'exposant

$$\left(e^{-a/RT} \cdot \mathcal{Z}^1 \cdot \mathcal{Z}^j = e^{-a/RT} \cdot \sum_x e^{-E_x/RT} \cdot \sum_y e^{-E_y/RT} \right)$$

$$= \sum_{x,y} e^{-a/RT} \cdot e^{-E_x/RT} \cdot e^{-E_y/RT} = \sum_{x,y} e^{-(a+E_x+E_y)/RT}$$

Exemple :



Fonction de partition

Fonction de partition = Comptage pondéré des structures compatibles

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

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

Validité de la fonction de partition :

- Exhaustivité/non ambiguïté du schéma
- Correction du facteur de Boltzmann

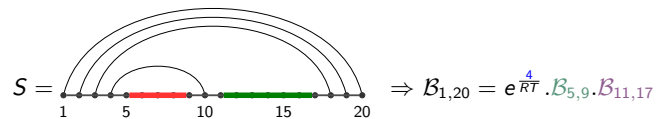
Facteur d'un backtrack = Produit des facteurs de ses parties

Contributions énergétiques passent à l'exposant

$$\left(e^{-a/RT} \cdot \mathcal{Z}^1 \cdot \mathcal{Z}^j = e^{-a/RT} \cdot \sum_x e^{-E_x/RT} \cdot \sum_y e^{-E_y/RT} \right)$$

$$= \sum_{x,y} e^{-a/RT} \cdot e^{-E_x/RT} \cdot e^{-E_y/RT} = \sum_{x,y} e^{-(a+E_x+E_y)/RT}$$

Exemple :



Fonction de partition

Fonction de partition = Comptage pondéré des structures compatibles

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

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

Validité de la fonction de partition :

- Exhaustivité/non ambiguïté du schéma
- Correction du facteur de Boltzmann

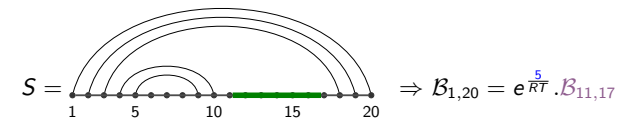
Facteur d'un backtrack = Produit des facteurs de ses parties

Contributions énergétiques passent à l'exposant

$$\left(e^{-a/RT} \cdot \mathcal{Z}^1 \cdot \mathcal{Z}^j = e^{-a/RT} \cdot \sum_x e^{-E_x/RT} \cdot \sum_y e^{-E_y/RT} \right)$$

$$= \sum_{x,y} e^{-a/RT} \cdot e^{-E_x/RT} \cdot e^{-E_y/RT} = \sum_{x,y} e^{-(a+E_x+E_y)/RT}$$

Exemple :



Fonction de partition

Fonction de partition = Comptage pondéré des structures compatibles

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

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

Validité de la fonction de partition :

- Exhaustivité/non ambiguïté du schéma
- Correction du facteur de Boltzmann

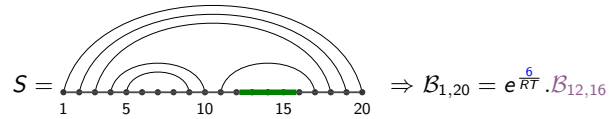
Facteur d'un backtrack = Produit des facteurs de ses parties

Contributions énergétiques passent à l'exposant

$$\left(e^{-a/RT} \cdot \mathcal{Z}^1 \cdot \mathcal{Z}' = e^{-a/RT} \cdot \sum_x e^{-E_x/RT} \cdot \sum_y e^{-E_y/RT} \right)$$

$$= \sum_{x,y} e^{-a/RT} \cdot e^{-E_x/RT} \cdot e^{-E_y/RT} = \sum_{x,y} e^{-(a+E_x+E_y)/RT}$$

Exemple :



Fonction de partition

Fonction de partition = Comptage pondéré des structures compatibles

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

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

Validité de la fonction de partition :

- Exhaustivité/non ambiguïté du schéma
- Correction du facteur de Boltzmann

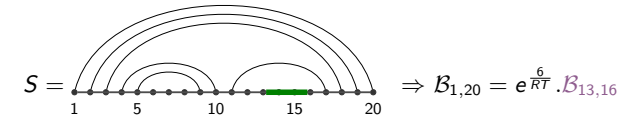
Facteur d'un backtrack = Produit des facteurs de ses parties

Contributions énergétiques passent à l'exposant

$$\left(e^{-a/RT} \cdot \mathcal{Z}^1 \cdot \mathcal{Z}' = e^{-a/RT} \cdot \sum_x e^{-E_x/RT} \cdot \sum_y e^{-E_y/RT} \right)$$

$$= \sum_{x,y} e^{-a/RT} \cdot e^{-E_x/RT} \cdot e^{-E_y/RT} = \sum_{x,y} e^{-(a+E_x+E_y)/RT}$$

Exemple :



Fonction de partition

Fonction de partition = Comptage pondéré des structures compatibles

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

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

Validité de la fonction de partition :

- Exhaustivité/non ambiguïté du schéma
- Correction du facteur de Boltzmann

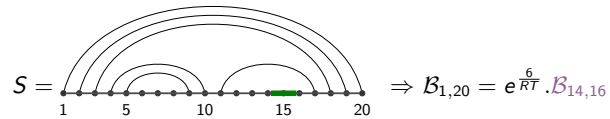
Facteur d'un backtrack = Produit des facteurs de ses parties

Contributions énergétiques passent à l'exposant

$$\left(e^{-a/RT} \cdot \mathcal{Z}^1 \cdot \mathcal{Z}' = e^{-a/RT} \cdot \sum_x e^{-E_x/RT} \cdot \sum_y e^{-E_y/RT} \right)$$

$$= \sum_{x,y} e^{-a/RT} \cdot e^{-E_x/RT} \cdot e^{-E_y/RT} = \sum_{x,y} e^{-(a+E_x+E_y)/RT}$$

Exemple :



Fonction de partition

Fonction de partition = Comptage pondéré des structures compatibles

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

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

Validité de la fonction de partition :

- Exhaustivité/non ambiguïté du schéma
- Correction du facteur de Boltzmann

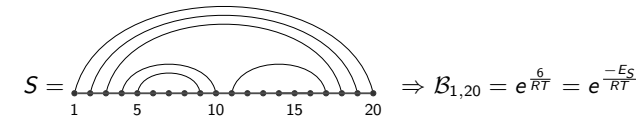
Facteur d'un backtrack = Produit des facteurs de ses parties

Contributions énergétiques passent à l'exposant

$$\left(e^{-a/RT} \cdot \mathcal{Z}^1 \cdot \mathcal{Z}' = e^{-a/RT} \cdot \sum_x e^{-E_x/RT} \cdot \sum_y e^{-E_y/RT} \right)$$

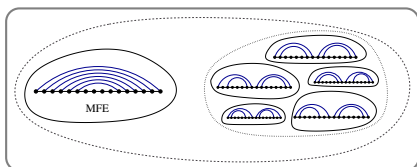
$$= \sum_{x,y} e^{-a/RT} \cdot e^{-E_x/RT} \cdot e^{-E_y/RT} = \sum_{x,y} e^{-(a+E_x+E_y)/RT}$$

Exemple :



Échantillonnage statistique de structures d'ARN

La MFE (Probabilité maximale) peut être largement dominée par un ensemble \mathcal{B} de sous-optimaux structurellement similaires.
 ⇒ Conformation fonctionnelle trouvée plus probablement dans \mathcal{B} .



Expérience : [DCL05]

- Échantillonner des structures selon une probabilité de Boltzmann
- Effectuer un clustering
- Construire structure consensus dans le plus lourd cluster

⇒ Amélioration relative pour spécificité (+17.6%) et sensibilité (+21.74%, sauf Introns du groupe II)

Problème

Comment engendrer des structures dans la distribution de Boltzmann ?

Remontée stochastique

Algorithme (Reformulation SFo1d [DL03])

Précalcul : Calculer les matrices (\mathcal{Z} , \mathcal{Z}' , \mathcal{Z}^1) des fonctions de partition.

Remontée stochastique :

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

Remontée stochastique

Algorithme (Reformulation SFo1d [DL03])

Précalcul : Calculer les matrices (\mathcal{Z} , \mathcal{Z}' , \mathcal{Z}^1) des fonctions de partition.

Remontée stochastique :

- Générer un nombre aléatoire r dans $[0, \mathcal{Z}'(i,j))$

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

Remontée stochastique

Algorithme (Reformulation SFo1d [DL03])

Précalcul : Calculer les matrices (\mathcal{Z} , \mathcal{Z}' , \mathcal{Z}^1) des fonctions de partition.

Remontée stochastique :

- Générer un nombre aléatoire r dans $[0, \mathcal{Z}'(i,j))$

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

$A_1 - A_2 - B_i - B_{i+1} - \dots - B_{j-1} - B_j - \underset{\substack{\downarrow \\ r}}{C_i} - C_{i+1} - \dots - C_{j-1} - C_j$

Algorithme (Reformulation SFo1d [DL03])

Précalcul : Calculer les matrices (\mathcal{Z} , \mathcal{Z}' , \mathcal{Z}^1) des fonctions de partition.

Remontée stochastique :

- 1 Générer un nombre aléatoire r dans $[0, \mathcal{Z}'(i, j))$
- 2 Retirer à r les contributions à $\mathcal{Z}'(i, j)$, jusqu'à ce que $r < 0$

$$\mathcal{Z}'(i, j) = \sum \left\{ \begin{array}{l} e^{-\frac{E_H(i, j)}{RT}} + e^{-\frac{E_S(i, j)}{RT}} \mathcal{Z}'(i+1, j-1) \quad \text{A} \\ \sum \left(e^{-\frac{E_{B_l}(i, i', j', j)}{RT}} \mathcal{Z}'(i', j') \right) \quad \text{B} \\ e^{-\frac{(g+c)}{RT}} \sum (\mathcal{Z}(i+1, k-1) \mathcal{Z}^1(k, j-1)) \quad \text{C} \end{array} \right.$$

Algorithme (Reformulation SFo1d [DL03])

Précalcul : Calculer les matrices (\mathcal{Z} , \mathcal{Z}' , \mathcal{Z}^1) des fonctions de partition.

Remontée stochastique :

- 1 Générer un nombre aléatoire r dans $[0, \mathcal{Z}'(i, j))$
- 2 Retirer à r les contributions à $\mathcal{Z}'(i, j)$, jusqu'à ce que $r < 0$

$$\mathcal{Z}'(i, j) = \sum \left\{ \begin{array}{l} e^{-\frac{E_H(i, j)}{RT}} + e^{-\frac{E_S(i, j)}{RT}} \mathcal{Z}'(i+1, j-1) \quad \text{A} \\ \sum \left(e^{-\frac{E_{B_l}(i, i', j', j)}{RT}} \mathcal{Z}'(i', j') \right) \quad \text{B} \\ e^{-\frac{(g+c)}{RT}} \sum (\mathcal{Z}(i+1, k-1) \mathcal{Z}^1(k, j-1)) \quad \text{C} \end{array} \right.$$

Algorithme (Reformulation SFo1d [DL03])

Précalcul : Calculer les matrices (\mathcal{Z} , \mathcal{Z}' , \mathcal{Z}^1) des fonctions de partition.

Remontée stochastique :

- 1 Générer un nombre aléatoire r dans $[0, \mathcal{Z}'(i, j))$
- 2 Retirer à r les contributions à $\mathcal{Z}'(i, j)$, jusqu'à ce que $r < 0$

$$\mathcal{Z}'(i, j) = \sum \left\{ \begin{array}{l} e^{-\frac{E_H(i, j)}{RT}} + e^{-\frac{E_S(i, j)}{RT}} \mathcal{Z}'(i+1, j-1) \quad \text{A} \\ \sum \left(e^{-\frac{E_{B_l}(i, i', j', j)}{RT}} \mathcal{Z}'(i', j') \right) \quad \text{B} \\ e^{-\frac{(g+c)}{RT}} \sum (\mathcal{Z}(i+1, k-1) \mathcal{Z}^1(k, j-1)) \quad \text{C} \end{array} \right.$$

Algorithme (Reformulation SFo1d [DL03])

Précalcul : Calculer les matrices (\mathcal{Z} , \mathcal{Z}' , \mathcal{Z}^1) des fonctions de partition.

Remontée stochastique :

- 1 Générer un nombre aléatoire r dans $[0, \mathcal{Z}'(i, j))$
- 2 Retirer à r les contributions à $\mathcal{Z}'(i, j)$, jusqu'à ce que $r < 0$

$$\mathcal{Z}'(i, j) = \sum \left\{ \begin{array}{l} e^{-\frac{E_H(i, j)}{RT}} + e^{-\frac{E_S(i, j)}{RT}} \mathcal{Z}'(i+1, j-1) \quad \text{A} \\ \sum \left(e^{-\frac{E_{B_l}(i, i', j', j)}{RT}} \mathcal{Z}'(i', j') \right) \quad \text{B} \\ e^{-\frac{(g+c)}{RT}} \sum (\mathcal{Z}(i+1, k-1) \mathcal{Z}^1(k, j-1)) \quad \text{C} \end{array} \right.$$

Algorithme (Reformulation SFo1d [DL03])

Précalcul : Calculer les matrices (\mathcal{Z} , \mathcal{Z}' , \mathcal{Z}^1) des fonctions de partition.

Remontée stochastique :

- 1 Générer un nombre aléatoire r dans $[0, \mathcal{Z}'(i, j))$
- 2 Retirer à r les contributions à $\mathcal{Z}'(i, j)$, jusqu'à ce que $r < 0$
- 3 Répéter sur les sous-structures

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

Correction : Chaque terme de la décomposition engendre $\mathcal{T} \in \{\mathcal{A}_1, \dots, \mathcal{C}_j\}$, et est choisi selon son facteur de Boltzmann cumulé $\mathcal{B}(\mathcal{T})/\mathcal{Z} = \sum_{S \in \mathcal{T}} e^{-E/RT}/\mathcal{Z}$ (Par récurrence).

Chaque structure $S \in \mathcal{S}_w$ est engendrée uniquement (Unambiguïté de Turner) par une séquence de choix d'ensembles $\mathcal{S}_w \supset E_1 \supset E_2 \supset \dots \supset \{S\}$.

La probabilité d'engendrer S est donc $p_S = \frac{\mathcal{B}(E_1)}{\mathcal{B}(\mathcal{S}_w)} \cdot \frac{\mathcal{B}(E_2)}{\mathcal{B}(E_1)} \cdot \frac{\mathcal{B}(E_3)}{\mathcal{B}(E_2)} \dots \frac{\mathcal{B}(\{S\})}{\mathcal{B}(E_m)}$

Algorithme (Reformulation SFo1d [DL03])

Précalcul : Calculer les matrices (\mathcal{Z} , \mathcal{Z}' , \mathcal{Z}^1) des fonctions de partition.

Remontée stochastique :

- 1 Générer un nombre aléatoire r dans $[0, \mathcal{Z}'(i, j))$
- 2 Retirer à r les contributions à $\mathcal{Z}'(i, j)$, jusqu'à ce que $r < 0$
- 3 Répéter sur les sous-structures

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

Correction : Chaque terme de la décomposition engendre $\mathcal{T} \in \{\mathcal{A}_1, \dots, \mathcal{C}_j\}$, et est choisi selon son facteur de Boltzmann cumulé $\mathcal{B}(\mathcal{T})/\mathcal{Z} = \sum_{S \in \mathcal{T}} e^{-E/RT}/\mathcal{Z}$ (Par récurrence).

Chaque structure $S \in \mathcal{S}_w$ est engendrée uniquement (Unambiguïté de Turner) par une séquence de choix d'ensembles $\mathcal{S}_w \supset E_1 \supset E_2 \supset \dots \supset \{S\}$.

La probabilité d'engendrer S est donc $p_S = \frac{1}{\mathcal{B}(\mathcal{S}_w)} \cdot \frac{1}{1} \cdot \frac{1}{1} \dots \frac{\mathcal{B}(\{S\})}{1}$

Algorithme (Reformulation SFo1d [DL03])

Précalcul : Calculer les matrices (\mathcal{Z} , \mathcal{Z}' , \mathcal{Z}^1) des fonctions de partition.

Remontée stochastique :

- 1 Générer un nombre aléatoire r dans $[0, \mathcal{Z}'(i, j))$
- 2 Retirer à r les contributions à $\mathcal{Z}'(i, j)$, jusqu'à ce que $r < 0$
- 3 Répéter sur les sous-structures

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

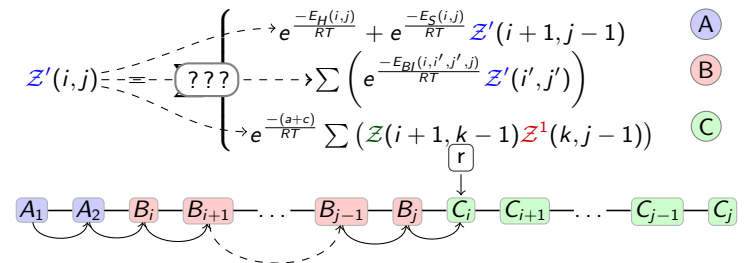
Correction : Chaque terme de la décomposition engendre $\mathcal{T} \in \{\mathcal{A}_1, \dots, \mathcal{C}_j\}$, et est choisi selon son facteur de Boltzmann cumulé $\mathcal{B}(\mathcal{T})/\mathcal{Z} = \sum_{S \in \mathcal{T}} e^{-E/RT}/\mathcal{Z}$ (Par récurrence).

Chaque structure $S \in \mathcal{S}_w$ est engendrée uniquement (Unambiguïté de Turner) par une séquence de choix d'ensembles $\mathcal{S}_w \supset E_1 \supset E_2 \supset \dots \supset \{S\}$.

La probabilité d'engendrer S est donc $p_S = \frac{\mathcal{B}(\{S\})}{\mathcal{B}(\mathcal{S}_w)} = \frac{e^{-E_S/RT}}{\mathcal{Z}} = P_{S, w}$

Algorithme (Reformulation SFo1d [DL03])

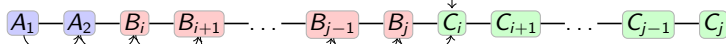
- 1 Générer un nombre aléatoire r dans $[0, \mathcal{Z}'(i, j))$
- 2 Retirer à r les contributions à $\mathcal{Z}'(i, j)$, jusqu'à ce que $r < 0$
- 3 Répéter sur les sous-structures



Algorithme (Reformulation SFo1d [DL03])

- 1 Générer un nombre aléatoire r dans $[0, Z'(i, j))$
- 2 Retirer à r les contributions à $Z'(i, j)$, jusqu'à ce que $r < 0$
- 3 Répéter sur les sous-structures

$$Z'(i, j) = \sum \left\{ \begin{array}{l} e^{-\frac{E_H(i, j)}{RT}} + e^{-\frac{E_S(i, j)}{RT}} Z'(i+1, j-1) \quad \text{A} \\ \sum \left(e^{-\frac{E_B(i, i', j', j)}{RT}} Z'(i', j') \right) \quad \text{B} \\ e^{-\frac{(g+c)}{RT}} \sum \left(Z(i+1, k-1) Z^1(k, j-1) \right) \quad \text{C} \end{array} \right.$$

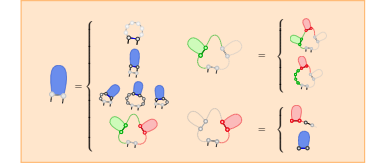
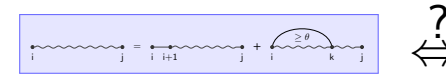


Après $\Theta(n)$ opérations, on réitère sur un interval de taille $n-1$
 \Rightarrow Complexité du cas au pire en $\mathcal{O}(n^2k)$ pour k échantillons

Remarque : Instance pondérée d'un problème de génération aléatoire par la méthode *réursive* [Pon08].
 Complexité en moyenne en $\Theta(n\sqrt{n})$ dans l'hypothèse *tout appariement*.
 Adaptation d'un parcours *Boustrophedon* $\Rightarrow \mathcal{O}(n \log nk)$ au pire.

Une preuve de correction possible :

- Calcul correct localement
- + Toutes les conformations sont parcourues
- \Rightarrow Algorithme correct (Induction)



Forte certitude *mais* pas encore preuve (Séries génératrices).

Une preuve de correction possible :

- Calcul correct localement
- + Toutes les conformations sont parcourues
- \Rightarrow Algorithme correct (Induction)

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

$$C_{i,j} = \sum \left\{ \begin{array}{l} C_{i+1,j} \\ \sum_{k=i+\theta+1}^j 1 \times C_{i+1,k-1} \times C_{k+1,j} \end{array} \right.$$

Homopolymère (Toute paire autorisée) + $\theta = 1$
 $\Rightarrow C_{1,n} = 1, 1, 1, 2, 4, 8, 17, 32, 82, 185, 423, \dots$



$$C'_{i,j} = \sum \left\{ \begin{array}{l} 1 \\ \sum_{i',j'} C'_{i',j'} \\ \sum_k C'_{i+1,k-1} \times C'_{k,j-1} \end{array} \right.$$

$$C_{i,j} = \sum_k ((C_{i,k-1} + 1) \times C'_{k,j})$$

$$C'_{i,j} = C'_{i,j-1} + C'_{i,j}$$

Homopolymère + $\theta = 1$
 $\Rightarrow C'_{1,n} = 0, 1, 1, 2, 4, 8, 17, 32, 82, 185, 423, \dots$

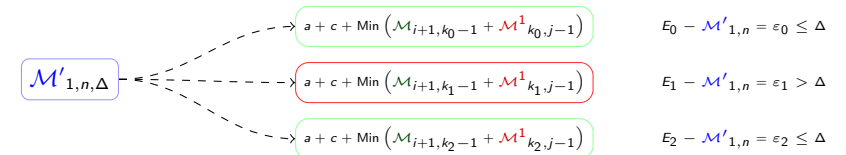
Forte certitude *mais* pas encore preuve (Séries génératrices).

Prob. : Simplifications de l'énergie (Pseudo-noeuds, non-can.)

\Rightarrow La structure *native* (fonctionnelle) pourrait être *ignorée*.

\Rightarrow Engendrer des repliements sous-optimaux (RNASubopt [WFHS99]),
 i.e. construire toutes les structures à Δ KCal.mol⁻¹ de la MFE :

- Calculer la matrice des énergies minimales
- Effectuer un Backtrack sur toutes les contributions à $\leq \Delta$ de la MFE



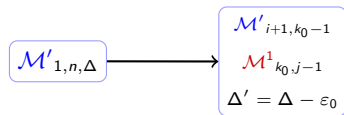
Repliement sous-optimal

Prob. : Simplifications de l'énergie (Pseudo-noeuds, non-can.)

⇒ La structure *native* (fonctionnelle) pourrait être *ignorée*.

⇒ Engendrer des repliements sous-optimaux (RNASubopt [WFHS99]),
i.e. construire toutes les structures à Δ KCal.mol⁻¹ de la MFE :

- Calculer la matrice des énergies minimales
- Effectuer un Backtrack sur toutes les contributions à $\leq \Delta$ de la MFE
- Mettre à jour Δ t.q. les futurs backtracks donnent ≥ 1 struct.



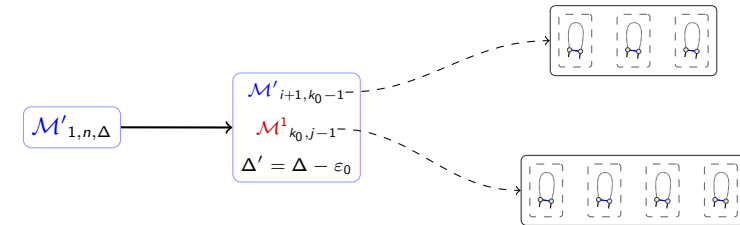
Repliement sous-optimal

Prob. : Simplifications de l'énergie (Pseudo-noeuds, non-can.)

⇒ La structure *native* (fonctionnelle) pourrait être *ignorée*.

⇒ Engendrer des repliements sous-optimaux (RNASubopt [WFHS99]),
i.e. construire toutes les structures à Δ KCal.mol⁻¹ de la MFE :

- Calculer la matrice des énergies minimales
- Effectuer un Backtrack sur toutes les contributions à $\leq \Delta$ de la MFE
- Mettre à jour Δ t.q. les futurs backtracks donnent ≥ 1 struct.
- Engendrer (Rec.) les sous-ensembles et combiner (brutal ou Tri)



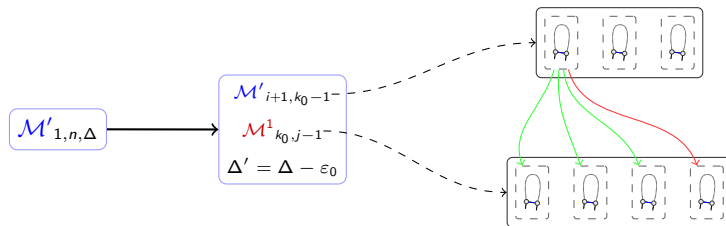
Repliement sous-optimal

Prob. : Simplifications de l'énergie (Pseudo-noeuds, non-can.)

⇒ La structure *native* (fonctionnelle) pourrait être *ignorée*.

⇒ Engendrer des repliements sous-optimaux (RNASubopt [WFHS99]),
i.e. construire toutes les structures à Δ KCal.mol⁻¹ de la MFE :

- Calculer la matrice des énergies minimales
- Effectuer un Backtrack sur toutes les contributions à $\leq \Delta$ de la MFE
- Mettre à jour Δ t.q. les futurs backtracks donnent ≥ 1 struct.
- Engendrer (Rec.) les sous-ensembles et combiner (brutal ou Tri)



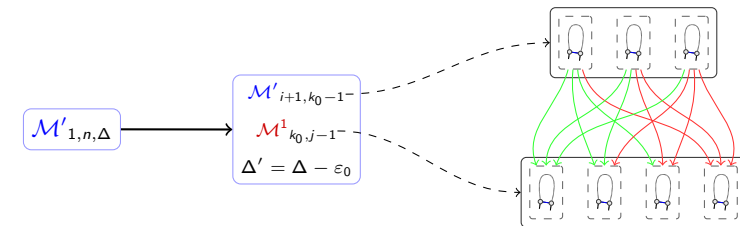
Repliement sous-optimal

Prob. : Simplifications de l'énergie (Pseudo-noeuds, non-can.)

⇒ La structure *native* (fonctionnelle) pourrait être *ignorée*.

⇒ Engendrer des repliements sous-optimaux (RNASubopt [WFHS99]),
i.e. construire toutes les structures à Δ KCal.mol⁻¹ de la MFE :

- Calculer la matrice des énergies minimales
- Effectuer un Backtrack sur toutes les contributions à $\leq \Delta$ de la MFE
- Mettre à jour Δ t.q. les futurs backtracks donnent ≥ 1 struct.
- Engendrer (Rec.) les sous-ensembles et combiner (brutal ou Tri)



Repliement sous-optimal

Prob. : Simplifications de l'énergie (Pseudo-noeuds, non-can.)

⇒ La structure *native* (fonctionnelle) pourrait être *ignorée*.

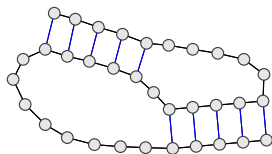
⇒ Engendrer des repliements sous-optimaux (RNASubopt [WFHS99]),
i.e. construire toutes les structures à Δ KCal.mol⁻¹ de la MFE :

- Calculer la matrice des énergies minimales
- Effectuer un Backtrack sur toutes les contributions à $\leq \Delta$ de la MFE
- Mettre à jour Δ t.q. les futurs backtracks donnent ≥ 1 struct.
- Engendrer (Rec.) les sous-ensembles et combiner (**brutal** ou **Tri**)

⇒ Complexité en temps (**Tri**) : $\mathcal{O}(n^3 + nk \log(k))$
(k croît exponentiellement sur Δ !)

Algorithme d'Akutsu/Uemura

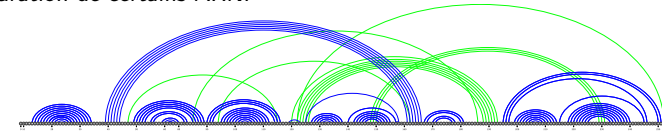
But : Capturer des catégorie de pseudo-noeuds *simples*, mais très représentés.



Idée : Quand on *retourne* ce type de pseudonoeuds, il suffit de précalculer les meilleures configurations *en dessous* d'un triplet (i, j, k) pour obtenir son énergie minimale.

Repliement avec pseudo-noeuds

Les pseudo-noeuds (et vrais noeuds) sont des constituants essentiels à la structuration de certains ARN.



Ribozyme du groupe I

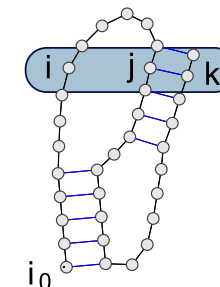
Leur absence historique au sein algorithmes de repliement est liée à la difficulté algorithmique des problèmes associés.

(Présence de croisements interdit une hypothèse d'indépendance des repliements).

Type	Complexité	Référence
Structures secondaires	$\mathcal{O}(n^3)$	[MSZT99]
L&P	$\mathcal{O}(n^5)$	[LP00]
D&P	$\mathcal{O}(n^5)$	[DP03]
A&U	$\mathcal{O}(n^5)$	[Aku00]
R&E	$\mathcal{O}(n^6)$	[RE99]
Généraux	NP-complet	[LP00]

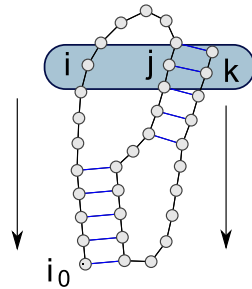
Algorithme d'Akutsu/Uemura

But : Capturer des catégorie de pseudo-noeuds *simples*, mais très représentés.

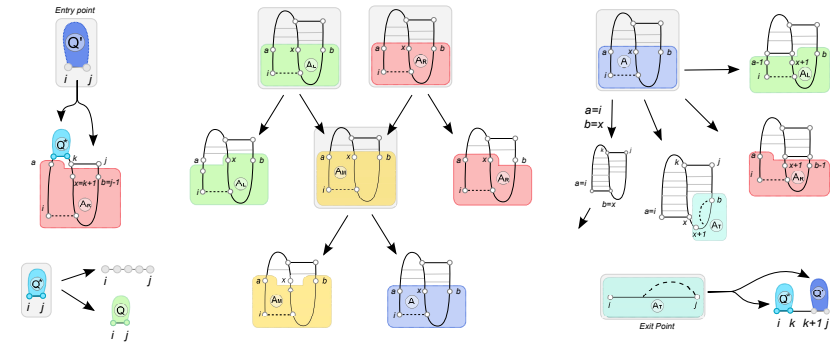


Idée : Quand on *retourne* ce type de pseudonoeuds, il suffit de précalculer les meilleures configurations *en dessous* d'un triplet (i, j, k) pour obtenir son énergie minimale.

But : Capturer des catégorie de pseudo-noeuds *simples*, mais très représentés.



Idée : Quand on *retourne* ce type de pseudonoeuds, il suffit de précalculer les meilleures configurations *en dessous* d'un triplet (i, j, k) pour obtenir son énergie minimale.



Application/Problème	Weight fun.	Time/Space	Ref.
Minimisation d'énergie	$\frac{-E}{RT}$	$\mathcal{O}(n^4)/\mathcal{O}(n^4)$	[Aku00]
Fonction de partition	$e^{\frac{-E}{RT}}$	$\mathcal{O}(n^4)/\mathcal{O}(n^4)$	$\Theta(n^6)$ [CC09]
Probabilité de paires de bases	$e^{\frac{-E}{RT}}$	$\mathcal{O}(n^4)/\mathcal{O}(n^4)$	-
Échantillonnage (k -struct.)	$e^{\frac{-E}{RT}}$	$\mathcal{O}(n^4 + kn \log n)/\mathcal{O}(n^4)$	-

Exercice : Ecrire l'équation de programmation dynamique associée pour le repliement, le comptage et la fonction de partition.

References I

- Tatsuya Akutsu.
Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots.
Discrete Appl. Math., 104(1-3) :45-62, 2000.
- S. Cao and S-J Chen.
Predicting structured and stabilities for h-type pseudoknots with interhelix loop.
RNA, 15 :696-706, 2009.
- 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.
- Robert M Dirks and Niles A Pierce.
A partition function algorithm for nucleic acid secondary structure including pseudoknots.
J Comput Chem, 24(13) :1664-1677, Oct 2003.
- R. B. Lyngsø and C. N. S. Pedersen.
RNA pseudoknot prediction in energy-based models.
Journal of Computational Biology, 7(3-4) :409-427, 2000.
- D.H. Mathews, J. Sabina, M. Zuker, and D.H. Turner.
Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure.
J Mol Biol, 288 :911-940, 1999.
- 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

- E. Rivas and S.R. Eddy.
A dynamic programming algorithm for RNA structure prediction including pseudoknots.
J Mol Biol, 285 :2053-2068, 1999.
- S. Wuchty, W. Fontana, I.L. Hofacker, and P. Schuster.
Complete suboptimal folding of RNA and the stability of secondary structures.
Biopolymers, 49 :145-164, 1999.