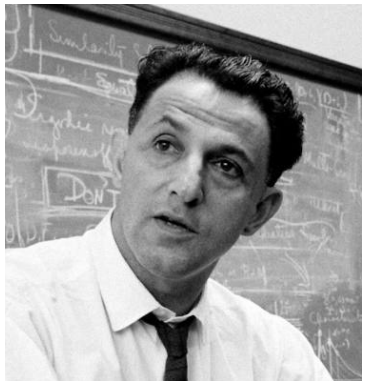
Combinatoire et repliement algorithmique des ARN

Yann Ponty

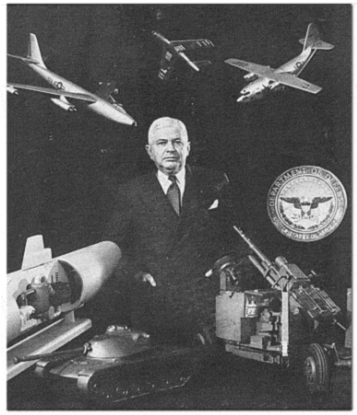
Polytechnique/CNRS/INRIA AMIB – France

22 Mars 2010

Programmation dynamique



R. Bellman



C. E. Wilson

Programmation dynamique = Technique algorithmique de résolution des problèmes d'optimisation.

Principe : Partant d'une équation liant le score optimal d'un (sous-)problème à celui de certains de ses sous-problèmes,

- Calculer efficacement le score optimal du problème initial
- Reconstruire la (ou les) solution(s) associée(s) au score optimal

Technique robuste face à des altérations *locales* de la fonction objectif
$\Rightarrow$ Omniprésente en bioinformatique : Alignement, repliement...

Difficulté : Trouver l'équation de programmation dynamique

Propriétés requises d'une équation de programmation dynamique =

1. Complétude* dans le parcours de l'espace des solutions
2. Correction du score optimal hérité des sous-problèmes
3. Non-ambiguité de la décomposition

Modéles combinatoires peuvent aider :

Grammaire/spécification $\rightsquigarrow$ Récurrences de comptage
$\Rightarrow$ Complétude et non-ambiguité

\+ Correction
$\Rightarrow$ Équation de programmation dynamique

Difficulté : Trouver l'équation de programmation dynamique

Propriétés souhaitables d'une équation de programmation dynamique =

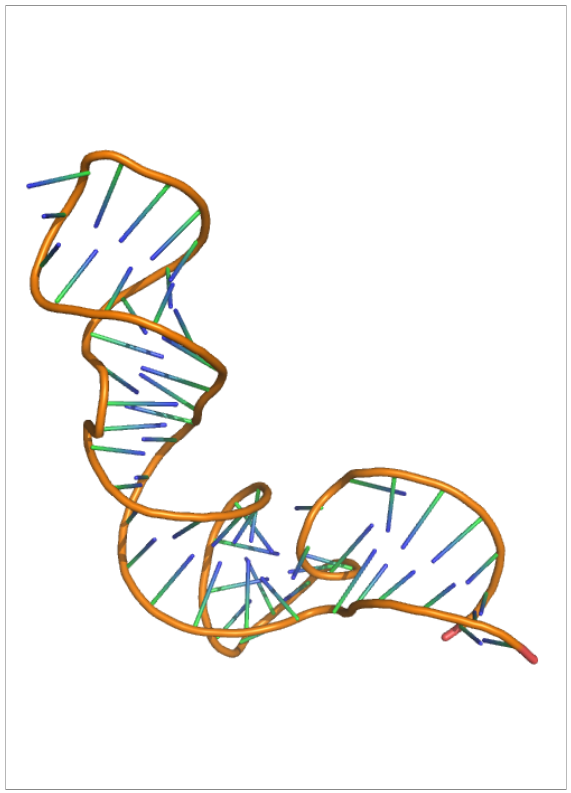
1. Complétude* dans le parcours de l'espace des solutions
2. Correction du score optimal hérité des sous-problèmes
3. Non-ambiguité de la décomposition

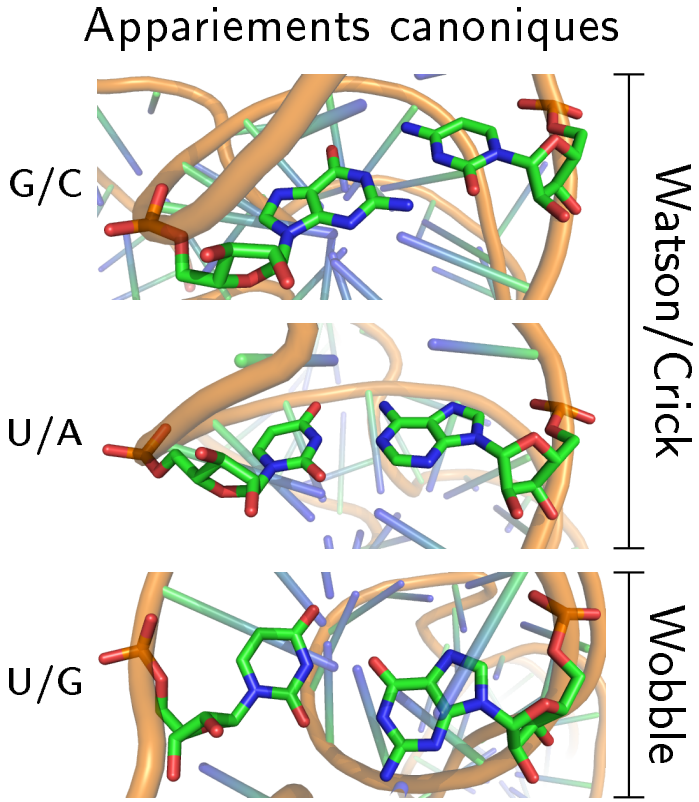
Modéles combinatoires peuvent aider :

Grammaire/spécification $\rightsquigarrow$ Récurrences de comptage
$\Rightarrow$ Complétude et non-ambiguité

$+$ Correction
$\Rightarrow$ Équation de programmation dynamique

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

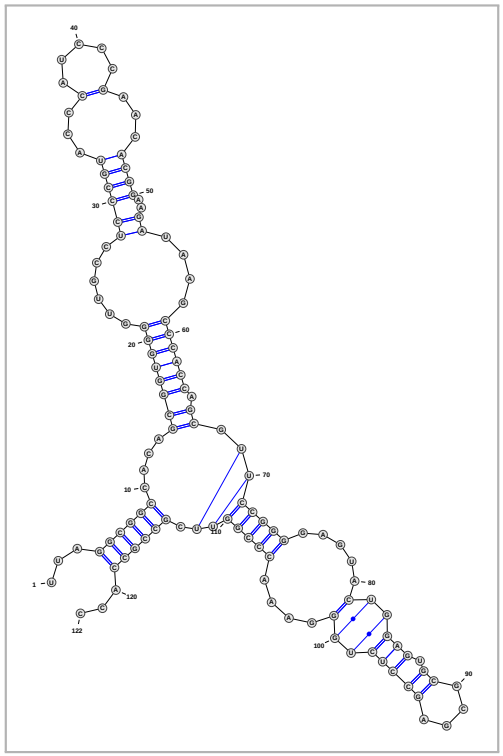


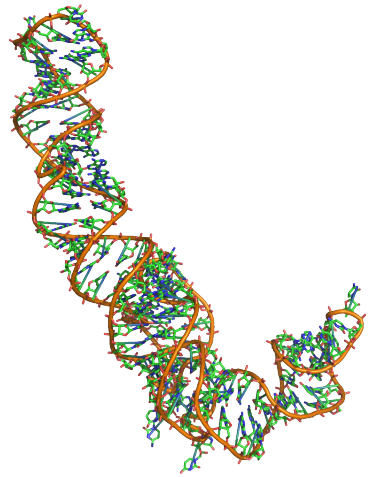


Repliement de l'ARN = Processus stochastique continu dirigé par (résultant en) un appariement des nucléotides.

Trois niveaux de représentation :

UUAGGCGGCCACAGC
GGUGGGGUUGCCUCC
CGUACCCAUCCCGAA
CACGGAAGAUAAGCC
CACCAGCGUUCCGGG
GAGUACUGGAGUGCG
CGAGCCUCUGGGAAA
CCCGGUUCGCCGCCA
CC



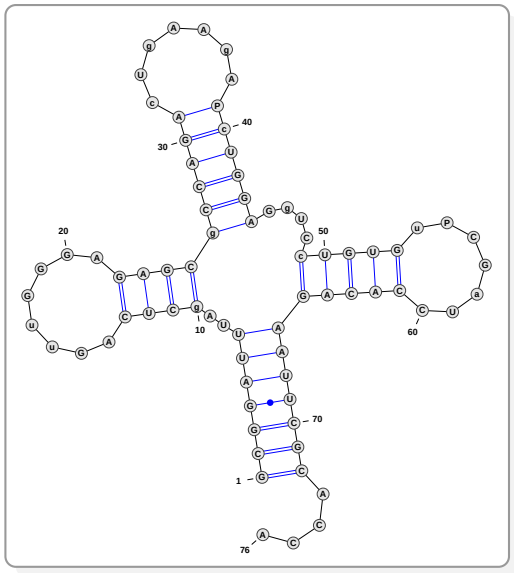


Source: 5s rRNA

(PDBID: 1K73:B)

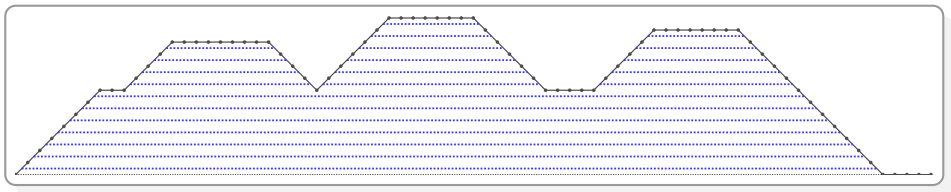
Structure primaire | Structure secondaire | Structure tertiaire

Structure secondaire

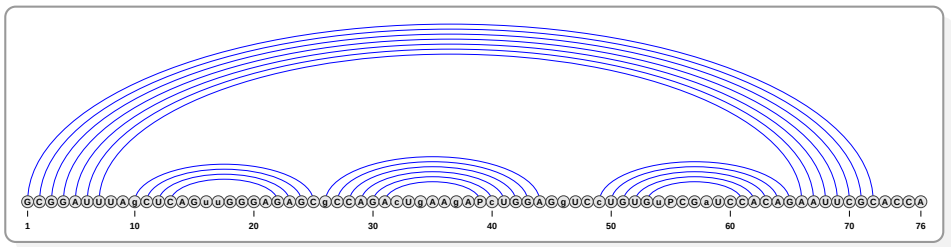


Graphe planaire (outer planar)

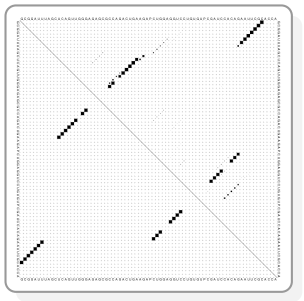
Expression bien parenthésée



Mountain view



Linéaire



Dot plot

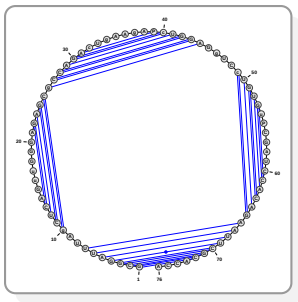


Diagramme de Feynman

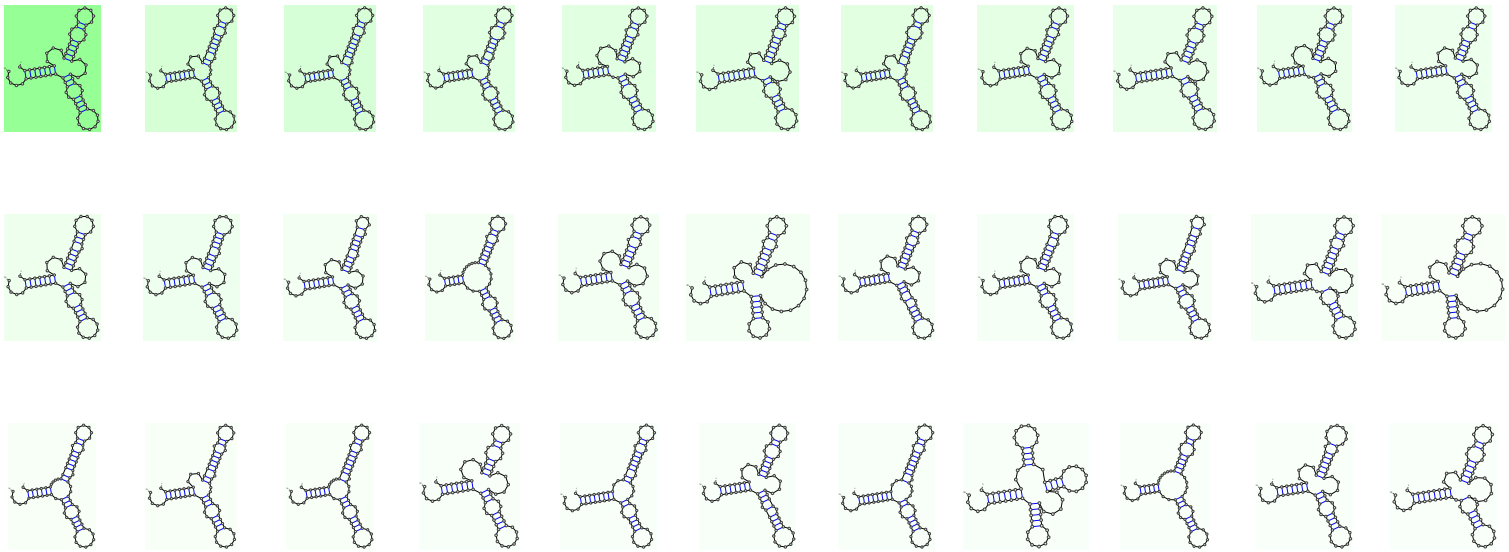
Représentation différentes **mais**
Structure combinatoire commune

Objectif du repliement

Constat : Données génomiques (séquences) beaucoup plus faciles à obtenir expérimentalement que données structurelles (Cristallographie,. . .).
⇒ Intérêt à prédire la structure *in silico* (repliement).

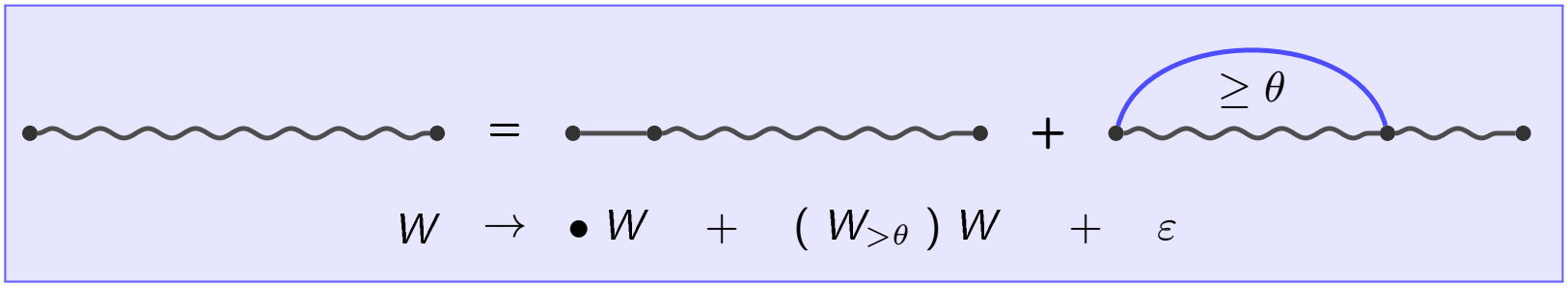
Mais une séquence → Nombreuses ($\approx 1.8^n$ [ZS84]) str. sec. compatibles.

Exemple : ARN de transfert



Problème : Quelle est la (ou les) structure(s) fonctionnelle(s), ou native(s) ?

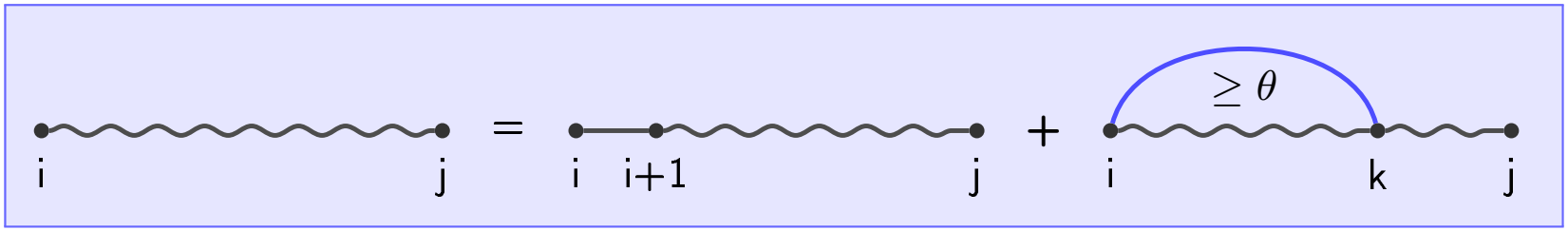
Structures secondaires = Mots bien-parenthèsés + *Contrainte stérique* θ



$\Rightarrow$ Série génératrice [Wat78] $(\theta = 1)$:

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

Structure secondaire = Mots bien-parenthèsés + *Contrainte stérique* θ



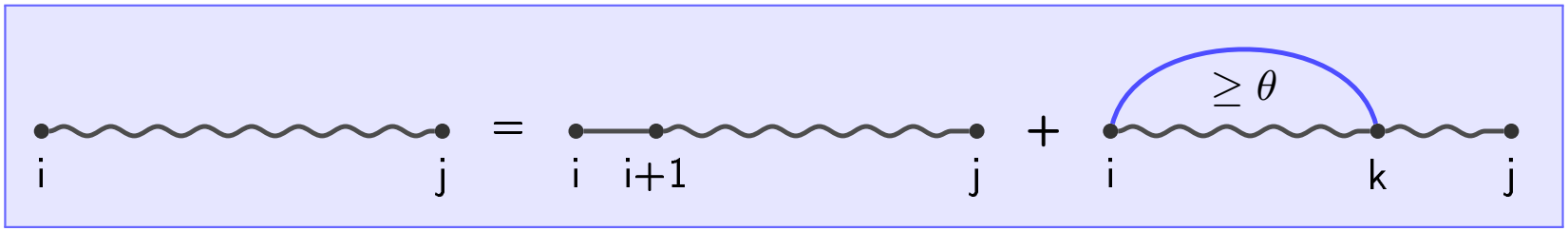
$\Rightarrow$ Comptage des str. sec. compatibles avec un ARN ω :

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

$$N_{i,j} = \sum \begin{cases} N_{i+1,j} & i \text{ non apparié} \\ \sum_{k=i+\theta+1}^{j} \psi_{i,k} \cdot N_{i+1,k-1} \cdot N_{k+1,j} & i \text{ apparié à } k \end{cases}$$

où $\psi_{i,k} = 1$ ssi appariement possible des bases ω_i/ω_k, et 0 sinon.

Structure secondaire = Mots bien-parenthèsés + *Contrainte stérique* θ



Hypothèse [NJ80] : Parmi toutes les struc. sec., le repliement *fonctionnel* d'un ARN maximise le #paires de bases.

Décomposition engendre toutes les str. sec. compatibles avec ω
+ *Sous-structure optimale* $\Rightarrow$ Changement d'algèbre $(+, \times) \to (\max, +)$:

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

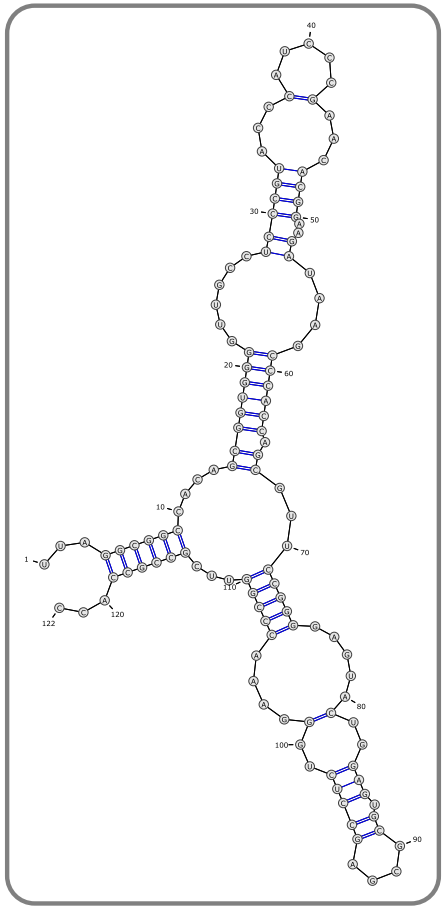
$$N_{i,j} = \max \begin{cases} N_{i+1,j} & i \text{ non apparié} \\ \max\limits_{k=i+\theta+1}^{j} \psi_{i,k} + N_{i+1,k-1} + N_{k+1,j} & i \text{ apparié à } k \end{cases}$$

Algorithme en $\Theta(n^3)$, 60% des paires prédites sont valides.

Modèle d'énergie libre de Turner

Énergie libre additive sur celle des éléments d'une décomposition non-ambiguë en boucles de la structure 2^{aire} :

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

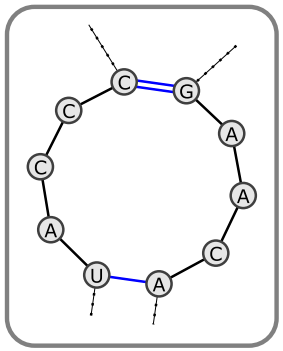


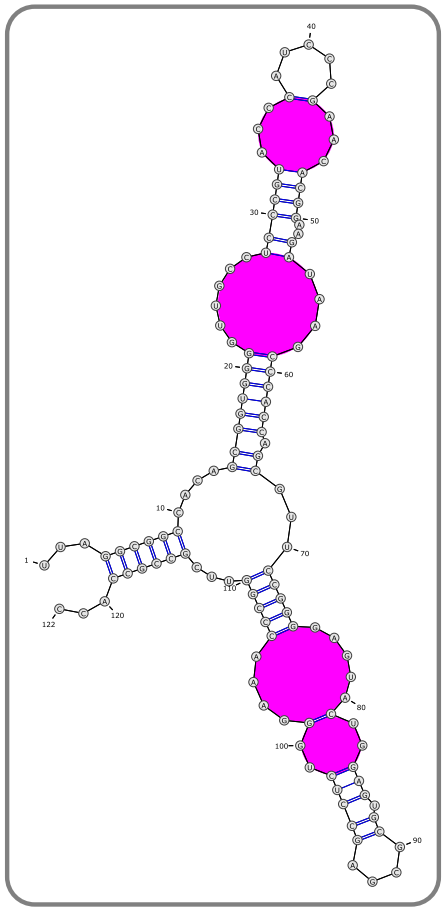
Énergies libres déterminées expérimentalement
+ Interpolation pour les grandes boucles
Énergie libre faible → repliement stable
⇒ Minimiser l'énergie libre.

Modèle d'énergie libre de Turner

Énergie libre additive sur celle des éléments d'une décomposition non-ambiguë en boucles de la structure 2^{aire} :

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



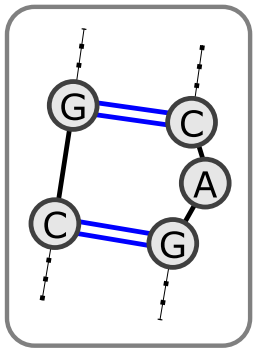


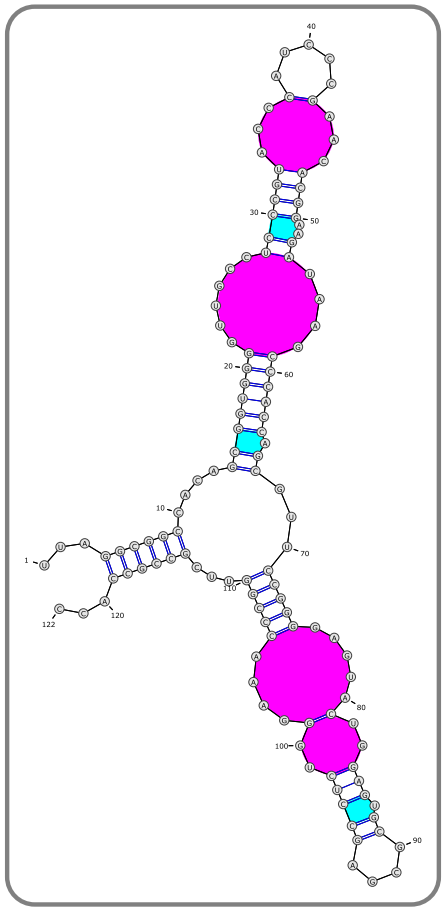
Énergies libres déterminées expérimentalement
+ Interpolation pour les grandes boucles
Énergie libre faible → repliement stable
⇒ Minimiser l'énergie libre.

Modèle d'énergie libre de Turner

Énergie libre additive sur celle des éléments d'une décomposition non-ambiguë en boucles de la structure 2^{aire} :

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



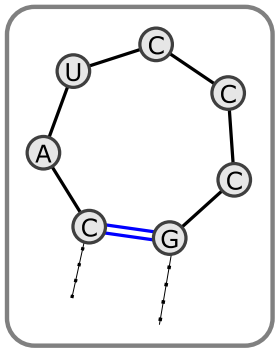


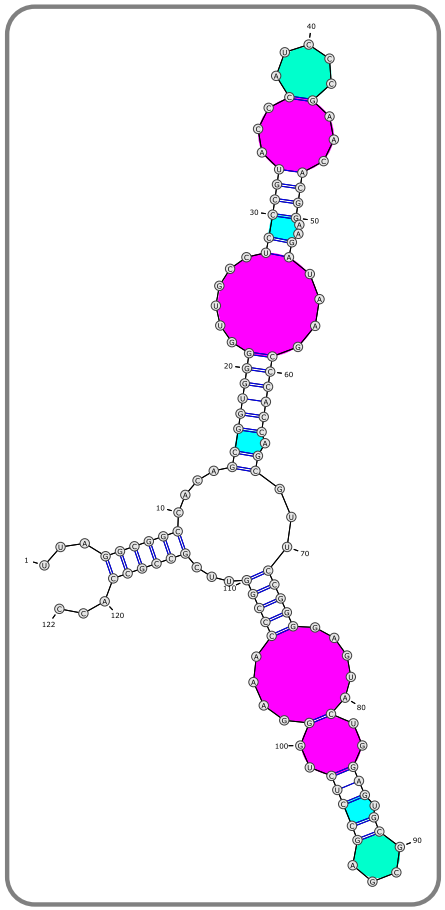
Énergies libres déterminées expérimentalement
+ Interpolation pour les grandes boucles
Énergie libre faible → repliement stable
⇒ Minimiser l'énergie libre.

Modèle d'énergie libre de Turner

Énergie libre additive sur celle des éléments d'une décomposition non-ambiguë en boucles de la structure 2^{aire} :

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



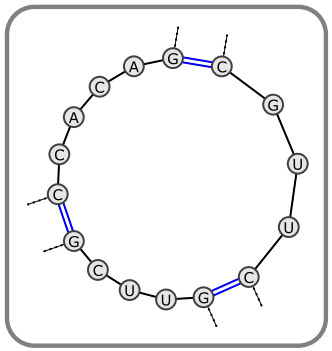


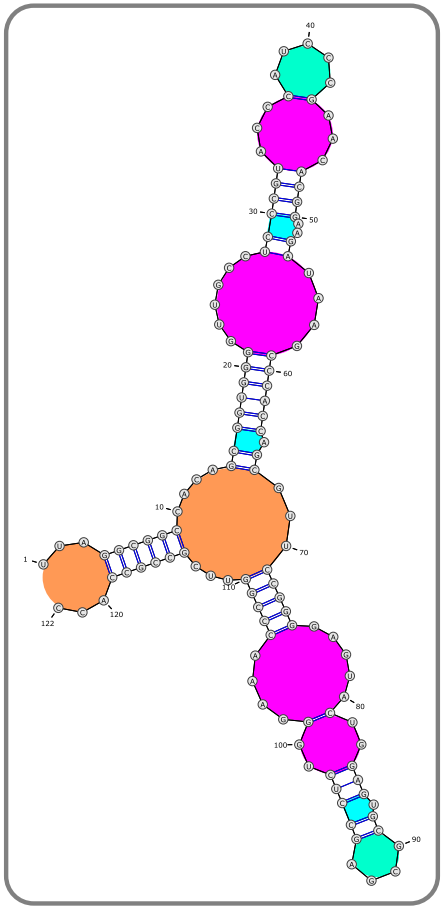
Énergies libres déterminées expérimentalement
+ Interpolation pour les grandes boucles
Énergie libre faible → repliement stable
⇒ Minimiser l'énergie libre.

Modèle d'énergie libre de Turner

Énergie libre additive sur celle des éléments d'une décomposition non-ambiguë en boucles de la structure 2^{aire} :

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



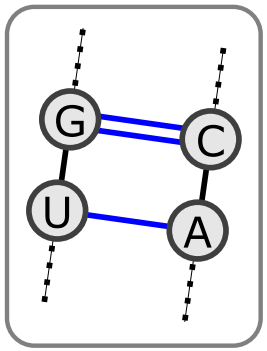


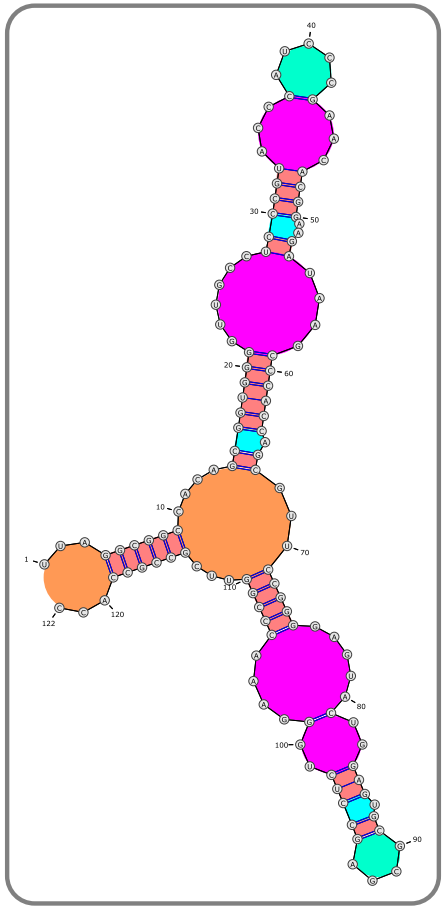
Énergies libres déterminées expérimentalement
+ Interpolation pour les grandes boucles
Énergie libre faible → repliement stable
⇒ Minimiser l'énergie libre.

Modèle d'énergie libre de Turner

Énergie libre additive sur celle des éléments d'une décomposition non-ambiguë en boucles de la structure 2^{aire} :

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

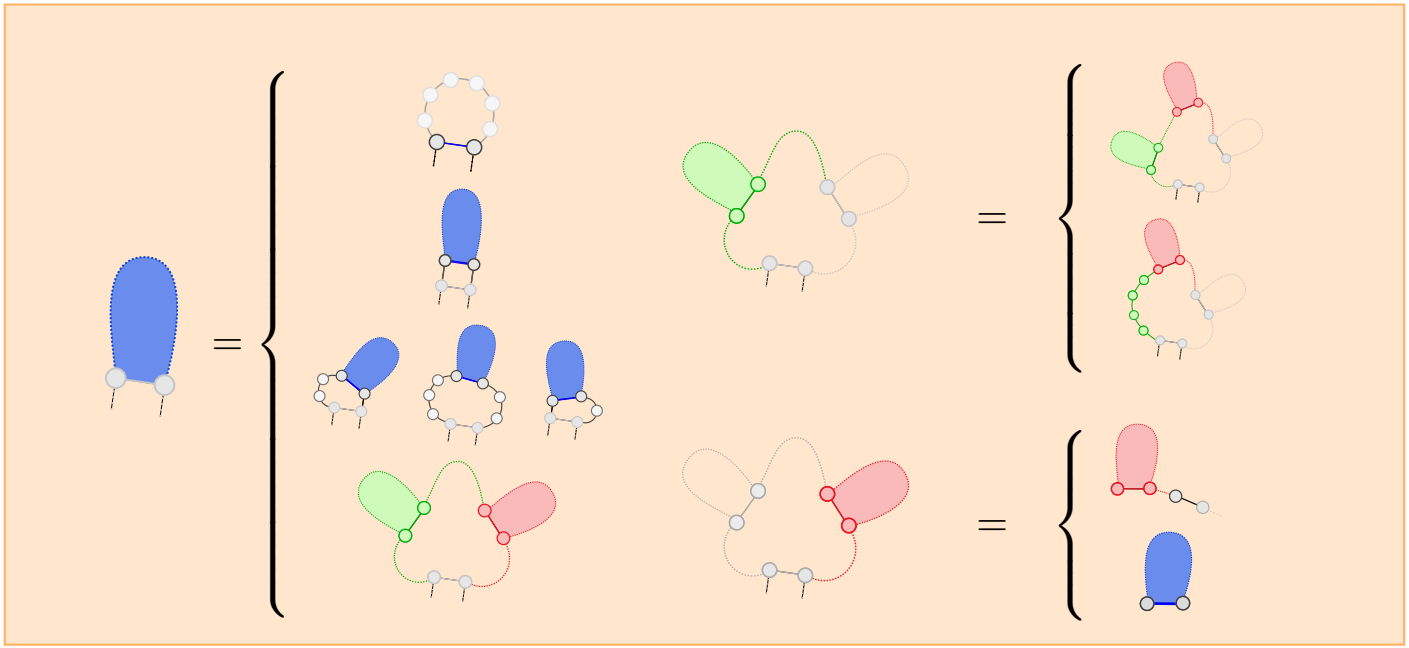




Énergies libres déterminées expérimentalement
+ Interpolation pour les grandes boucles
Énergie libre faible → repliement stable
⇒ Minimiser l'énergie libre.

Hypothèse [ZS81] : Le repliement *fonctionnel* d'un ARN minimise l'énergie libre.

Grammaire proposée par Zuker *et al* :



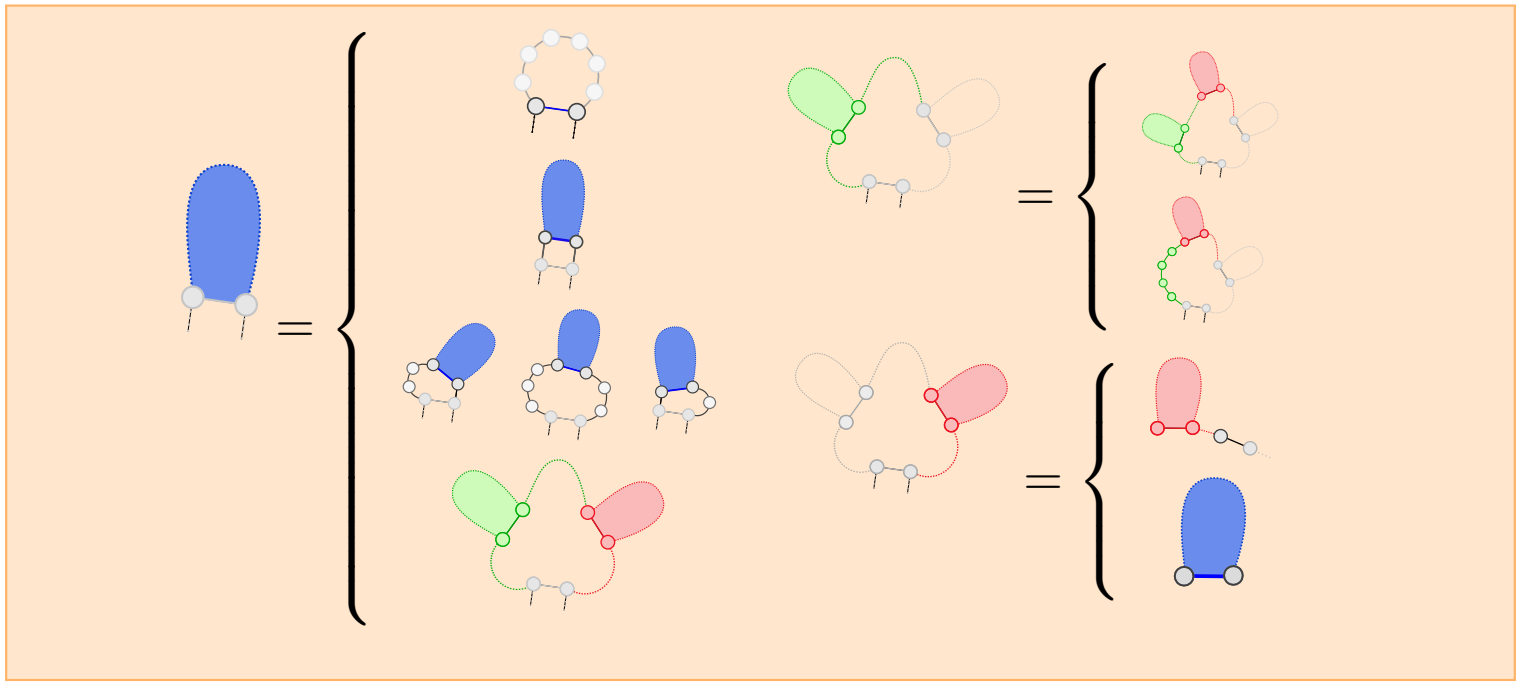
Grammaire non-ambiguë : OK !

Complète ? Comparons les séries ordinaires . . .

Hypothèse [ZS81] : Le repliement *fonctionnel* d'un ARN minimise l'énergie libre.

Rappel : Série génératrice des structures secondaires [Wat78]

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



Hypothèse [ZS81] : Le repliement *fonctionnel* d'un ARN minimise l'énergie libre.

Rappel : Série génératrice des structures secondaires [Wat78]

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

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

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

$$\begin{aligned} \mathcal{A}(z) &= \frac{1 - z - z^2 - \sqrt{1 - 2z - z^2 - 2z^3 + z^4}}{2z^2} \\ &= \mathcal{W}(z) - 1 \quad \text{(Oubli de la str. sec. de longueur 0)} \end{aligned}$$

$\Rightarrow$ Schéma complet, correct par construction $\Rightarrow$ Minimisation accessible.

Hypothèse [ZS81] : Le repliement *fonctionnel* d'un ARN minimise l'énergie libre.

Méthode : Grammaire $\rightarrow$ Récurrences de comptage
+ Changement d'algèbre $(+, \times) \rightarrow (\min, +)$

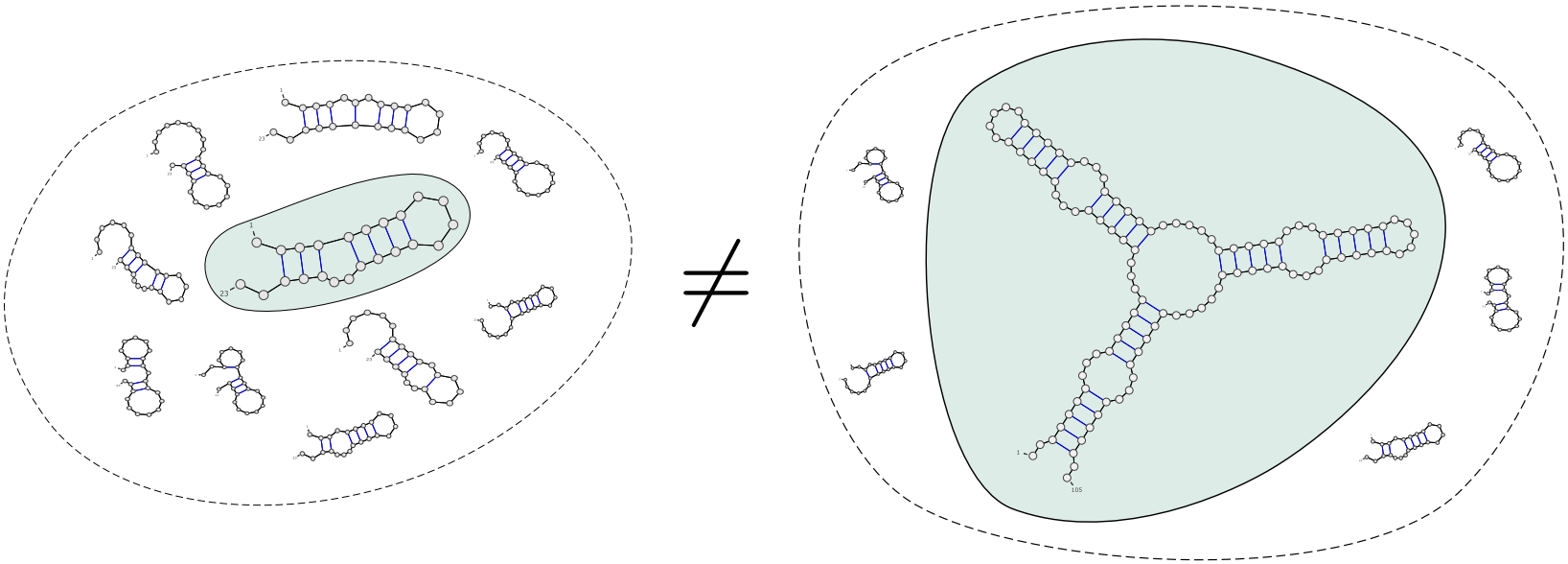
Algorithme MFold [ZS81] :

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

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

Prédit 73% des paires de bases correctement.
Méthode publiée puis étendue dans 5 articles cités $\sim$10 000 fois !

Constat : Repliement d'énergie libre minimale n'est pas toujours prédominant.
Pire, le repliement optimal est parfois faiblement représentatif de l'ensemble.



Idée [McC90] : Supposer une distribution de Boltzmann sur l'ensemble $\mathcal{S}_\omega$ des structures compatibles avec une séquence d'ARN ω.

$$\text{Str. sec. } S, \text{ énergie libre } E_{S,\omega} \quad \rightarrow \quad \text{Facteur de Boltzmann } \mathcal{B}_{S,\omega} = e^{\frac{-E_{S,\omega}}{RT}}$$

$$\text{Probabilité de Boltzmann } P_{S,\omega} \quad = \quad \frac{\mathcal{B}_{S,\omega}}{\mathcal{Z}_\omega}$$

où $\mathcal{Z}_\omega := \sum_{S \in \mathcal{S}_\omega} e^{\frac{-E_{S,\omega}}{RT}}$ est la fonction de partition d'un ARN ω.

Fonction de partition = Comptage pondéré des structures compatibles.

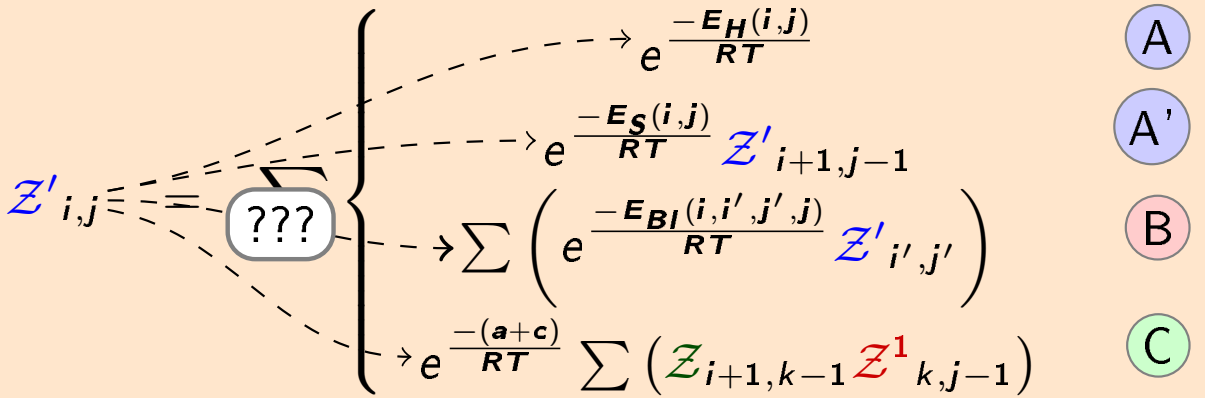
Principe : Pondérer chaque boucle d'énergie libre Δ par son facteur de Boltzmann *local* $e^{\frac{-\Delta}{RT}}$ (Grammaires pondérées [DRT00]).

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

Bonus : Extraction des probabilités d'appariement pour tout couple de bases.

Quelle structure(s) représentante(s) pour l'ensemble de Boltzmann ?

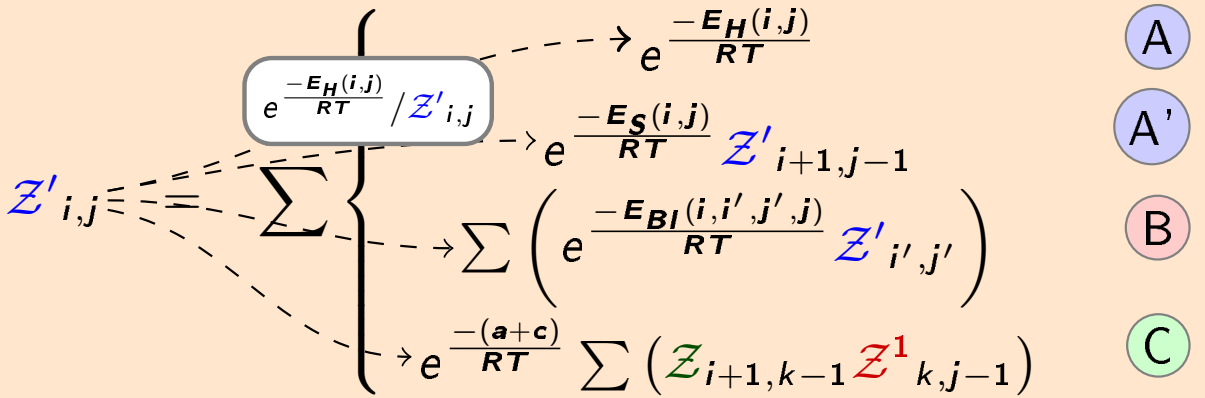
Algorithme SFold [DL03]
Précalcul : Calculer les matrices des fonctions de partition ($\mathcal{Z}, \mathcal{Z}', \mathcal{Z}^1$).
Remontée stochastique : Choisir un des cas proportionnellement à sa contribution, et réitérer récursivement.



$\Leftrightarrow$ Méthode récursive [Wil77] pour la génération aléatoire pondérée [DRT00].
$\Rightarrow$ Optimisations : Boustrophedon [FZV94], génération non-redondante ...

Ding *et al* [DCL05] engendrent puis calculent une structure consensus.
$\Rightarrow$ Même sensibilité que `MFold` mais gain de 30% sur la spécificité.

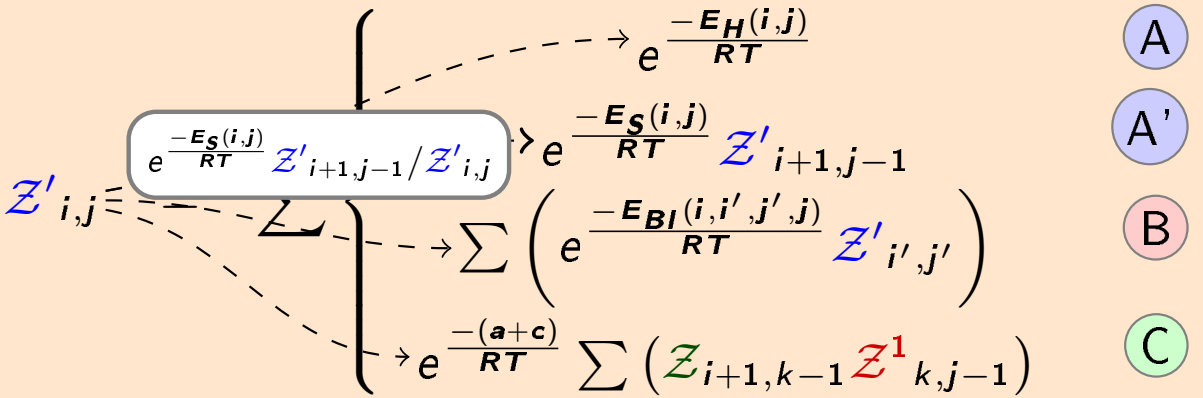
Algorithme SFold [DL03]
Précalcul : Calculer les matrices des fonctions de partition ($\mathcal{Z}, \mathcal{Z}', \mathcal{Z}^1$).
Remontée stochastique : Choisir un des cas proportionnellement à sa contribution, et réitérer récursivement.



$\Leftrightarrow$ Méthode récursive [Wil77] pour la génération aléatoire pondérée [DRT00].
$\Rightarrow$ Optimisations : Boustrophedon [FZV94], génération non-redondante ...

Ding *et al* [DCL05] engendrent puis calculent une structure consensus.
$\Rightarrow$ Même sensibilité que `MFold` mais gain de 30% sur la spécificité.

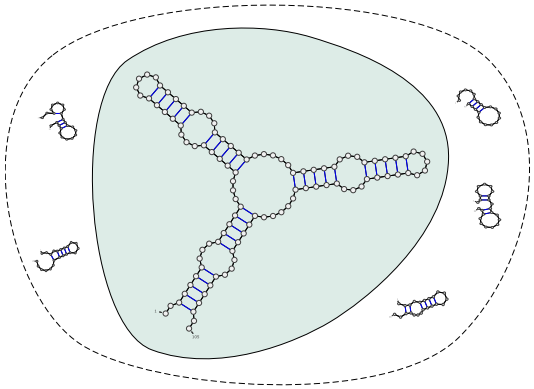
Algorithme SFold [DL03]
Précalcul : Calculer les matrices des fonctions de partition ($\mathcal{Z}, \mathcal{Z}', \mathcal{Z}^1$).
Remontée stochastique : Choisir un des cas proportionnellement à sa contribution, et réitérer récursivement.



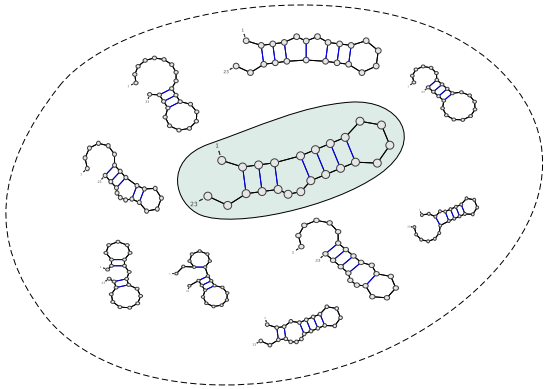
$\Leftrightarrow$ Méthode récursive [Wil77] pour la génération aléatoire pondérée [DRT00].
$\Rightarrow$ Optimisations : Boustrophedon [FZV94], génération non-redondante ...

Ding *et al* [DCL05] engendrent puis calculent une structure consensus.
$\Rightarrow$ Même sensibilité que `MFold` mais gain de 30% sur la spécificité.

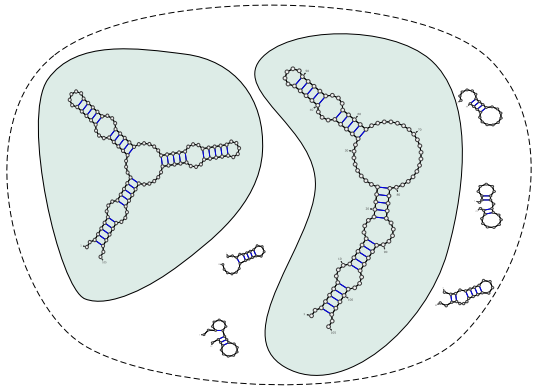
Statistique de la distribution de Boltzmann



Repliement fonctionnel ?



Repliement mal défini : ARNm ?



ARN bi-stable ou phénomène cinétique ?

Constat : Plusieurs scénarios possibles pour l'ensemble de Boltzmann.
$\Rightarrow$ Extraire des indicateurs numériques discriminant les scénarios.

Idée : Marquer occurrences de boucles et extraire les espérances, variances ...

Problème : Contraintes d'appariements induisent un *manque de symétrie*.
Série génératrices et dérivées partielles difficilement accessibles ...

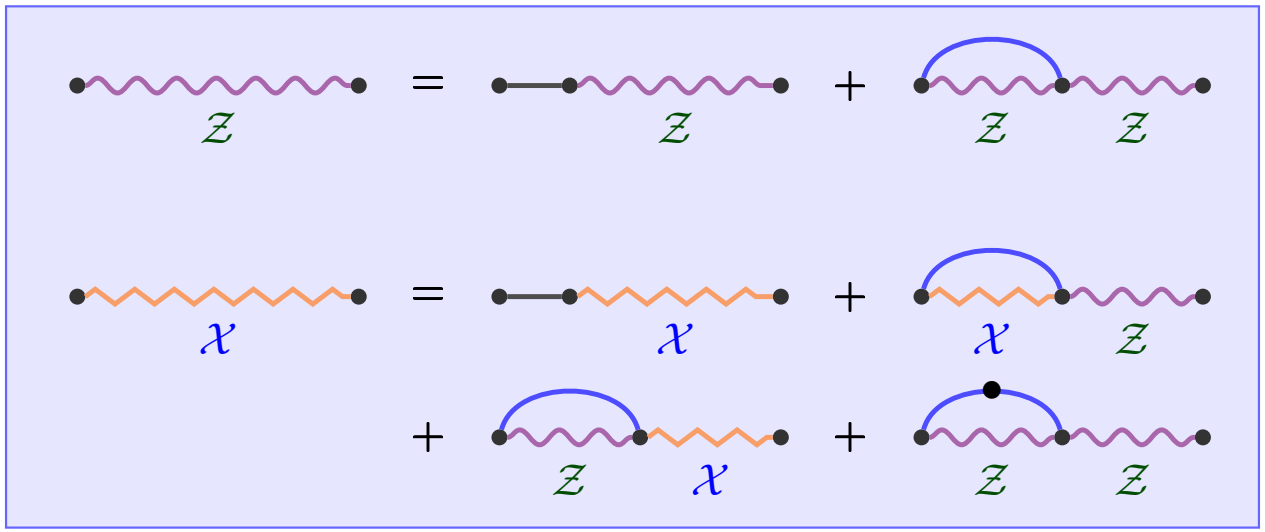
$\Rightarrow$ Utiliser un (ou plusieurs) pointage(s) partiel(s) [DFLS04] de la grammaire.

Objectif : Espérance E_ω du nombre de paires de bases $bp(\cdot)$.

$$E_\omega = \sum_{s \in \mathcal{S}_\omega} bp(s) \cdot p_s = \frac{\sum_{s \in \mathcal{S}_\omega} bp(s) \cdot e^{\frac{-E_{\omega,s}}{RT}}}{\mathcal{Z}} = \frac{\mathcal{X}_\omega}{\mathcal{Z}_\omega}.$$

où $\mathcal{Z}_\omega$ est la fonction de partition et $\mathcal{X}_\omega$ le coeff. d'une dérivée partielle.

Dérivée connaît un analogue au niveau des grammaires hors-contexte [DFLS04].

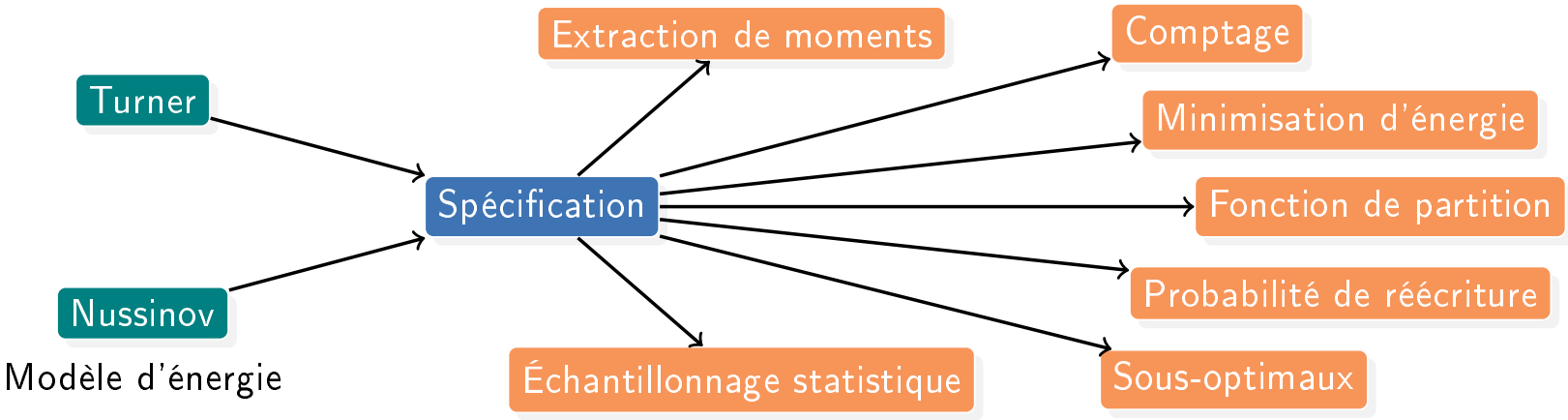


$\Rightarrow$ Calcul en $\Theta(n^3)$ de $\mathcal{X}_\omega$ puis E_ω est à portée de main.

(Grammaire $\rightarrow$ Réc. comptage $\rightarrow$ Contraintes d'appariements)

Remarque : Généralise un algorithme similaire pour l'énergie libre [MMN05].

Conclusion/perspectives



Limites liées à l'absence de symétrie (appariements) $\rightarrow$ Difficile de :

- Générer par Boltzmann [DFLS04] dans l'ensemble de Boltzmann.
- Récurrences linéaires à coeff. polynômiaux pour la fonction de partition.
- Supposer des distributions normales (Grammaire fortement connexe)

Perspectives :

- Quel sous-ensemble de structures domine l'ensemble de Boltzmann ?
  $\sim$ Collisions lors de l'échantillonnage (En cours avec D. Gardy)
- Étendre les espaces de conformations : Pseudonoeuds (cf C. Saule)
- Design, i.e. génération d'une séquence ayant un repliement donné.

References I

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

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

P. Duchon, P. Flajolet, G. Louchard, and G. Schaeffer.

Boltzmann samplers for the random generation of combinatorial structures.
Combinatorics, Probablity, and Computing, 13(4–5):577–625, 2004.
Special issue on Analysis of Algorithms.

Y. Ding and E. Lawrence.

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

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

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

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

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

J.S. McCaskill.

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

István Miklós, Irmtraud M Meyer, and Borbála Nagy.

Moments of the boltzmann distribution for rna secondary structures.
Bull Math Biol, 67(5):1031–1047, Sep 2005.

References II

R. Nussinov and A.B. Jacobson.

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

M. S. Waterman.

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

H. S. Wilf.

A unified setting for sequencing, ranking, and selection algorithms for combinatorial objects.
Advances in Mathematics, 24:281–291, 1977.

M. Zuker and P. Stiegler.

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

M. Zuker and D. Sankoff.

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