Cours M2 BIM - Séance 2

Équilibre de Boltzmann et comparaison

Yann Ponty

Bioinformatics Team
École Polytechnique/CNRS/INRIA AMIB – France

21 Janvier 2011

Plan

1. Performances et approches comparatives
   - Vers une prédiction ab-initio 3D
2. Ensemble de Boltzmann
   - Ensemble de Boltzmann
   - Nussinov : Minimisation $\Rightarrow$ Comptage
   - Calcul de la fonction de partition
   - Échantillonnage statistique
3. Extensions
   - Validité d'un schéma
   - Structures sous-optimales
   - Pseudo-noeuds
4. Alignement et comparaison de structures d'ARN
   - Méthode géométrique
   - Alignement de structures secondaires
   - Méthodes hybrides

Deux approches

Definition (Repliement ab initio)

Partant de la séquence, trouver la conformation minimisant une fonction d'énergie.

Avantages :
- Explication mécanique
- Complexité raisonnable $\mathcal{O}(n^3)/\mathcal{O}(n^2)$ temps/mémoire
- Exploration exhaustive

Limites :
- Pas de cinétique
- Pas d'info évolutive
- Performances limitées

Definition (Approche comparative)

Partant de plusieurs séquences homologues ou d'un alignement, trouver une conformation de score (énergie+alignement) élevé.

Avantages :
- Meilleures performances
- Affinement permanent

Limites :
- Complexité élevée
- Exploration non-exhaustive

Performances

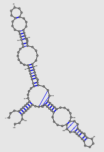
Rappel : $MCC = \frac{t^+t^- - f^+f^-}{\sqrt{(t^+ + f^+)(t^+ + f^-)(t^- + f^+)(t^- + f^-)}}$

Performances

UUAGGCGGCCACAGC
GGUGGGGUUGCCUCC
CGUACCCAUCCCGAA
CACGGAAGAUAAGCC
CACCAGCGUUCCGGG
GAGUACUGGAGUGCG
CGAGCCUCUGGGAAA
CCCGGUUCGCCGCCA
CC

Séquence

Taille	Sens.
<700	70-73%
[MSZT99, DCCG04]	
16s,23s	~50%
MCC~ 0.5 [GG04]	
Thermodynamique	



Structure 2^{aire}

Rappel : $MCC = \frac{t^+ t^- - f^+ f^-}{\sqrt{(t^+ + f^+)(t^+ + f^-)(t^- + f^+)(t^- + f^-)}}$

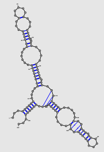
Performances

UUAGGCGGCCACAGC
GGUGGGGUUGCCUCC
CGUACCCAUCCCGAA
CACGGAAGAUAAGCC
CACCAGCGUUCCGGG
GAGUACUGGAGUGCG
CGAGCCUCUGGGAAA
CCCGGUUCGCCGCCA
CC

Séquence

Taille	Sens.
<700	70-73%
[MSZT99, DCCG04]	
16s,23s	~50%
MCC~ 0.5 [GG04]	
Thermodynamique	

Sens.	Spé.	MCC.
~75%	~75%	0.8
Comparative [GG04]		



Structure 2^{aire}

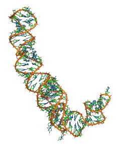
Rappel : $MCC = \frac{t^+ t^- - f^+ f^-}{\sqrt{(t^+ + f^+)(t^+ + f^-)(t^- + f^+)(t^- + f^-)}}$

Futur (proche) : Vers une prédiction 3D

But : De la séquence à des modèles tri-dimensionnels ! ! !

UUAGGCGGCCACAGC
GGUGGGGUUGCCUCC
CGUACCCAUCCCGAA
CACGGAAGAUAAGCC
CACCAGCGUUCCGGG
GAGUACUGGAGUGCG
CGAGCCUCUGGGAAA
CCCGGUUCGCCGCCA
CC

Séquence



Structure 3^{aire}

Futur (proche) : Vers une prédiction 3D

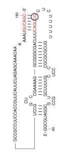
But : De la séquence à des modèles tri-dimensionnels ! ! !

- Models comparatifs + Dynamique moléculaires : `RNA2D3D` [SYKB07]

UUAGGCGGCCACAGC
GGUGGGGUUGCCUCC
CGUACCCAUCCCGAA
CACGGAAGAUAAGCC
CACCAGCGUUCCGGG
GAGUACUGGAGUGCG
CGAGCCUCUGGGAAA
CCCGGUUCGCCGCCA
CC

Séquence

Comparative
`KNetFold`



Structure 2^{ary}
+ Pseudonoeuds

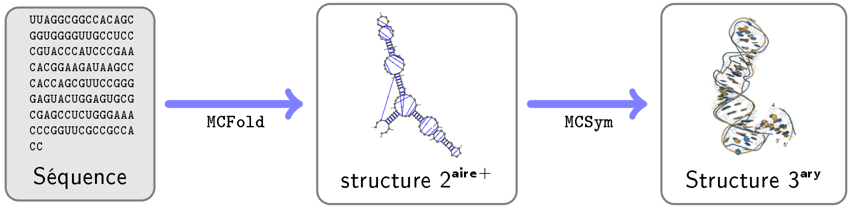
Dynamique
moléculaire

Structure 3^{aire}

Futur (proche) : Vers une prédiction 3D

But : De la séquence à des modèles tri-dimensionnels ! ! !

- Pipeline MC-Fold/MC-sym [PM08]



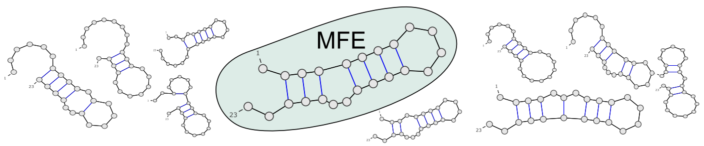
Résumé

Ensemble canonique de Boltzmann

L'ARN *respire* ⇒ 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.
⇒ Comprendre les modes d'actions de l'ARN exige de prendre en considération l'ensemble des structures.

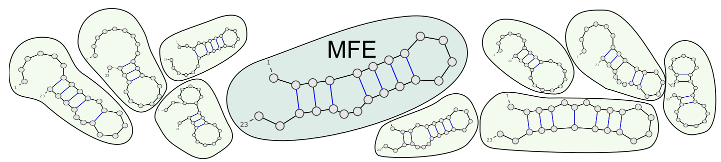
En particulier, des structures proches peuvent se *grouper* et devenir l'hypothèse la plus réaliste dans la recherche d'une conformation fonctionnelle.

Ensemble canonique de Boltzmann

L'ARN *respire* ⇒ 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.
⇒ 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.

Distribution de Boltzmann : Définition

Une distribution de Bolzmann 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 (1.986.10^{-3} kCal.K^{-1}.mol^{-1})

Distribution renormalisée sur $\mathcal{S}_\omega$ par la fonction de partition

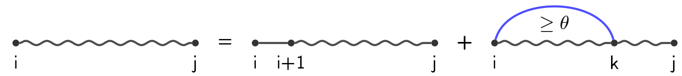
$$\mathcal{Z}_\omega = \sum_{S\in\mathcal{S}_\omega} e^{\frac{-E_{S,\omega}}{RT}}.$$

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

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

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

Décomposition de Nussinov/Jacobson



Récurrence sur l'énergie minimale d'un repliement :

$$\begin{aligned} N_{i,t} &= 0, \quad \forall t\in[i,i+\theta] \\ N_{i,j} &= \min\begin{cases} N_{i+1,j} & (i \text{ non apparié}) \\ \min_{k=i+\theta+1}^{j} E_{i,k}+N_{i+1,k-1}+N_{k+1,j} & (i \text{ comp. avec } k) \end{cases} \end{aligned}$$

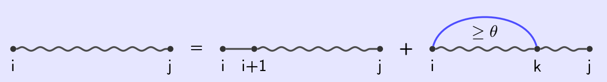
Récurrence de comptage des structures compatibles :

$$\begin{aligned} C_{i,t} &= 1, \quad \forall t\in[i,i+\theta] \\ C_{i,j} &= \sum\begin{cases} C_{i+1,j} & (i \text{ non apparié}) \\ \sum_{k=i+\theta+1}^{j} 1\times C_{i+1,k-1}\times C_{k+1,j} & (i \text{ comp. avec } k) \end{cases} \end{aligned}$$

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

Fonction de partition

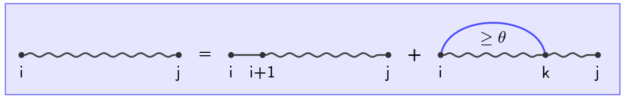
Fonction de partition = Comptage pondéré des structures compatibles



$$\begin{aligned} \mathcal{Z}_{i,t} &= 1, \quad \forall t\in[i,i+\theta] \\ \mathcal{Z}_{i,j} &= \sum\begin{cases} \mathcal{Z}_{i+1,j} \\ \sum_{k=i+\theta+1}^{j} 1\times \mathcal{Z}_{i+1,k-1}\times \mathcal{Z}_{k+1,j} \end{cases} \end{aligned}$$

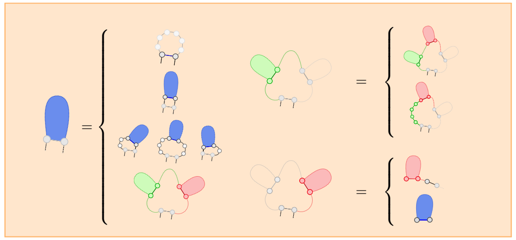
Fonction de partition

Fonction de partition = Comptage pondéré des structures compatibles



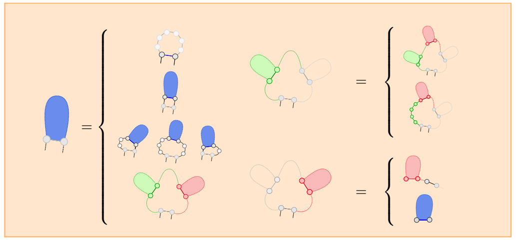
$$\begin{aligned} \mathcal{Z}_{i,t} &= 1, \quad \forall t\in[i,i+\theta] \\ \mathcal{Z}_{i,j} &= \sum\begin{cases} \mathcal{Z}_{i+1,j} \\ \sum_{k=i+\theta+1}^{j} e^{\frac{-E_{bp}(i,k)}{RT}}\times \mathcal{Z}_{i+1,k-1}\times \mathcal{Z}_{k+1,j} \end{cases} \end{aligned}$$

Fonction de partition = Comptage pondéré des structures compatibles



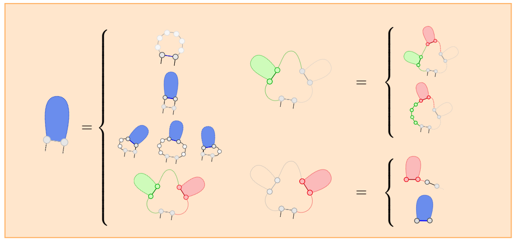
$$
\begin{aligned}
\mathcal{M}'_{i,j} &= \text{Min}\begin{cases} E_H(i,j) \\ E_S(i,j) + \mathcal{M}'_{i+1,j-1} \\ \text{Min}(E_{BI}(i,i',j',j) + \mathcal{M}'_{i',j'}) \\ a + c + \text{Min}\left(\mathcal{M}_{i+1,k-1} + \mathcal{M}^1{}_{k,j-1}\right) \end{cases} \\
\mathcal{M}_{i,j} &= \text{Min}\left\{\text{Min}\left(\mathcal{M}_{i,k-1},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

Fonction de partition = Comptage pondéré des structures compatibles



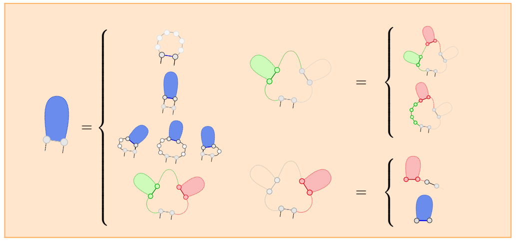
$$
\begin{aligned}
\mathcal{M}'_{i,j} &= \text{Min}\begin{cases} e^{\frac{-E_H(i,j)}{RT}} \\ e^{\frac{-E_S(i,j)}{RT}} + \mathcal{M}'_{i+1,j-1} \\ \text{Min}\left(e^{\frac{-E_{BI}(i,i',j',j)}{RT}} + \mathcal{M}'_{i',j'}\right) \\ e^{\frac{-(a+c)}{RT}} + \text{Min}\left(\mathcal{M}_{i+1,k-1} + \mathcal{M}^1{}_{k,j-1}\right)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

Fonction de partition = Comptage pondéré des structures compatibles



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

Fonction de partition = Comptage pondéré des structures compatibles



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

Fonction de partition

Fonction de partition = Comptage pondéré des structures compatibles

$$\begin{aligned}\mathcal{Z}_{i,t} &= 1,\quad \forall t\in[i,i+\theta]\\ \mathcal{Z}_{i,j} &= \sum\left\{\begin{array}{l}\mathcal{Z}_{i+1,j}\\ \displaystyle\sum_{k=i+\theta+1}^{j} e^{\frac{-E_{bp}(i,k)}{RT}}\times\mathcal{Z}_{i+1,k-1}\times\mathcal{Z}_{k+1,j}\end{array}\right.\end{aligned}$$

Validité de la fonction de partition :

- Exhaustivité/non ambiguïté du schéma

Fonction de partition

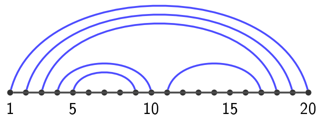
Fonction de partition = Comptage pondéré des structures compatibles

$$\begin{aligned}\mathcal{Z}_{i,t} &= 1,\quad \forall t\in[i,i+\theta]\\ \mathcal{Z}_{i,j} &= \sum\left\{\begin{array}{l}\mathcal{Z}_{i+1,j}\\ \displaystyle\sum_{k=i+\theta+1}^{j} e^{\frac{-E_{bp}(i,k)}{RT}}\times\mathcal{Z}_{i+1,k-1}\times\mathcal{Z}_{k+1,j}\end{array}\right.\end{aligned}$$

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
  $\left(e^{-a/RT}.\mathcal{Z}^{1}.\mathcal{Z}' = e^{-a/RT}.\sum_x e^{-E_x/RT}.\sum_y e^{-E_y/RT} = \sum_{x,y} e^{-a/RT}.e^{-E_x/RT}.e^{-E_y/RT} = \sum_{x,y} e^{-(a+E_x+E_y)/RT}\right)$

Exemple :

$S =$ 

$\Rightarrow \mathcal{B}_{1,20} = \mathcal{B}_{1,20}$

Fonction de partition

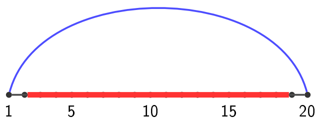
Fonction de partition = Comptage pondéré des structures compatibles

$$\begin{aligned}\mathcal{Z}_{i,t} &= 1,\quad \forall t\in[i,i+\theta]\\ \mathcal{Z}_{i,j} &= \sum\left\{\begin{array}{l}\mathcal{Z}_{i+1,j}\\ \displaystyle\sum_{k=i+\theta+1}^{j} e^{\frac{-E_{bp}(i,k)}{RT}}\times\mathcal{Z}_{i+1,k-1}\times\mathcal{Z}_{k+1,j}\end{array}\right.\end{aligned}$$

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
  $\left(e^{-a/RT}.\mathcal{Z}^{1}.\mathcal{Z}' = e^{-a/RT}.\sum_x e^{-E_x/RT}.\sum_y e^{-E_y/RT} = \sum_{x,y} e^{-a/RT}.e^{-E_x/RT}.e^{-E_y/RT} = \sum_{x,y} e^{-(a+E_x+E_y)/RT}\right)$

Exemple :

$S =$ 

$\Rightarrow \mathcal{B}_{1,20} = e^{\frac{1}{RT}}.\mathcal{B}_{2,19}$

Fonction de partition

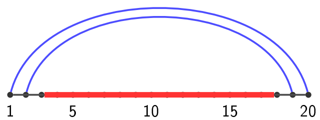
Fonction de partition = Comptage pondéré des structures compatibles

$$\begin{aligned}\mathcal{Z}_{i,t} &= 1,\quad \forall t\in[i,i+\theta]\\ \mathcal{Z}_{i,j} &= \sum\left\{\begin{array}{l}\mathcal{Z}_{i+1,j}\\ \displaystyle\sum_{k=i+\theta+1}^{j} e^{\frac{-E_{bp}(i,k)}{RT}}\times\mathcal{Z}_{i+1,k-1}\times\mathcal{Z}_{k+1,j}\end{array}\right.\end{aligned}$$

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
  $\left(e^{-a/RT}.\mathcal{Z}^{1}.\mathcal{Z}' = e^{-a/RT}.\sum_x e^{-E_x/RT}.\sum_y e^{-E_y/RT} = \sum_{x,y} e^{-a/RT}.e^{-E_x/RT}.e^{-E_y/RT} = \sum_{x,y} e^{-(a+E_x+E_y)/RT}\right)$

Exemple :

$S =$ 

$\Rightarrow \mathcal{B}_{1,20} = e^{\frac{-1}{RT}}.e^{\frac{1}{RT}}.\mathcal{B}_{3,18}$

Fonction de partition

Fonction de partition = Comptage pondéré des structures compatibles

$$\begin{aligned}\mathcal{Z}_{i,t} &= 1, \quad \forall t \in [i, i+\theta]\\ \mathcal{Z}_{i,j} &= \sum \begin{cases} \mathcal{Z}_{i+1,j} \\ \sum_{k=i+\theta+1}^{j} e^{\frac{-E_{\text{bp}}(i,k)}{RT}} \times \mathcal{Z}_{i+1,k-1} \times \mathcal{Z}_{k+1,j}\end{cases}\end{aligned}$$

Validité de la fonction de partition :

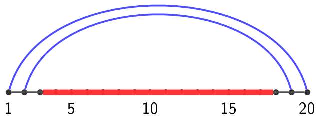
- Exhaustivité/non ambiguïté du schéma
- Correction du facteur de Boltzmann

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

  Contributions énergétiques passent à l'exposant

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

Exemple :

$S =$ 

$\Rightarrow \mathcal{B}_{1,20} = e^{\frac{2}{RT}}.\mathcal{B}_{3,18}$

Fonction de partition

Fonction de partition = Comptage pondéré des structures compatibles

$$\begin{aligned}\mathcal{Z}_{i,t} &= 1, \quad \forall t \in [i, i+\theta]\\ \mathcal{Z}_{i,j} &= \sum \begin{cases} \mathcal{Z}_{i+1,j} \\ \sum_{k=i+\theta+1}^{j} e^{\frac{-E_{\text{bp}}(i,k)}{RT}} \times \mathcal{Z}_{i+1,k-1} \times \mathcal{Z}_{k+1,j}\end{cases}\end{aligned}$$

Validité de la fonction de partition :

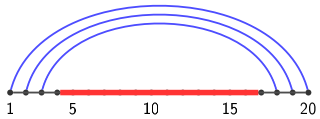
- Exhaustivité/non ambiguïté du schéma
- Correction du facteur de Boltzmann

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

  Contributions énergétiques passent à l'exposant

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

Exemple :

$S =$ 

$\Rightarrow \mathcal{B}_{1,20} = e^{\frac{3}{RT}}.\mathcal{B}_{4,17}$

Fonction de partition

Fonction de partition = Comptage pondéré des structures compatibles

$$\begin{aligned}\mathcal{Z}_{i,t} &= 1, \quad \forall t \in [i, i+\theta]\\ \mathcal{Z}_{i,j} &= \sum \begin{cases} \mathcal{Z}_{i+1,j} \\ \sum_{k=i+\theta+1}^{j} e^{\frac{-E_{\text{bp}}(i,k)}{RT}} \times \mathcal{Z}_{i+1,k-1} \times \mathcal{Z}_{k+1,j}\end{cases}\end{aligned}$$

Validité de la fonction de partition :

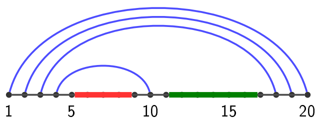
- Exhaustivité/non ambiguïté du schéma
- Correction du facteur de Boltzmann

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

  Contributions énergétiques passent à l'exposant

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

Exemple :

$S =$ 

$\Rightarrow \mathcal{B}_{1,20} = e^{\frac{4}{RT}}.\mathcal{B}_{5,9}.\mathcal{B}_{11,17}$

Fonction de partition

Fonction de partition = Comptage pondéré des structures compatibles

$$\begin{aligned}\mathcal{Z}_{i,t} &= 1, \quad \forall t \in [i, i+\theta]\\ \mathcal{Z}_{i,j} &= \sum \begin{cases} \mathcal{Z}_{i+1,j} \\ \sum_{k=i+\theta+1}^{j} e^{\frac{-E_{\text{bp}}(i,k)}{RT}} \times \mathcal{Z}_{i+1,k-1} \times \mathcal{Z}_{k+1,j}\end{cases}\end{aligned}$$

Validité de la fonction de partition :

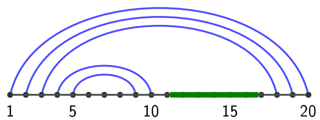
- Exhaustivité/non ambiguïté du schéma
- Correction du facteur de Boltzmann

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

  Contributions énergétiques passent à l'exposant

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

Exemple :

$S =$ 

$\Rightarrow \mathcal{B}_{1,20} = e^{\frac{5}{RT}}.\mathcal{B}_{11,17}$

Fonction de partition

Fonction de partition = Comptage pondéré des structures compatibles

$$\mathcal{Z}_{i,t} = 1, \quad \forall t \in [i, i+\theta]$$
$$\mathcal{Z}_{i,j} = \sum \begin{cases} \mathcal{Z}_{i+1,j} \\ \sum_{k=i+\theta+1}^{j} e^{\frac{-E_{bp}(i,k)}{RT}} \times \mathcal{Z}_{i+1,k-1} \times \mathcal{Z}_{k+1,j} \end{cases}$$

Validité de la fonction de partition :

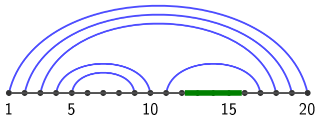
- Exhaustivité/non ambiguïté du schéma
- Correction du facteur de Boltzmann

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

  Contributions énergétiques passent à l'exposant

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

Exemple :

$S =$ 

$\Rightarrow \mathcal{B}_{1,20} = e^{\frac{6}{RT}} . \mathcal{B}_{12,16}$

Fonction de partition

Fonction de partition = Comptage pondéré des structures compatibles

$$\mathcal{Z}_{i,t} = 1, \quad \forall t \in [i, i+\theta]$$
$$\mathcal{Z}_{i,j} = \sum \begin{cases} \mathcal{Z}_{i+1,j} \\ \sum_{k=i+\theta+1}^{j} e^{\frac{-E_{bp}(i,k)}{RT}} \times \mathcal{Z}_{i+1,k-1} \times \mathcal{Z}_{k+1,j} \end{cases}$$

Validité de la fonction de partition :

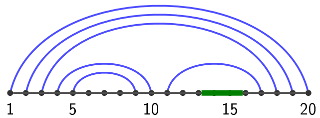
- Exhaustivité/non ambiguïté du schéma
- Correction du facteur de Boltzmann

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

  Contributions énergétiques passent à l'exposant

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

Exemple :

$S =$ 

$\Rightarrow \mathcal{B}_{1,20} = e^{\frac{6}{RT}} . \mathcal{B}_{13,16}$

Fonction de partition

Fonction de partition = Comptage pondéré des structures compatibles

$$\mathcal{Z}_{i,t} = 1, \quad \forall t \in [i, i+\theta]$$
$$\mathcal{Z}_{i,j} = \sum \begin{cases} \mathcal{Z}_{i+1,j} \\ \sum_{k=i+\theta+1}^{j} e^{\frac{-E_{bp}(i,k)}{RT}} \times \mathcal{Z}_{i+1,k-1} \times \mathcal{Z}_{k+1,j} \end{cases}$$

Validité de la fonction de partition :

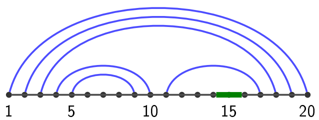
- Exhaustivité/non ambiguïté du schéma
- Correction du facteur de Boltzmann

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

  Contributions énergétiques passent à l'exposant

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

Exemple :

$S =$ 

$\Rightarrow \mathcal{B}_{1,20} = e^{\frac{6}{RT}} . \mathcal{B}_{14,16}$

Fonction de partition

Fonction de partition = Comptage pondéré des structures compatibles

$$\mathcal{Z}_{i,t} = 1, \quad \forall t \in [i, i+\theta]$$
$$\mathcal{Z}_{i,j} = \sum \begin{cases} \mathcal{Z}_{i+1,j} \\ \sum_{k=i+\theta+1}^{j} e^{\frac{-E_{bp}(i,k)}{RT}} \times \mathcal{Z}_{i+1,k-1} \times \mathcal{Z}_{k+1,j} \end{cases}$$

Validité de la fonction de partition :

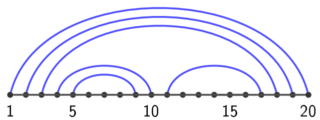
- Exhaustivité/non ambiguïté du schéma
- Correction du facteur de Boltzmann

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

  Contributions énergétiques passent à l'exposant

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

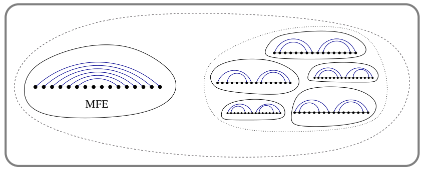
Exemple :

$S =$ 

$\Rightarrow \mathcal{B}_{1,20} = e^{\frac{6}{RT}} = e^{\frac{-E_S}{RT}}$

Échantillonnage statistique de structures d'ARN

La MFE (Probabilité maximale) peut être *écrasée* 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 `SFold` [DL03])
Précalcul : Calculer les matrices ($\mathcal{Z}, \mathcal{Z}', \mathcal{Z}^1$) des fonctions de partition.
Remontée stochastique :

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

Remontée stochastique

Algorithme (Reformulation `SFold` [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 $[0, \mathcal{Z}'(i,j))$

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

Remontée stochastique

Algorithme (Reformulation `SFold` [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 $[0, \mathcal{Z}'(i,j))$

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

r

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

Remontée stochastique

Algorithme (Reformulation SFold [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 $r < 0$

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

r

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

Remontée stochastique

Algorithme (Reformulation SFold [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 $r < 0$

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

r

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

Yann Ponty Cours M2 BIM - Séance 2 - Boltzmann et comparaison

Remontée stochastique

Algorithme (Reformulation SFold [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 $r < 0$

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

r

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

Yann Ponty Cours M2 BIM - Séance 2 - Boltzmann et comparaison

Remontée stochastique

Algorithme (Reformulation SFold [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 $r < 0$

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

r

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

Yann Ponty Cours M2 BIM - Séance 2 - Boltzmann et comparaison

Remontée stochastique

Algorithme (Reformulation SFold [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. Réitérer sur les sous-structures

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

Correction : Chaque terme de la décomposition engendre $\mathcal{T} \in \{\mathcal{A}_1, \ldots, \mathcal{C}_j\}$, 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 $S \in \mathcal{S}_\omega$ est engendrée uniquement (Unambiguité de Turner) par une séquence de choix d'ensembles $\mathcal{S}_w \supset E_1 \supset E_2 \supset \ldots \supset \{S\}$.
La probabilité d'engendrer S est donc $p_S = \frac{\mathcal{B}(E_1)}{\mathcal{B}(\mathcal{S}_w)} \cdot \frac{\mathcal{B}(E_2)}{\mathcal{B}(E_1)} \cdot \frac{\mathcal{B}(E_3)}{\mathcal{B}(E_2)} \cdots \frac{\mathcal{B}(\{S\})}{\mathcal{B}(E_m)}$

Remontée stochastique

Algorithme (Reformulation SFold [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. Réitérer sur les sous-structures

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

Correction : Chaque terme de la décomposition engendre $\mathcal{T} \in \{\mathcal{A}_1, \ldots, \mathcal{C}_j\}$, 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 $S \in \mathcal{S}_\omega$ est engendrée uniquement (Unambiguité de Turner) par une séquence de choix d'ensembles $\mathcal{S}_w \supset E_1 \supset E_2 \supset \ldots \supset \{S\}$.
La probabilité d'engendrer S est donc $p_S = \frac{1}{\mathcal{B}(\mathcal{S}_w)} \cdot \frac{1}{1} \cdot \frac{1}{1} \cdots \frac{\mathcal{B}(\{S\})}{1}$

Remontée stochastique

Algorithme (Reformulation SFold [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. Réitérer sur les sous-structures

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

Correction : Chaque terme de la décomposition engendre $\mathcal{T} \in \{\mathcal{A}_1, \ldots, \mathcal{C}_j\}$, 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 $S \in \mathcal{S}_\omega$ est engendrée uniquement (Unambiguité de Turner) par une séquence de choix d'ensembles $\mathcal{S}_w \supset E_1 \supset E_2 \supset \ldots \supset \{S\}$.
La probabilité d'engendrer S est donc $p_S = \frac{\mathcal{B}(\{S\})}{\mathcal{B}(\mathcal{S}_w)} = \frac{e^{-E_s/RT}}{\mathcal{Z}} = P_{S,\omega}$

Complexité

Algorithme (Reformulation SFold [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. Réitérer sur les sous-structures

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

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

Complexité

Algorithme (Reformulation `SFold` [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. Réitérer sur les sous-structures

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

$\boxed{r}$

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

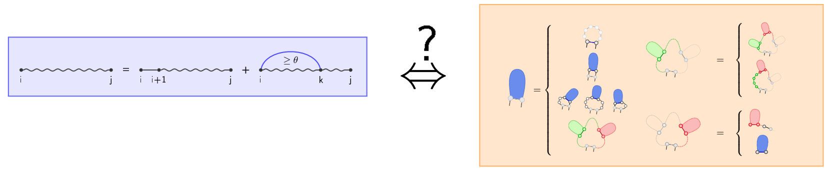
Après $\Theta(n)$ opérations, on réitère sur un interval de taille $n-1$
$\Rightarrow$ Complexité du cas au pire en $\mathcal{O}(n^2 k)$ pour k échantillons

Remarque : Instance pondérée d'un problème de génération aléatoire par la méthode *récursive* [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.

Validité d'un schéma

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).

Validité d'un schéma

Une preuve de correction possible :

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

$$C_{i,t} = 1, \quad \forall t \in [i, i+\theta]$$
$$C_{i,j} = \sum \begin{cases} C_{i+1,j} \\ \sum_{k=i+\theta+1}^{j} 1 \times C_{i+1,k-1} \times C_{k+1,j} \end{cases}$$

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

$\overset{?}{\Leftrightarrow}$

$$\mathcal{C}'_{i,j} = \sum \begin{cases} 1 \\ \mathcal{C}'_{i+1,j-1} \\ \sum_{i',j'} \mathcal{C}'_{i',j'} \\ \sum_k \mathcal{C}_{i+1,k-1} \times \mathcal{C}^1_{k,j-1} \end{cases}$$
$$\mathcal{C}_{i,j} = \sum_k \left((\mathcal{C}_{i,k-1} + 1) \times \mathcal{C}^1_{k,j} \right)$$
$$\mathcal{C}^1_{i,j} = \mathcal{C}^1_{i,j-1} + \mathcal{C}'_{i,j}$$

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

Forte certitude mais pas encore preuve (Séries génératrices).

Repliement sous-optimal

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

$\mathcal{M}'_{1,n,\Delta}$

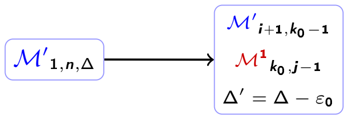
- $a + c + \text{Min}\left(\mathcal{M}_{i+1,k_0-1} + \mathcal{M}^1_{k_0,j-1}\right)$ — $E_0 - \mathcal{M}'_{1,n} = \varepsilon_0 \leq \Delta$
- $a + c + \text{Min}\left(\mathcal{M}_{i+1,k_1-1} + \mathcal{M}^1_{k_1,j-1}\right)$ — $E_1 - \mathcal{M}'_{1,n} = \varepsilon_1 > \Delta$
- $a + c + \text{Min}\left(\mathcal{M}_{i+1,k_2-1} + \mathcal{M}^1_{k_2,j-1}\right)$ — $E_2 - \mathcal{M}'_{1,n} = \varepsilon_2 \leq \Delta$

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.

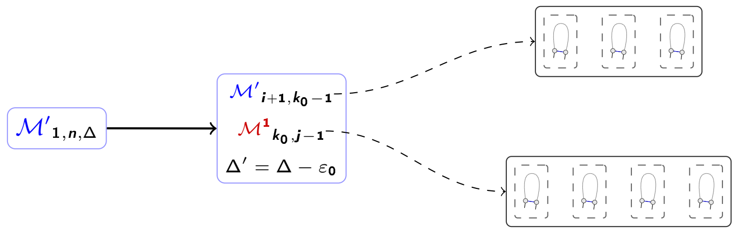


Repliement sous-optimal

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)

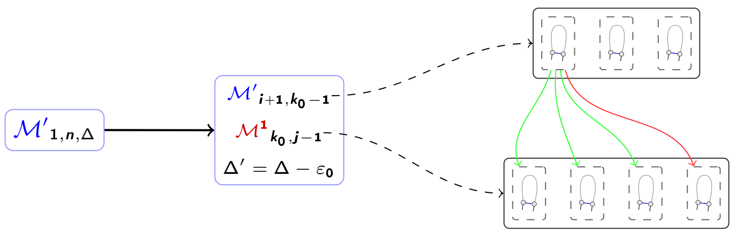


Repliement sous-optimal

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)

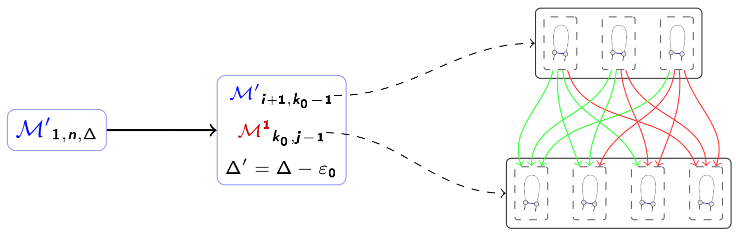


Repliement sous-optimal

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.

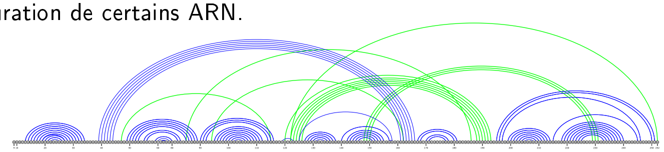
⇒ Engendrer des repliements sous-optimaux (`RNASubopt` [WFHS99]), i.e. construire toutes les structures à Δ KCal.mol^{-1} de la MFE :

- 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, mais bon...)

Repliement avec pseudo-noeuds

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



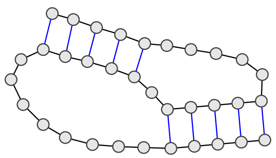
Ribozyme du groupe I

Leur absence historique au sein algorithmes de repliement est liée à la difficulté algorithmique des problèmes associés (Présence de pseudo-noeuds brise l'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

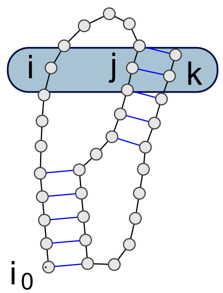
Le but est de capturer des catégorie de pseudo-noeuds *simples*, mais très représentée.



Idée : Quand on retourne ce type de pseudonoeuds, il suffit de précalculer les meilleures configurations *en dessous* d'un triplet (i, j, k), puis de regarder les configurations locales.

Algorithme d'Akutsu/Uemura

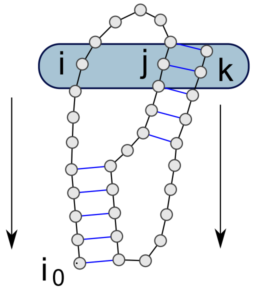
Le but est de capturer des catégorie de pseudo-noeuds *simples*, mais très représentée.



Idée : Quand on retourne ce type de pseudonoeuds, il suffit de précalculer les meilleures configurations *en dessous* d'un triplet (i, j, k), puis de regarder les configurations locales.

Algorithme d'Akutsu/Uemura

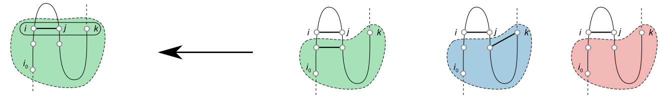
Le but est de capturer des catégorie de pseudo-noeuds *simples*, mais très représentée.



Idée : Quand on retourne ce type de pseudonoeuds, il suffit de précalculer les meilleures configurations *en dessous* d'un triplet (i, j, k), puis de regarder les configurations locales.

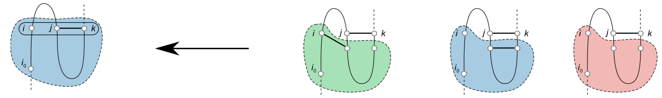
Akutsu/Uemura : Programmation dynamique

r : Séquence d'ARN
$\Delta_{i,j}$: Énergie de la paire (r_i, r_j).



$$S_L^{i_0}(i,j,k) = \Delta_{i,j} + \min \left\{ \begin{array}{c} S_L^{i_0}(i-1,j+1,k), \\ S_M^{i_0}(i-1,j+1,k), \\ S_R^{i_0}(i-1,j+1,k) \end{array} \right\},$$

$$S_L^{i_0}(i,j,j) = \Delta_{i,j}, \quad \forall i < j$$

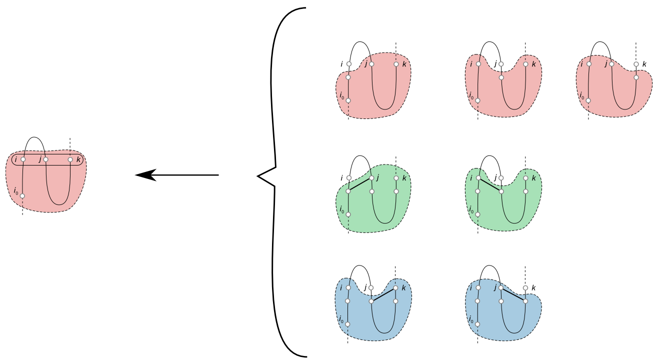


$$S_R^{i_0}(i,j,k) = \Delta_{j,k} + \min \left\{ \begin{array}{c} S_L^{i_0}(i,j+1,k-1), \\ S_M^{i_0}(i,j+1,k-1), \\ S_R^{i_0}(i,j+1,k-1) \end{array} \right\},$$

$$S_R^{i_0}(i_0-1,j,j+\theta+1) = \Delta_{i,j+\theta+1}, \quad \forall j$$

Akutsu/Uemura : Programmation dynamique

r : Séquence d'ARN
$\Delta_{i,j}$: Énergie de la paire (r_i, r_j).



$$S_M^{i_0}(i,j,k) = \min \left\{ \begin{array}{c} S_M^{i_0}(i-1,j,k), S_M^{i_0}(i,j+1,k), S_M^{i_0}(i,j,k-1), \\ S_L^{i_0}(i-1,j,k), S_L^{i_0}(i,j+1,k) \\ S_R^{i_0}(i,j+1,k), S_R^{i_0}(i,j,k-1) \end{array} \right\}$$

$$S_L^{i_0}(i_0-1,j,k) = S_R^{i_0}(i_0-1,j,k) = S_M^{i_0}(i_0-1,j,k) = 0, \quad \forall j,k \text{ tq } k-j \leq \theta$$

Akutsu/Uemura : Programmation dynamique

r : Séquence d'ARN
$\Delta_{i,j}$: Énergie de la paire (r_i, r_j).
L'équation générale sur le nombre de paires de bases pour la présence, sur l'intervalle (i_0, k_0), de ce type de pseudo-noeuds est alors donné par

$$S_P(i_0,k_0) = \min_{i_0 \leq i < j < k \leq k_0} \left(S_L^{i_0}(i,j,k), S_R^{i_0}(i,j,k), S_M^{i_0}(i,j,k) \right)$$

On *insère* ces pseudonoeuds au sein d'une structure secondaire classique au moyen d'une variante de Nussinov

$$S(i,j) = \min \left(S_P(i,j), S(i+1,j-1) + \Delta_{i,j}, \min_{i<k\leq j} (S(i,k-1) + S(k,j)) \right)$$

En utilisant une astuce dans l'ordre des calculs, on arrive à faire tomber la complexité à $\mathcal{O}(n^4)$ dans un modèle de Nussinov, mais on reste en $\mathcal{O}(n^5)$ dans le modèle de Turner.

Résumé

1. Performances et approches comparatives
   - Vers une prédiction ab-initio 3D
2. Ensemble de Boltzmann
   - Ensemble de Boltzmann
   - Nussinov : Minimisation $\Rightarrow$ Comptage
   - Calcul de la fonction de partition
   - Échantillonnage statistique
3. Extensions
   - Validité d'un schéma
   - Structures sous-optimales
   - Pseudo-noeuds
4. Alignement et comparaison de structures d'ARN
   - Méthode géométrique
   - Alignement de structures secondaires
   - Méthodes hybrides

Pourquoi aligner structurellement des ARN

Une pression évolutive commune permet d'identifier une fonction commune. Chez certains organismes (et pour certaines familles d'ARN), très faible conservation de la séquence. Cependant, la structure peut être bien plus conservée, et connue (Expérimentalement) ou déterminée par repliement.

Problèmes :

- Édition : Trouver la *distance* entre deux structures A et B. Quelle est la séquence d'opérations de coût minimal permettant de passer de A à B ? Déjà NP-complet pour deux structures secondaires [BFRS07].
- Alignement : Trouver une super-structure de coût minimal. Généralise la notion d'alignement de séquence. Polynomial pour des structures secondaires [BDD+08], NP-complet en 3D [SZS+08].
  Variantes : Alignement local ou global, Recherche de motifs.
- Superposition : Trouver une transformation géométrique (Rotation, translation, zoom) pour superposer *au mieux* les coordonnées de deux ARN de matching connu. Polynomial en 3D [McL82].

$\Rightarrow$ La difficulté algorithme provient de la recherche du matching initial.

FR3D : Un exemple d'approche géométrique

Quand les structures tertiaires (3D) des ARN sont connues, le problème de l'alignement peut être abordé de façon purement géométrique.

Problème

Donnée : Motif m et structure cible b (Ensembles de bases 3D).
Résultat : Matching de m et d'un sous-ensemble de b minimisant une divergence géométrique.

Divergence géométrique : Dans FR3D [SZS+08], une fonction D basée sur deux fonctions L et A d'erreur tenant compte respectivement de la superposabilité (L) et de l'orientation des bases (A) de m et b.

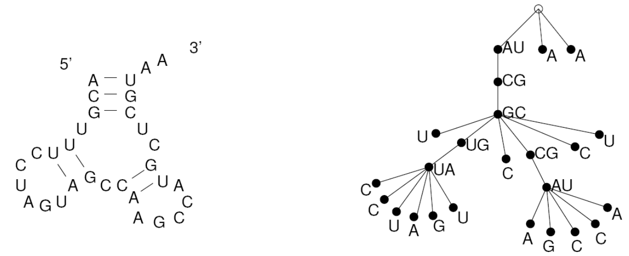
$$L = \sqrt{\min_{R,T} \sum_{i=1}^{m} \|b_i - R(T(m_i))\|^2} \qquad A = \sqrt{\sum_{i=1}^{m} \alpha_i^2} \qquad D = \frac{1}{m}\sqrt{L^2 + A^2}$$

R, T : Rotation et translation. c_i : Barycentre pour la base m_i. α_i : Écart entre les axes barycentre/bases dans m_i et b_i.

Exploration (Backtrack) + Élagage incrémental (Bornes sur D) $\Rightarrow$ Explosion. Mais recherche exacte pour des petits motifs.

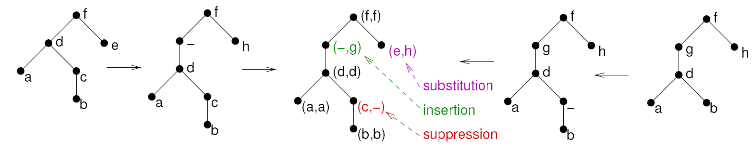
Structures vers arbres

L'alignement de deux structures secondaires est basé sur une représentation arborescente de la structure secondaire[1].



Paires de bases $\Rightarrow$ noeuds internes   Bases non-appariées $\Rightarrow$ Feuilles

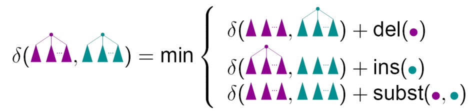
Alignement = Construction d'un matching complet de coût minimal.



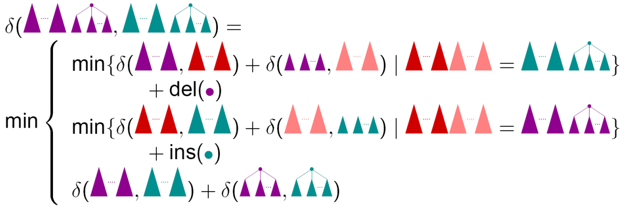
1. Illustrations empruntées à C. Herrbach

Alignement historique : Jiang, Wang & Zhang 95 [JWZ94]

Alignement d'arbre [2]



Alignement de forêt

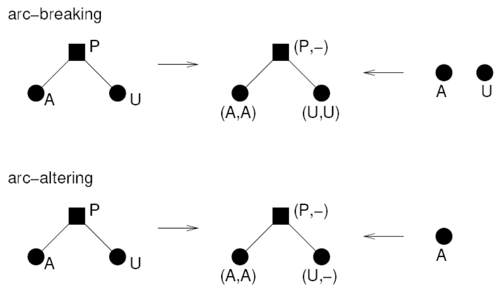


Complexité au pire en $\mathcal{O}(n^4)$ [JWZ94], en moyenne en $\mathcal{O}(n^2)$ [HDD07].
Mais opérations spécifiques à l'ARN manquantes.

2. Idem

RNAForester [HVG04]

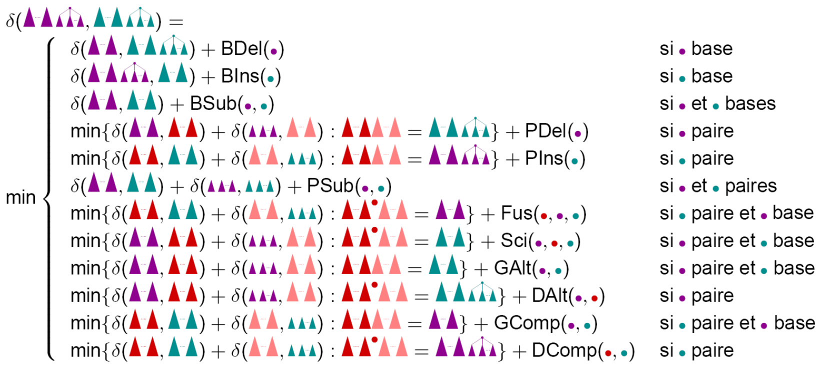
Basé sur l'algorithme de Jiang, Wang & Zhang
+ Intégrations d'opérations spécifiques à l'ARN [3].



Possibilité de paramètrer les coûts des opérations, mais certaines opérations atomiques dans un modèle réalistes doivent être recomposées à partir des opérations disponibles. Par exemple, la substitution d'un sommet par une feuille est interdite directement.

3. Idem

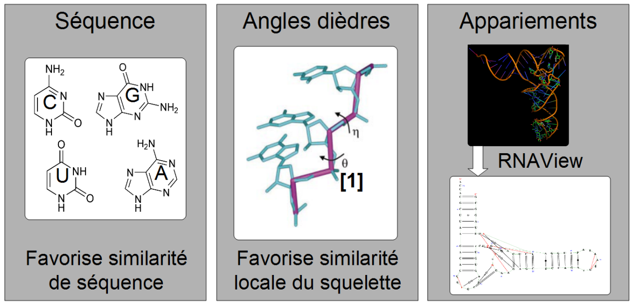
NestedAlign [BDD$^+$08]



Méthode hybride

DIAL [FPLC07] est une méthode hybride qui se concentre sur les comportements locaux.
Idée : L'ARN est flexible, petite variation locale peuvent entraîner des grandes déviations géométriques.

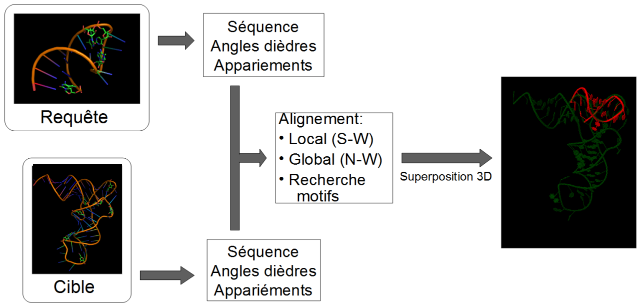
DIAL capture les similarités locales à trois niveau :



Méthode hybride

DIAL [FPLC07] est une méthode hybride qui se concentre sur les comportements locaux.
Idée : L'ARN est flexible, petite variation locale peuvent entraîner des grandes déviations géométriques.

Un algorithme d'alignement de séquence est alors utilisé



Conclusion

Tout dépend de ce que l'on a et veut :

- Modèle 3D :
  - Recherche d'un motif peu conservé en séquence : FR3D
  - Recherche d'un motif conservé : FR3D, DIAL ou DARTS
  - Recherche d'une structure entière : DIAL ou DARTS
- Structure secondaire :
  - Recherche d'un motif : NestedAlign
  - Alignement structure : RNAForester, NestedAlign

De nombreux autres programmes disponibles : Migal, Magnolia, ...
+ Explosion des approches *par fragments* : FASTR3D, RNA FRABASE, ...

References I

Tatsuya Akutsu.
Dynamic programming algorithms for rna secondary structure prediction with pseudoknots.
Discrete Appl. Math., 104(1-3) :45–62, 2000.

G. Blin, A. Denise, S. Dulucq, C. Herrbach, and H. Touzet.
Alignment of rna structures.
Transactions on Computational Biology and Bioinformatics, ... :..., 2008.
A paraître.

Guillaume Blin, Guillaume Fertin, Irena Rusu, and Christine Sinoquet.
Extending the Hardness of RNA Secondary Structure Comparison.
In Bo Chen, Mike Paterson, and Guochuan Zhang, editors, *ESCAPE'07*, volume 4614 of *LNCS*, pages 140–151, Hangzhou, China, Apr 2007.

K. Doshi, J. J. Cannone, C. Cobaugh, and R. R. Gutell.
Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for rna secondary structure prediction.
BMC Bioinformatics, 5(1) :105, 2004.

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.

References II

F. Ferrè, Y. Ponty, W. A. Lorenz, and Peter Clote.
Dial : A web server for the pairwise alignment of two RNA 3-dimensional structures using nucleotide, dihedral angle and base pairing similarities.
Nucleic Acids Research, 35(Web server issue) :W659–668, July 2007.

P. Gardner and R. Giegerich.
A comprehensive comparison of comparative rna structure prediction approaches.
BMC Bioinformatics, 5(1) :140, 2004.

Claire Herrbach, Alain Denise, and Serge Dulucq.
Average complexity of the jiang-wang-zhang pairwise tree alignment algorithm and of a rna secondary structure alignment algorithm.
In *Proceedings of MACIS 2007, Second International Conference on Mathematical Aspects of Computer and Information Sciences*, 2007.

M. Hochsmann, B. Voss, and R. Giegerich.
Pure multiple RNA secondary structure alignments : A progressive profile approach.
IEEE/ACM Transactions on Computational Biology and Bioinformatics, 01(1) :53–62, 2004.

Tao Jiang, Lusheng Wang, and Kaizhong Zhang.
Alignment of trees - an alternative to tree edit.
In *CPM '94 : Proceedings of the 5th Annual Symposium on Combinatorial Pattern Matching*, pages 75–86, London, UK, 1994. Springer-Verlag.

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. McLachlan.
Rapid comparison of protein structures.
Acta cristallographica A, 38(6) :871–873, 1982.

References III

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.

M. Parisien and F. Major.
The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data.
Nature, 452(7183) :51–55, 2008.

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.

E. Rivas and S.R. Eddy.
A dynamic programming algorithm for RNA structure prediction including pseudoknots.
J Mol Biol, 285 :2053–2068, 1999.

B. A. Shapiro, Y. G. Yingling, W. Kasprzak, and E. Bindewald.
Bridging the gap in rna structure prediction.
Curr Opin Struct Biol, 17(2) :157–165, Apr 2007.

M. Sarver, C. Zirbel, J. Stombaugh, A. Mokdad, and N. B. Leontis.
FR3D : Finding local and composite recurrent structural motifs in RNA 3D.
Journal of Mathematical Biology, 56(1–2) :215–252, January 2008.

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.