

Structure secondaire de l'ARN - BIF7101

Mohamed Amine Remita
Abdoulaye Baniré Diallo *Ph.D.*

17 février 2016

1 Introduction

- Structure de l'ARN
- Les types d'ARN
- ARN et bioinformatique

2 Repliement par minimisation d'énergie

- Le problème
- Un critère de choix : l'énergie
- Technique : la programmation dynamique
- MFOLD
- Énergie et probabilités : Vienna

3 Analyse de covariation de séquences

4 Atelier

La structure primaire

L'ARN

Les mots sur $\{A, C, G, U\}$

GUCCUCAUAGCUUACAAACCUCAAAGCGCGGCACUG
AAGAUGCCAAGACGGUAACCACCAUACCUGAGGACA

Différence ADN et ARN

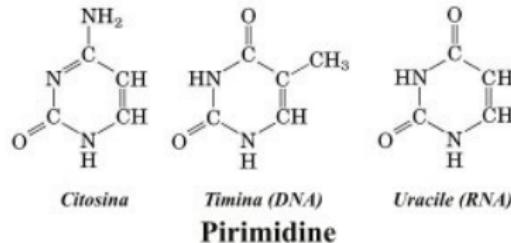
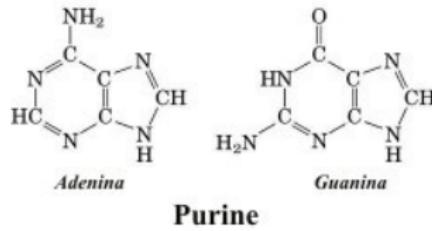
Uracile : pyrimidine \simeq Thymine

Le sucre des nucléotides = ribose au lieu de désoxyribose dans l'ADN

ARN : simple brin \Rightarrow plus de souplesse dans les structures 2D et 3D

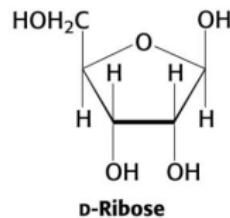
La structure primaire

ARN

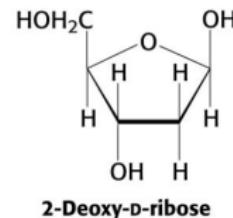


La structure primaire

ARN



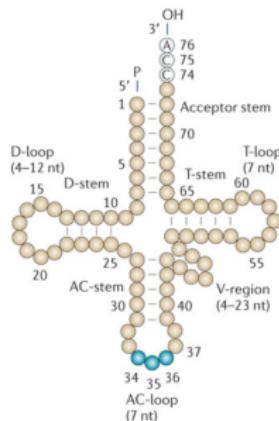
D-Ribose



2-Deoxy-D-ribose

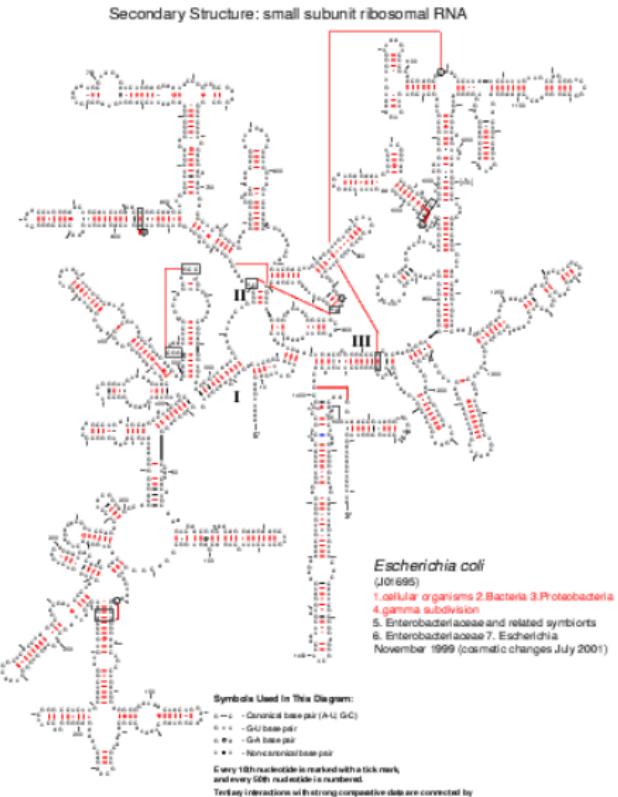
La structure secondaire

Repliement 2D par création de liens entre paires de bases



A – U } Watson-Crick
C – G }
G – U } Wobble
Pas de croisements

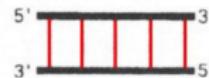
Éléments de structure secondaire



Éléments de structure secondaire



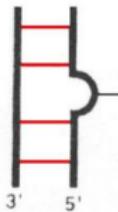
single strand



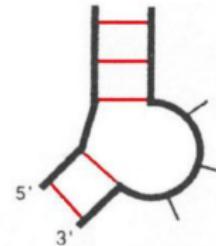
A-form double helix



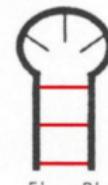
Double helix with
5'-dangling end



single nucleotide bulge

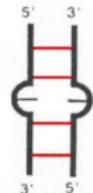


three nucleotide bulge



hairpin loop

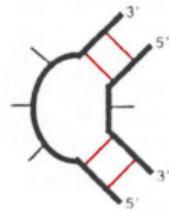
Éléments de structure secondaire



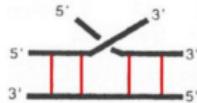
mismatch pair or,
symmetric internal loop
of 2 nucleotides



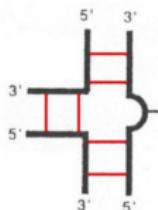
symmetric internal loop



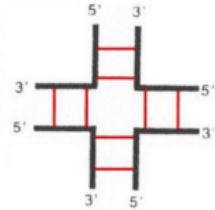
asymmetric internal loop



two-stem junction
or, coaxial stack



three-stem junction



four-stem junction

La structure tertiaire

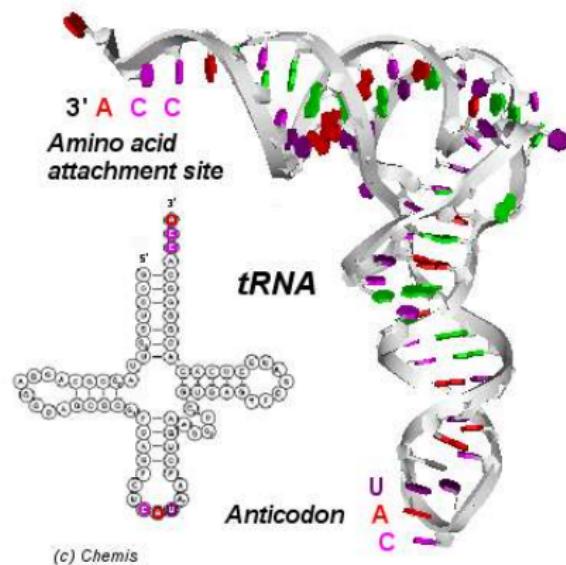
Repliement dans l'espace (3D) de la structure secondaire

Remarque :

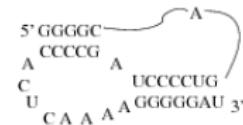
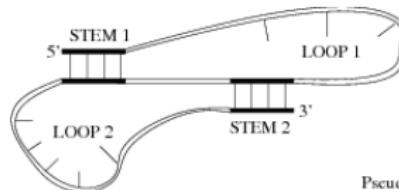
- Les ARN non traduits (ARNt, ARNr, ...) ont une structure « fixe » pour chaque famille
- Les ARNm ont une structure très variable
- La structure est très fortement liée à la fonction

En général, les méthodes ne tiennent pas compte de la structure tertiaire au départ

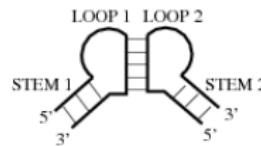
Structure secondaire vs. Structure tertiaire



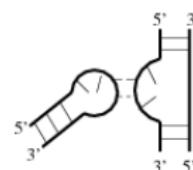
Éléments de structure tertiaire



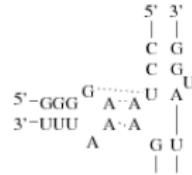
Pseudoknot



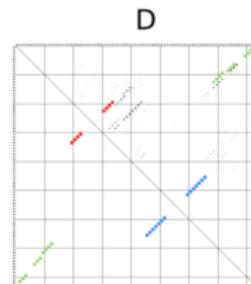
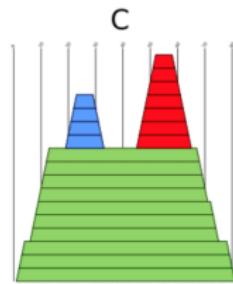
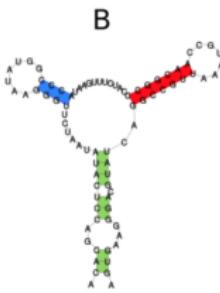
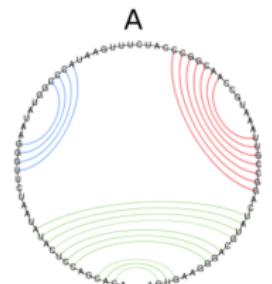
Kissing hairpins



Hairpin loop - bulge contact



Représentations de la structure secondaire



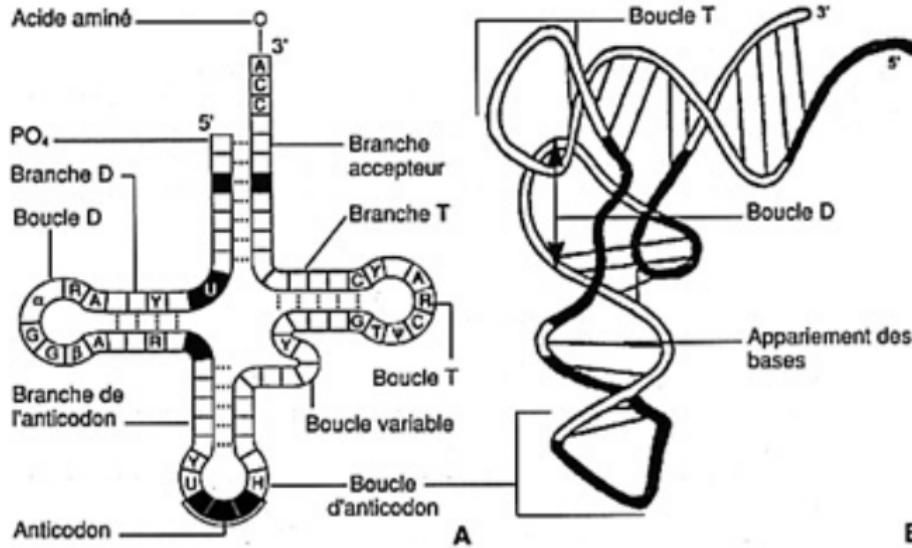
E
ACACGACCUCUAUAAAUCUUGGGAAUAUGGCCAUAGUUUCUACCGGGCAACCGUAAAUGCACUAUGCAGGGAAUGGA
.....(((((.....(((((.....))))))).....((((((.....))))))).....)))).....)))).....)))).....

Les types d'ARN

- ARN de transfert
- ARN ribosomal
- ARN messager
- petits ARN nucléolaires (snoARN)
- piARN, microARN

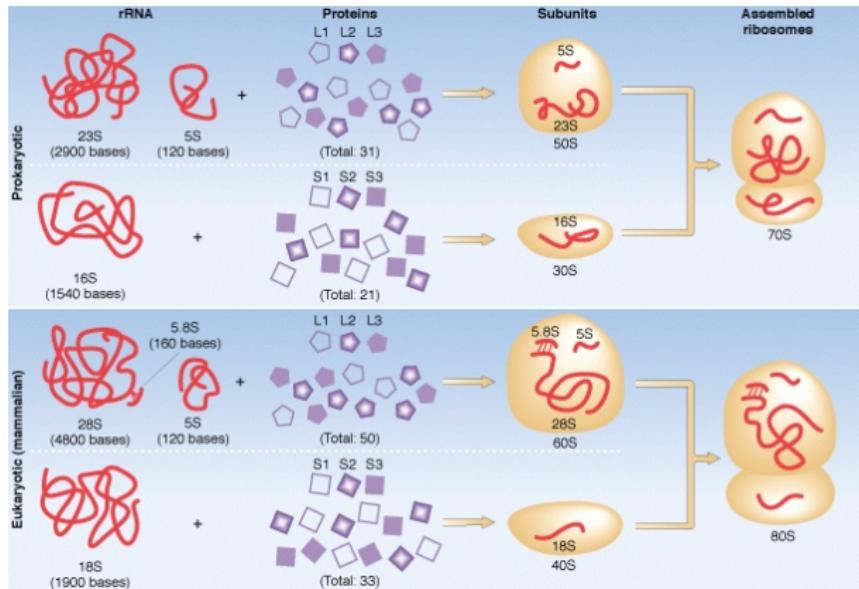
ARNt

Motifs d'un ARNt



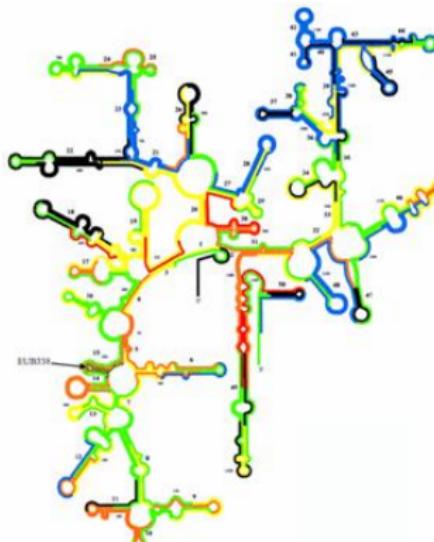
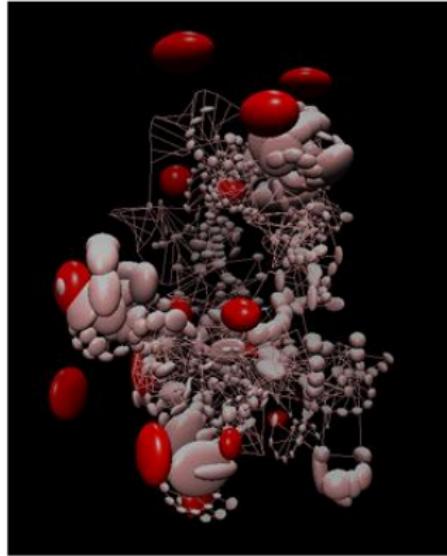
ARNr

Composition des sous unités



ARNr

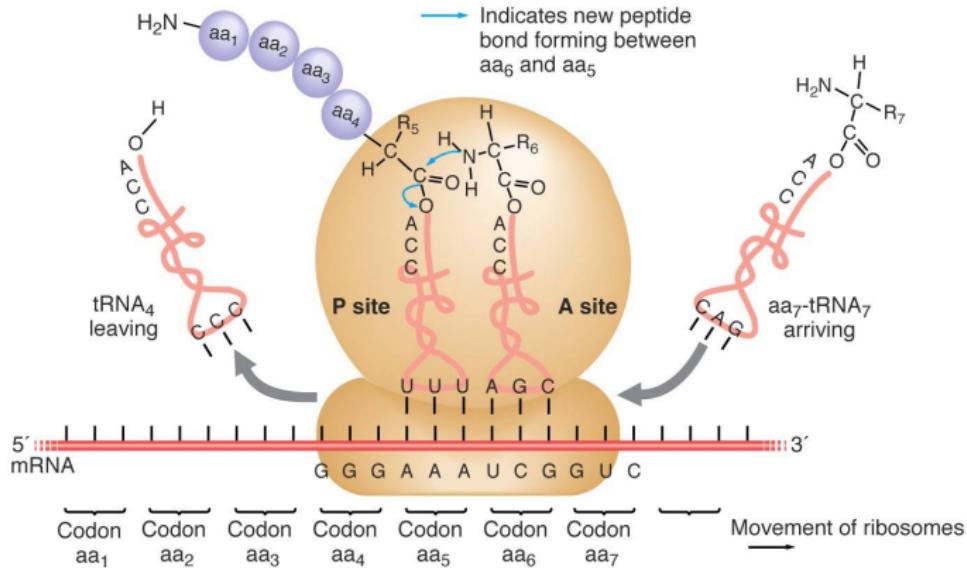
Structure tertiaire et secondaire



16S rRNA

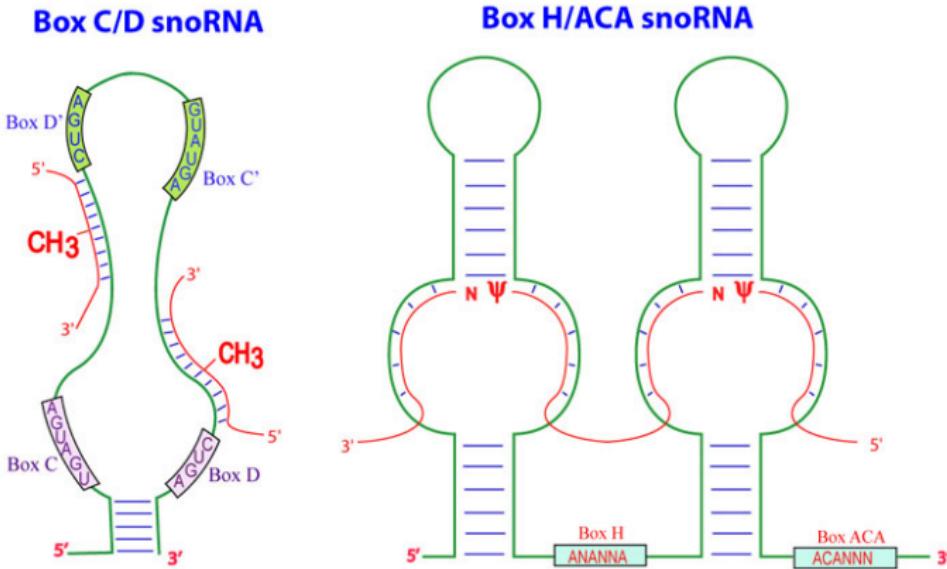
ARNm

Traduction de l'ARNm



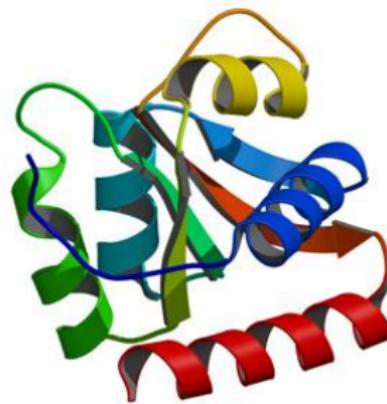
petits ARN nucléolaires (snoARN)

Structure et motifs communs



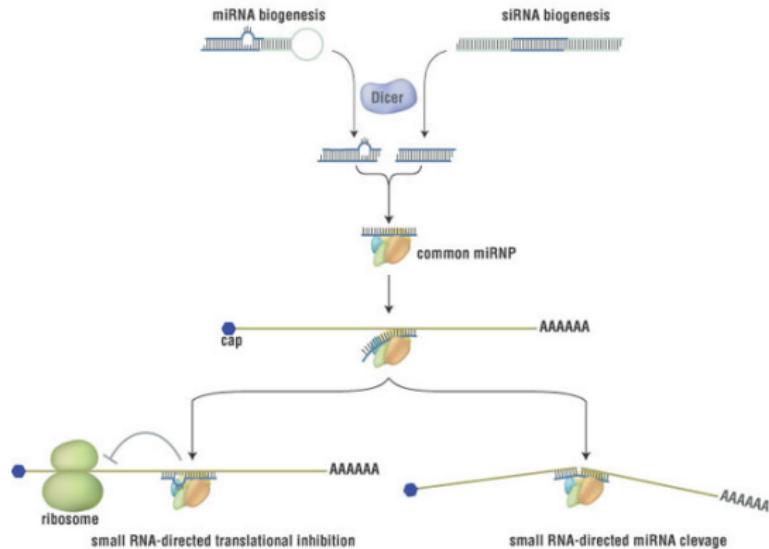
snoARN

Crystal structure of the L7 snoRNA-binding protein from *Methanococcus jannaschii*



microARN et piARN

Mécanisme d'action des miRNAs et siRNAs



ARN et bioinformatique

Manipulation de structures

bases de données, sites web, format de fichiers, visualisation

Prédiction / Repliement

Structure primaire d'ARN

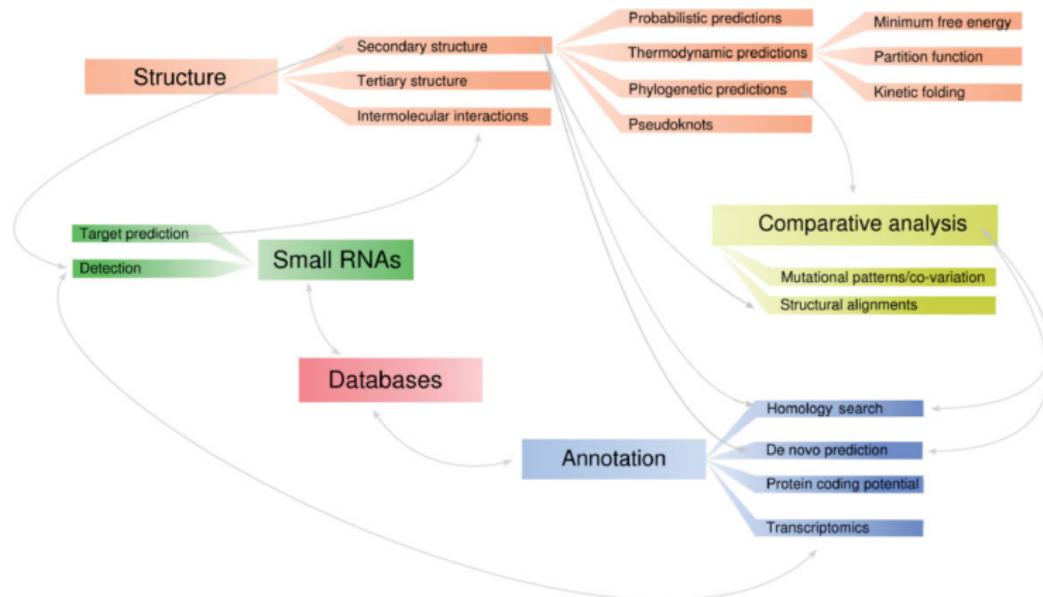
↓ prédiction

Structure secondaire

Détection d'ARN

A partir d'une séquence d'ARN et d'un schéma de structure secondaire, retrouver dans l'ARN les sous-séquences pouvant se replier selon le schéma donné

ARN et bioinformatique



Le problème

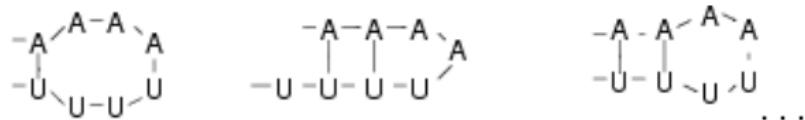
Définition

$S = \text{séquence d'ARN}$

prédire sa structure secondaire

Il y a un nombre exponentiel de repliements possibles

Exemple : AAAAUUUU



Lesquels sont « censés » ?

Un critère de choix : l'énergie

Une paire de bases renforce la stabilité de la structure : il faut beaucoup d'énergie pour la casser

Une boucle déstabilise la structure

Une structure stable demande beaucoup d'énergie pour être modifiée
⇒ on cherche à replier en une structure stable du point de vue énergétique

Plus précisément

Une paire de bases « cachée » dans un gros groupe de paires de base est protégée et dure à casser

Une paire de bases adjacente à une boucle est plus facile à casser

Une paire de bases adjacente à une grosse boucle l'est encore plus

Technique : la programmation dynamique

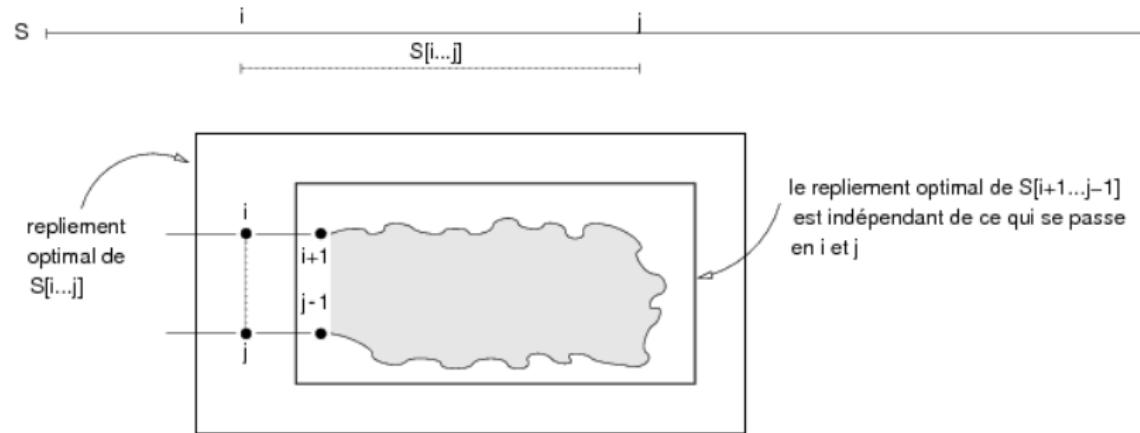
Premier critère :

on essaie de maximiser le nombre de paires de bases en tenant compte du fait que les liaisons A-U et G-C sont très stables, les liaisons G-U sont stables et les autres ne sont pas stables.

Avant de passer au calcul :

il faut travailler encore sur la partie « modèle », à savoir déterminer un critère de stabilité.

Principe de la programmation dynamique



Algorithme de Nussinov

Principe :

calculer le repliement qui maximise le nombre de paires de bases
(approximation du maximum d'énergie)

Programmation dynamique :

1. Calcul d'un tableau W : $W_{i,j}$ = nombre maximal de paires de bases parmi tous les repliements possibles du segment $S[i..j]$
 $\Rightarrow W_{1,n}$ = nombre de paires de bases d'une structure optimale
2. Calcul d'un chemin dans W pour en déduire une structure optimale.

Algorithme de Nussinov

Calcul de W :

programmation dynamique

Cas de base :

L = taille minimale d'une boucle

$W_{i,j} = 0$ si $j \leq i + L$

Récursion :

4 cas pour le calcul de $W_{i,j}$. On suppose $W_{k,l}$ connu pour

$$\begin{cases} k = i, & l < j \\ k > i, & l = j \\ k > i, & l < j \end{cases}$$

Algorithme de Nussinov

a). i et j forment une paire de bases :

$$W_{i,j} = 1 + W_{i+1,j-1}$$

b). i et j ne sont dans aucune paire de bases :

$$W_{i,j} = W_{i+1,j-1}$$

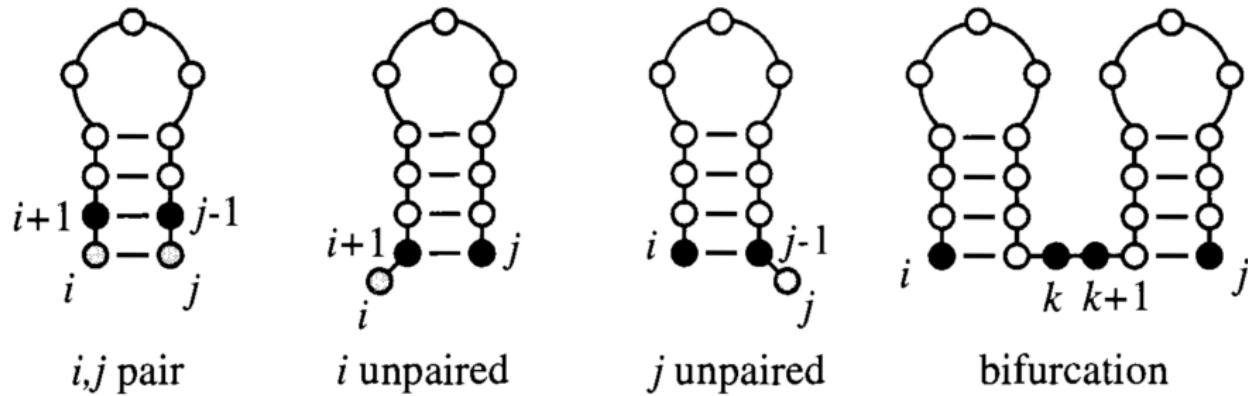
c). i (resp. j) est dans une paire de bases mais pas j (resp. i) :

$$W_{i,j} = W_{i,j-1} \text{ (resp. } W_{i,j} = W_{i+1,j})$$

d). i et j sont dans deux paires de bases :

$$W_{i,j} = \max_{k \in [i+1, j-1]} \{W_{i,k} + W_{k+1,j}\}$$

Algorithme de Nussinov



Algorithme de Nussinov

Algorithme de Nussinov et al. (repliement d'ARN)

1. Calcul de la matrice W

Pour j de 1 à n faire

 Pour i de 1 à n-j+1 faire

 Si ($j \leq s$ i+L) alors // L longueur minimale d'une boucle

 W[i,j] := 0;

 Sinon

 w := W[i,i+1] + W[i+2,j];

 Pour k de i+2 à j faire

 Si W[i,k] + W[k+1,j] > w alors

 w := W[i,k] + W[k+1,j];

 W[i,j] := MAX{

 W[i+1,j],

 W[i,j-1],

$\delta(i,j) + W[i+1,j-1]$,

 w

}

$\delta(i,j) = 1$ si $W[i]$ et $W[j]$ peuvent former une paire de bases, 0 sinon

Algorithme de Nussinov

Exemple de remplissage de la matrice

(a)

	$j \rightarrow$									
		G	G	G	A	A	A	U	C	C
$i \downarrow$	G	0								
	G	0	0							
A	A	0	0							
	A	0	0							
U	U				0	0				
	C				0	0				
C	C				0	0				

(b)

	$j \rightarrow$									
		G	G	G	A	A	A	U	C	C
$i \downarrow$	G	0	0							
	G	0	0	0						
A	A	0	0	0						
	A	0	0	0						
U	U				0	0				
	C				0	0				
C	C				0	0				

(c)

	$j \rightarrow$									
		G	G	G	A	A	A	U	C	C
$i \downarrow$	G	0	0	0	0					
	G	0	0	0	0	0				
A	A	0	0	0	0	0				
	A	0	0	0	0	0	①			
U	U				0	0	1	1		
	C				0	0	0	0		
C	C				0	0	0	0		

(d)

	$j \rightarrow$									
		G	G	G	A	A	A	U	C	C
$i \downarrow$	G	0	0	0	0	0	1	2	③	
	G	0	0	0	0	0	0	1	2	3
A	A	0	0	0	0	0	0	1	2	2
	A	0	0	0	0	0	0	1	1	1
U	U				0	0	1	1	1	1
	C				0	0	0	0	0	0
C	C				0	0	0	0	0	0

Algorithme de Nussinov

2. Calcul des paires de bases d'une structure secondaire

```
Soit P une pile vide;  
Empiler (1,n) dans P;  
Tant que P n'est pas vide faire  
    Soit (i,j) le sommet de P;  
    Dépiler (i,j) de P;  
    Si i >= j ne rien faire;  
    Sinon Si W[i,j] = W[i+1,j]  
        Empiler (i+1,j) dans P;  
    Sinon Si W[i,j] = W[i,j-1]  
        Empiler (i,j-1) dans P;  
    Sinon Si W[i+1,j-1] + delta(i,j) = W[i,j]  
        Enregistrer (i,j) comme paire de base de la structure secondaire;  
        Empiler (i+1,j-1) dans P;  
    Sinon  
        Pour k de i+1 à j-1 faire  
            Si W[i,k] + W[k+1,j] = W[i,j] alors  
                Empiler (k+1,j) dans P;  
                Empiler (i,k) dans P;  
  
            Sortir de la boucle Pour;
```

Algorithme de Nussinov

Exemple de retraçage

j →

	G	G	G	A	A	A	U	C	C
	G	0	0	0	0	0	1	2	3
	G	0	0	0	0	0	1	2	3
<i>i</i> ↓	G	0	0	0	0	0	1	2	2
	A	0	0	0	0	0	1	1	1
	A	0	0	0	0	0	1	1	1
	A	0	0	0	0	0	1	1	1
	U	0	0	0	0	0	0	0	0
	C	0	0	0	0	0	0	0	0
	C	0	0	0	0	0	0	0	0

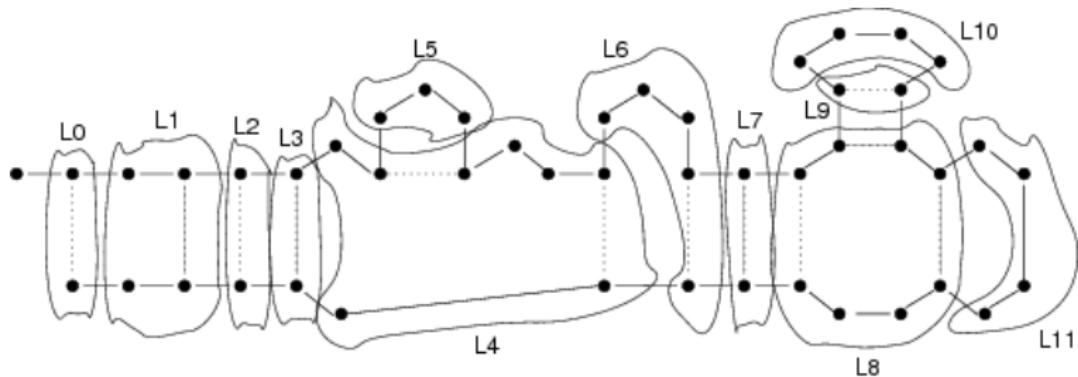
A ↗ A
 A • U
 G • C
 G • C
 G

L'algorithme MFOLD

Exploiter seulement le nombre de paires de bases est insuffisant : une paire de bases isolée stabilise moins que si elle est groupée avec d'autres.

Mais le principe de la programmation dynamique est intéressant
⇒ il faut améliorer le modèle de stabilité énergétique : MFOLD (Zucker)

L'algorithme MFOLD



Idée :

décomposer une structure en éléments moins grossiers que de simples paires de bases et associer une énergie à chacun.

$$G(S) = \sum_{i=0}^{11} G(L_i)$$

Éléments structuraux : Loops

Soit (i, j) une paire de bases

- la base i' (resp. paire de base (i', j')) est accessible depuis (i, j) si $\forall (k, l)$ paire de bases, on n'a pas $i < k < i'$ (resp. $< j' < l < j$)
- une loop est fermée par (i, j) si toutes ses bases et paires de bases sont accessibles

k–loop

une loop à $k - 1$ paire de bases

Hairpin :

1–loop (L_5, L_{10}, L_{11})

BasePairs :

2–loop fermée par (i, j) et avec une seule paire de bases (i', j') : $i' = i + 1$ et $j' = j - 1$ (L_2, L_3, L_7, L_9)

Éléments structuraux : Loops

Bulge :

2-loop $\{(i,j), (i',j')\}$ telle que : $(i' = i + 1, j' < j - 1)$ ou
 $(i' > i + 1, j' = j - 1)$ (L_6)

InteriorLoop :

2-loop $\{(i,j), (i',j')\}$ telle que : $i' > i + 1, j' < j - 1$ (L_1)

MultiLoop :

k -loop, pour $k \geq 3$ (L_4, L_8)

Stem (ou Stack) : suite de BasePairs (L_2L_3)

Éléments structuraux : Loops

Energie d'une hairpin (exemple de calcul)

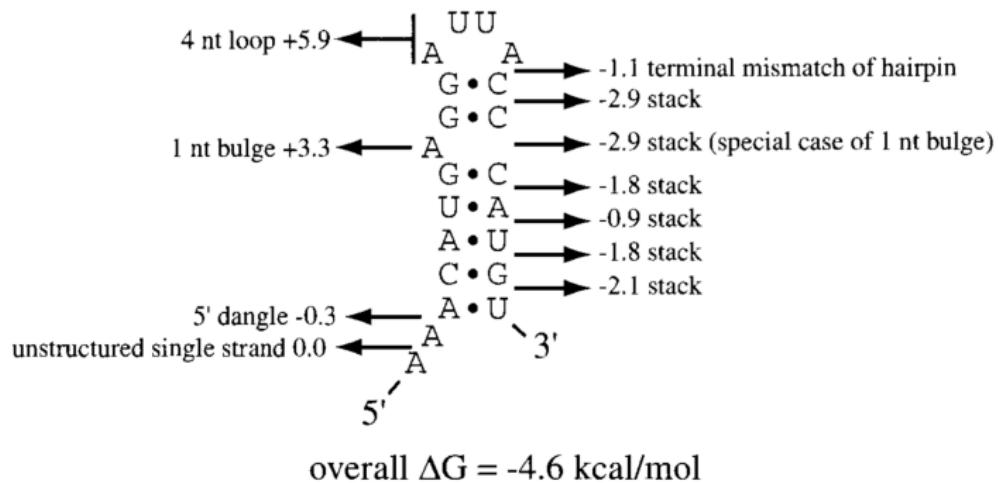
Exemple :



calcul :

① Loop penalty (loop size=4)	+5.60 kcal/mol
② Stacking GC/AU	-2.20 kcal/mol
③ Tetraloop bonus	-3.00 kcal/mol
Total :	+0.40 kcal/mol

Example de calcul de l'énergie d'un stem loop



Énergie

nécessaire à la stabilité de la structure à une température donnée

données stockées dans des fichiers distribués avec le logiciel MFOLD
(modifiables)

autres loops :

- stack : $e_s(i,j)$
- bulge et interior loop : $e_{bi}(i,j,i',j')$

Fichiers

SIZE	INTERNAL	BULGE	HAIRPIN
1	.	3.80	.
2	.	2.80	.
3	.	3.20	5.70
4	1.70	3.60	5.60
5	1.80	4.00	5.60
6	2.00	4.40	5.40
7	2.20	4.60	5.90
8	2.30	4.70	5.60
9	2.40	4.80	6.40
10	2.50	4.90	6.50
11	2.60	5.00	6.60
12	2.70	5.10	6.70
13	2.80	5.20	6.80
14	2.90	5.30	6.90
15	3.00	5.40	6.90
16	3.00	5.40	7.00
17	3.10	5.50	7.10
18	3.10	5.50	7.10
19	3.20	5.60	7.20
20	3.30	5.70	7.20
21	3.30	5.70	7.30
22	3.40	5.80	7.30
23	3.40	5.80	7.40
24	3.40	5.80	7.40
25	3.50	5.90	7.50
26	3.50	5.90	7.50
27	3.60	6.00	7.50
28	3.60	6.00	7.60
29	3.60	6.00	7.60
30	3.70	6.10	7.70

Fichier loop
de MFOLD
énergie de
déstabilisation
d'une loop à
37 degrés C
(kCal/mol)

Fichiers

Fichier tloops de
MFOLD
Tetraloops :
Énergie à
37 degrés C
(kCal/mol)

Seq	Energy
GGGGAC	-3.00
GGUGAC	-3.00
CGAAAG	-3.00
GGAGAC	-3.00
CGCAAG	-3.00
GGAAAC	-3.00
CGGAAG	-3.00
CUIUCG	-3.00
CGUGAG	-3.00
CGAAGG	-2.50
CUACGG	-2.50
GGCAAC	-2.50
CGCGAG	-2.50
UGAGAG	-2.50
CGAGAG	-2.00
AGAAAU	-2.00
CGUAAG	-2.00
CUAACG	-2.00
UGAAAG	-2.00
GGAAAGC	-1.50
GGGAAC	-1.50
UGAAAA	-1.50
AGCAAU	-1.50
AGUAAU	-1.50
CGGGAG	-1.50
AGUGAU	-1.50
GGCGAC	-1.50
GGGAGC	-1.50
GUGAAC	-1.50
UGGAAA	-1.50

Énergie d'empilement (stacking)

- Les énergies d'empilement sont données en 16 (4x4) tableaux de 16 (4x4) nombres
- Par convention, A, C, G, T/U correspondent à 1, 2, 3 et 4 respectivement

Pour un empilement :

5' -WX- 3'
3' -ZY- 5' , l'énergie cor-

respondante est dans le tableau de la Wième ligne et la Zième colonne, et dans ce tableau, à la Xième ligne et la Yième colonne. Par exemple, pour W=1 et Z=4 :

A	C	G	U
5'	--> 3'		
AX			
UY			
3'	<- 5'		
.	.	.	-0.90
.	.	-2.20	.
.	-2.10	.	-0.60
-1.10	.	-1.40	.

Énergie d'empilement (stacking)

Y				Y				Y				Y			
A	C	G	U	A	C	G	U	A	C	G	U	A	C	G	U
5' --> 3'				5' --> 3'				5' --> 3'				5' --> 3'			
AX				AX				AX				AX			
AY				CY				GY				UY			
3' <- 5'				3' <- 5'				3' <- 5'				3' <- 5'			
.	-2.28	.	-0.90
.	-2.10	.	-0.60	
.	-1.10	.	-1.40	
Y				Y				Y				Y			
A	C	G	U	A	C	G	U	A	C	G	U	A	C	G	U
5' --> 3'				5' --> 3'				5' --> 3'				5' --> 3'			
CX				CX				CX				CX			
AY				CY				GY				UY			
3' <- 5'				3' <- 5'				3' <- 5'				3' <- 5'			
.	-2.10	.	.	
.	-3.30	.	.	
.	-2.40	.	-1.40	
.	-2.10	.	-2.10	
Y				Y				Y				Y			
A	C	G	U	A	C	G	U	A	C	G	U	A	C	G	U
5' --> 3'				5' --> 3'				5' --> 3'				5' --> 3'			
GX				GX				GX				GX			
AY				CY				GY				UY			
3' <- 5'				3' <- 5'				3' <- 5'				3' <- 5'			
.	-1.30	.	
.	-3.40	.	.	
.	-3.30	.	-1.50	
.	-2.28	.	-2.50	
Y				Y				Y				Y			
A	C	G	U	A	C	G	U	A	C	G	U	A	C	G	U
5' --> 3'				5' --> 3'				5' --> 3'				5' --> 3'			
UX				UX				UX				UX			
AY				CY				GY				UY			
3' <- 5'				3' <- 5'				3' <- 5'				3' <- 5'			
.	-1.00	.	.	
.	-2.40	.	-1.50	
.	-2.10	.	-0.50	
.	-0.90	.	-1.30	

Fichier stack
de MFOLD
Enthalpies
d'empilement
à 37 degrés C
(kCal/mol)

Énergie pour les Bulges et Interior loops

Exemple de Bulge



- ① Internal loop energy penalty (loop size = 5) : +1.8 kcal/mol
 - ② Terminal stacking energies for the mismatched base pairs
 - adjacent to CG base pair (CG/CU) : 0.0 kcal/mol
 - adjacent to GC base pair (CG/AC) : 0.0 kcal/mol
 - ③ For non-symmetric interior loops, there is an asymmetric loop penalty : 0.5 kcal/mol
- Total : +2.3 kcal/mol

Énergie pour les Bulges et Interior loops

Fichier tstacki
de MFOLD
Interior Loops :
Enthalpies selon
les mismatchs
terminaux et les
paires de bases
à 37 degrés C
(kCal/mol)

Y	Y	Y	Y
A C G U	A C G U	A C G U	A C G U
5' -> 3'	5' -> 3'	5' -> 3'	5' -> 3'
AX	AX	AX	AX
AY	CY	GY	UY
3' <- 5'	3' <- 5'	3' <- 5'	3' <- 5'
.	.	.	0.70 0.70 -0.40 0.70
.	.	.	0.70 0.70 0.70 0.70
.	.	.	-0.40 0.70 0.70 0.70
.	.	.	0.70 0.70 0.70 0.00
Y	Y	Y	Y
A C G U	A C G U	A C G U	A C G U
5' -> 3'	5' -> 3'	5' -> 3'	5' -> 3'
CX	CX	CX	CX
AY	CY	GY	UY
3' <- 5'	3' <- 5'	3' <- 5'	3' <- 5'
.	.	-0.00 -0.00 -1.10 -0.00	.
.	.	-0.00 -0.00 -0.00 -0.00	.
.	.	-1.10 -0.00 -0.00 -0.00	.
.	.	-0.00 -0.00 -0.00 -0.70	.
Y	Y	Y	Y
A C G U	A C G U	A C G U	A C G U
5' -> 3'	5' -> 3'	5' -> 3'	5' -> 3'
GX	GX	GX	GX
AY	Y	GY	UY
3' <- 5'	3' <- 5'	3' <- 5'	3' <- 5'
.	-0.00 -0.00 -1.10 -0.00	.	0.70 0.70 -0.40 0.70
.	-0.00 -0.00 -0.00 -0.00	.	0.70 0.70 0.70 0.70
.	-1.10 -0.00 -0.00 -0.00	.	-0.40 0.70 0.70 0.70
.	-0.00 -0.00 -0.00 -0.70	.	0.70 0.70 0.70 0.00
Y	Y	Y	Y
A C G U	A C G U	A C G U	A C G U
5' -> 3'	5' -> 3'	5' -> 3'	5' -> 3'
UX	UX	UX	UX
AY	CY	GY	UY
3' <- 5'	3' <- 5'	3' <- 5'	3' <- 5'
0.70 0.70 -0.40 0.70	.	0.70 0.70 -0.40 0.70	.
0.70 0.70 0.70 0.70	.	0.70 0.70 0.70 0.70	.
-0.40 0.70 0.70 0.70	.	-0.40 0.70 0.70 0.70	.
0.70 0.70 0.70 0.00	.	0.70 0.70 0.70 0.00	.

MFOLD : L'algorithme

- Principe similaire à Nussinov, mais on a 2 tableaux : W et V
 $W_{i,j}$ = énergie (minimum) du repliement optimal de $S[i..j]$
 $V_{i,j}$ = idem mais quand i et j forment une paire de bases ensemble
- étape de base de la programmation dynamique :
 $W_{i,j} = V_{i,j} = \infty$ si $j \leq i + L$

MFOLD : L'algorithme

- Récursion :

- $W_{i,j} = \min\{W_{i,j-1}; W_{i+1,j}; \min_{i < k < j}\{W_{i,k} + W_{k+1,j}\}; V_{i,j}\}$
- $V_{i,j} = e_h(i,j)$ si (i,j) forme une hairpin
ou $e_s(i,j) + V_{i+1,j-1}$ si (i,j) est une 1-loop
ou $e_{bi}(i,j, i',j')$ si (i,j) est une 2-loop

(multiloops non incluses, compliquées à traiter)

- Complexité : $\mathcal{O}(n^3)$

MFOLD : Le logiciel

Entrée :

- une séquence
- des paramètres énergétiques
- un ensemble de contraintes

F 23 87 3 va forcer les paires de bases 23.87, 24.86 et 25.85

Sortie :

- énergie dot-plot
- RNAML
- un ensemble de structures
- Annotations

MFOLD : Le logiciel

Output of mrgraph
by D. Stoye and M. Zuker



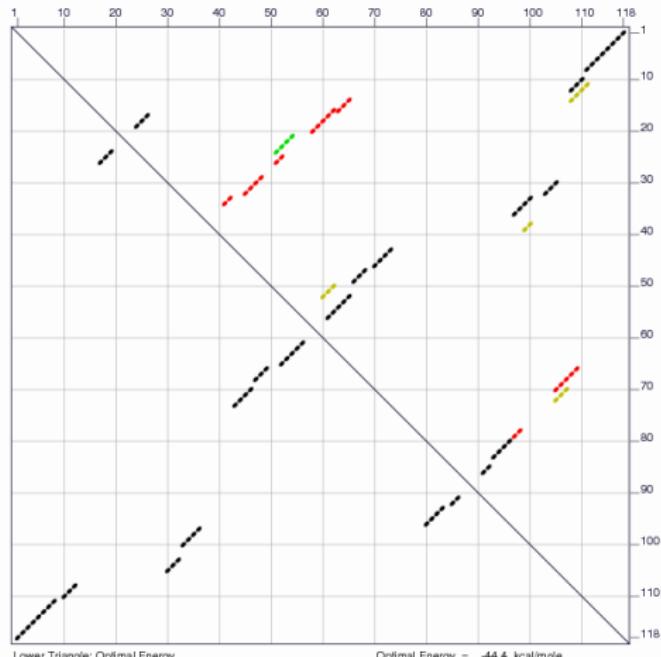
dG = -38.88 [initially -44.40] *Arabidopsis thaliana* 1

MFOLD : Le logiciel

Output of `arab1.g1`
by D. Slawson and M. Zuker

Fold of *Arabidopsis thaliana* 1 at 37° C.

8G In Plot File = 2.2 kcal/mole



Énergie et probabilités : Vienna

Idée : calculer la structure la plus probable (du point de vue énergie) en fonction des paramètres énergétiques choisis.

Mise en œuvre :

- fonction de partition : $\begin{cases} S = & \text{séquence} \\ \mathcal{R} = & \text{espace des repliements de } S \end{cases}$

$$Q_S = \sum_{r \in \mathcal{R}} e^{-G(r)/RT},$$

où $G(r)$ est l'énergie minimum de r , et R , T des constantes connues

Énergie et probabilités : Vienna

- probabilité d'un repliement r :

$$P(r|S) = e^{-G(r)/RT} / Q_S$$

- calcul de Q_S : « parallèle à MFOLD »

$$G(r) = \sum_{L_i} G(L_i)$$

$$e^{-G(r)/RT} = \prod_{L_i} e^{-G(L_i)/RT}$$

⇒ algorithme calqué sur celui de MFOLD

Corollaire : pour une paire de bases (i, j) donnée, on peut calculer la probabilité qu'elle appartienne à un repliement de S .

Énergie et probabilités : Vienna

- [RNAfold](#) -- predict minimum energy secondary structures and pair probabilities
- [RNACeval](#) -- evaluate energy of RNA secondary structures
- [RNAheat](#) -- calculate the specific heat (melting curve) of an RNA sequence
- [RNAinverse](#) -- inverse fold (design) sequences with predefined structure
- [RNAdistance](#) -- compare secondary structures
- [RNAPdist](#) -- compare base pair probabilities
- [RNAsubopt](#) -- complete suboptimal folding
- [RNAPlot](#) -- RNA structure drawings in PostScript, SVG, or GML
- [RNACofold](#) -- predict hybrid structure of two sequences
- [RNAduplex](#) -- predict possible hybridization sites between two sequences
- [RNAup](#) -- predict RNA-RNA interaction sites using accessibilities
- [RNAalifold](#) -- predict the consensus structure of several aligned sequences
- [RNAaliduplex](#) -- comparative (multiple alignment) version of RNAduplex
- [RNALfold](#) -- predict locally stable structure of long sequences
- [RNAlpfold](#) -- compute average pair probabilities for local base pairs in long sequences
- [RNAlpaln](#) -- fast structural alignment of RNA sequences using string alignments
- Several small but helpful Perl [Utilities](#)

Énergie et probabilités : Vienna

RNAfold WebServer

1 Enter Input
Parameters

2 View Results

[Home] [New Job] [Help]

The RNAfold web server will predict secondary structures of single stranded RNA or DNA sequences. Current limits are 7,500 nt for partition function calculations and 10,000 nt for minimum free energy only predictions.

Simply paste or upload your sequence below and click Proceed. To get more information on the meaning of the options click the ⓘ symbols. You can test the server using this sample sequence.

Paste or type your sequence here:

[clear]

GUCCUCUAGCUACAAACCUCAAAGCCCCCACUGAAGAUGCCAAGACGGUAACCACCAUACCUGAGGACA

Hide constraint folding

| : paired with another base
x : base must not pair
. : no constraint at all

> : base i is paired with a base j>i
< : base i is paired with a base j<i
matching brackets () : base i pairs base j

Paste or type your structure constraint using the symbols described above here:

[clear]

Note: The string for the structure constraint must be of the length of the sequence. Leave this field blank if no constraints should be applied during structure predictions.

Or upload a file in FASTA format: No file chosen

Fold algorithms and basic options

- minimum free energy (MFE) and partition function ⓘ
- minimum free energy (MFE) only ⓘ
- no GU pairs at the end of helices ⓘ
- avoid isolated base pairs ⓘ

Show advanced options

Output options

- Interactive RNA secondary structure plot ⓘ
- RNA secondary structure plots with reliability annotation (Partition function folding only) ⓘ
- Mountain plot ⓘ

Notification via e-mail upon completion of the job (optional):

Proceed »

Énergie et probabilités : Vienna

Results for minimum free energy prediction

The optimal secondary structure in dot-bracket notation with a minimum free energy of -17.20 kcal/mol is given below.

[color by base-pairing probability | color by positional entropy | no coloring]

1 GUCCUCAUAGCUUACAAACCUAANGCGGGCACUGAAGAUGCACAGCGGUUACCACCAUACCUGAGGACA

1 (((((.-((((.....))))..(((.....))))....((((.....))))))))..

You can download the minimum free energy (MFE) structure in [Vienna Format| Ct Format]. You can get thermodynamic details on this structure by submitting to our RI

Results for thermodynamic ensemble prediction

The free energy of the thermodynamic ensemble is -17.91 kcal/mol.

The frequency of the MFE structure in the ensemble is 31.59 %.

The ensemble diversity is 4.81 .

You may look at the dot plot containing the base pair probabilities [EPS|PDF|IMAGE CONVERTER].

The centroid secondary structure in dot-bracket notation with a minimum free energy of -17.20 kcal/mol is given below.

[color by base-pairing probability | color by positional entropy | no coloring]

1 GUCCUCAUAGCUUACAAACCUAANGCGGGCACUGAAGAUGCACAGCGGUUACCACCAUACCUGAGGACA

1 (((((.-((((.....))))..(((.....))))....((((.....))))))))..

You can download the minimum free energy (MFE) structure in [Vienna Format| Ct Format]. You can get thermodynamic details on this structure by submitting to our RI

Graphical output

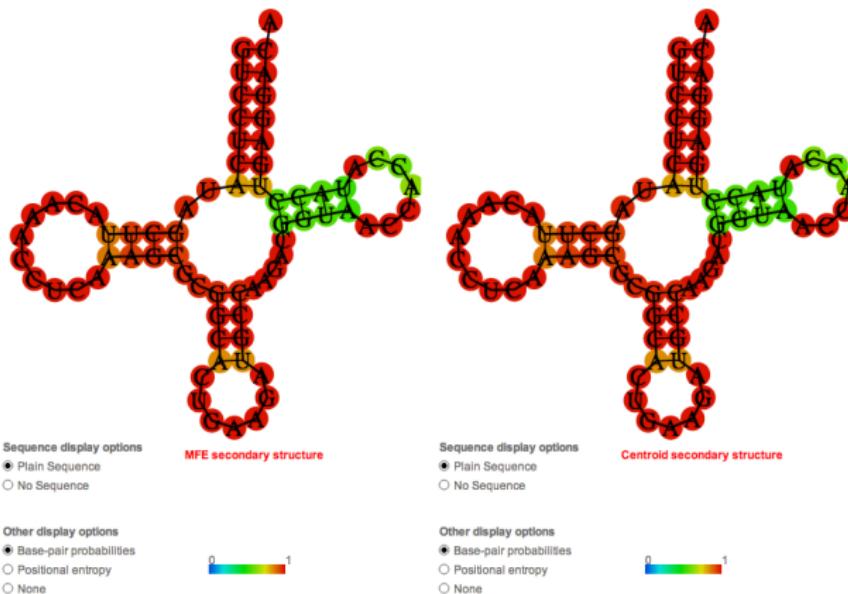
You may look at the interactive drawing of the MFE structure below. If you do not see the interactive drawing and you are using Internet Explorer, please install the Adobe

The structure below is colored by base-pairing probabilities. For unpaired regions the color denotes the probability of being unpaired.

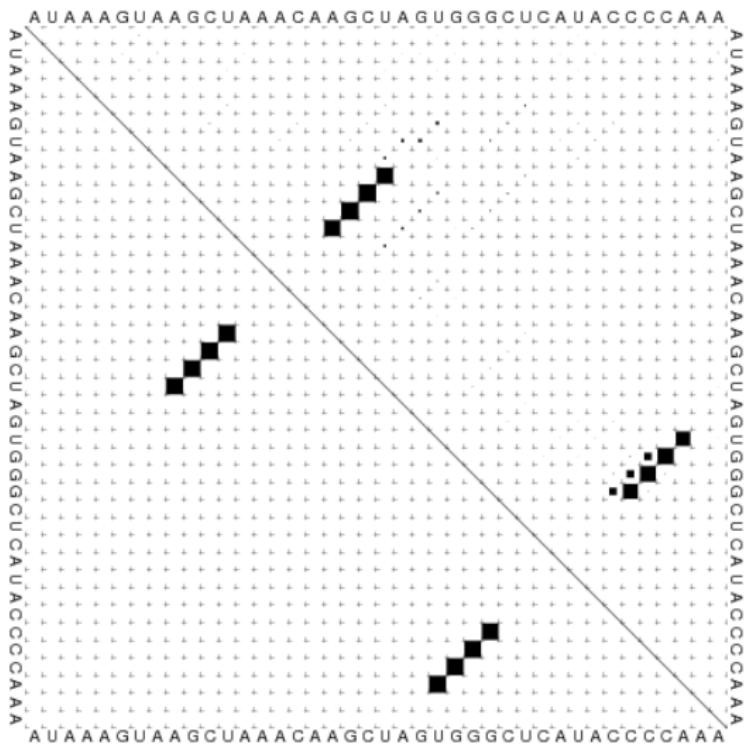
Énergie et probabilités : Vienna

Graphical output

You may look at the interactive drawing of the MFE structure below. If you do not see the interactive drawing and you are using Internet Explorer, please install the [Adobe SVG plugin](#). A note on base-pairing probabilities: The structure below is colored by base-pairing probabilities. For unpaired regions the color denotes the probability of being unpaired.



Énergie et probabilités : Vienna



Analyse de covariation de séquences

- ① La structure a plus d'importance que la séquence vis-à-vis de la fonction d'un gène : des séquences de même fonction dans plusieurs organismes auront même structure (à peu près) mais des séquences très différentes.
- ② La structure secondaire des ARN étant créée par des paires de bases, une mutation d'une base ne modifiant pas la structure devra être compensée par une mutation de l'autre base de la paire :



- ③ On va donc analyser ces covariations chez plusieurs organismes pour prédire la structure secondaire.

Covariation

	Acceptor Stem												
	D Stem				Anticodon Stem				T _Ψ C Stem				
Phe: <i>Amenellum quadruplicatum</i>	GCCAGGA	UA	GCNC	AGUUGGUA	GAAC	A	DAGGA	CUGAAA	UCCUC	GUGUC	GGGG	A	
Phe: <i>Spinacea oleacea</i>	GUCCCCA	UA	GCUC	AGCUUGGA	GAAC	A	GAGGA	CUGAAA	UCCUC	GUGUC	ACCAG	UCAAAU	A
RDGD	+ +			+			++	++	++	++	-	-	A
85.3% Similarity													
Phe: <i>S. cerevisiae</i>	GCGGAAU	UA	GCUC	AGUUGGGA	GAAC	G	CCAGA	CUGAAA	UCUQQ	AGGUC	CUQUG	UUCAAUC	A
Phe: <i>Bos taurus</i>	GCGGAAA	UA	GCUC	AGUUGGGA	GAAC	G	CCAGA	CUGAAA	UCUAA	AGGUC	CCUGG	UCCAUU	A
RDGD	- -						++	+	++	+	-	-	A
72.9% Similarity													
Phe: <i>S. cerevisiae</i>	GCGGAAU	UA	GCUC	AGUUGGGA	GAAC	G	CCAGA	CUGAGA	UCUQG	AGGUC	CUGUG	UUCCGAUD	A
Ala: <i>Thp. tenax</i>	GCGCCCG	UA	GUUC	AGC-GGA	GGAC	G	CCCAG	CUGAGA	UCUQG	AGGUC	CCGGG	UUUUCG	A
RDGD	- -		+++	+	++		-	-	-	+	-	-	A
53.4% Similarity							30	30	40	50	60	70	73

Figure 1. Reddot-greendot examples from tRNA. Symbols used: +: transition; -: transversion; |: deletion; *: ambiguous nucleotide. Experimentally verified helices from the secondary structure are boxed and connected with black lines. Nucleotide position numbers refer to the *S.cerevisiae* Phe reference sequence. Sequence names are shown as amino acid:organism.

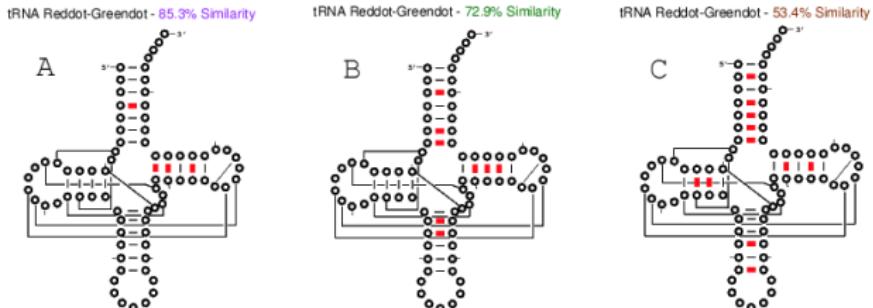


Figure 2. Results of the reddot-greendot analysis shown on tRNA secondary structure diagrams. Base pairs which are predicted with the method are shown with red tick marks. A. Sequences with 85.3% similarity. B. 72.9% similarity. C. 53.4% similarity.

Mutual Information content : MIX

- ① Le problème : on a un alignement de séquences dont on suppose qu'elles ont même structure (ou à peu près) : prédire cette structure
- ② Calcul : matrice M des scores MIX
 - i, j : colonnes de l'alignement
 - $f_i(X)$: fréquence de la base X en colonne i
 - (X, Y) : paire de bases AU, UA, GC, CG, GU, UG
 - $f_{i,j}(X, Y)$: fréquence de la paire de bases (X, Y) en i et j .
 - $$M_{i,j} = \sum_{X,Y} f_{i,j}(X, Y) \log_2 \left(\frac{f_{i,j}(X, Y)}{f_i(X)f_j(Y)} \right)$$
 - $M_{i,j} \Rightarrow$ matrice de scores $\in [0, 2]$: plus $M_{i,j}$ est élevé, plus les covariations en colonnes i et j supportent l'hypothèse d'une paire de bases entre ces deux positions.

MIX

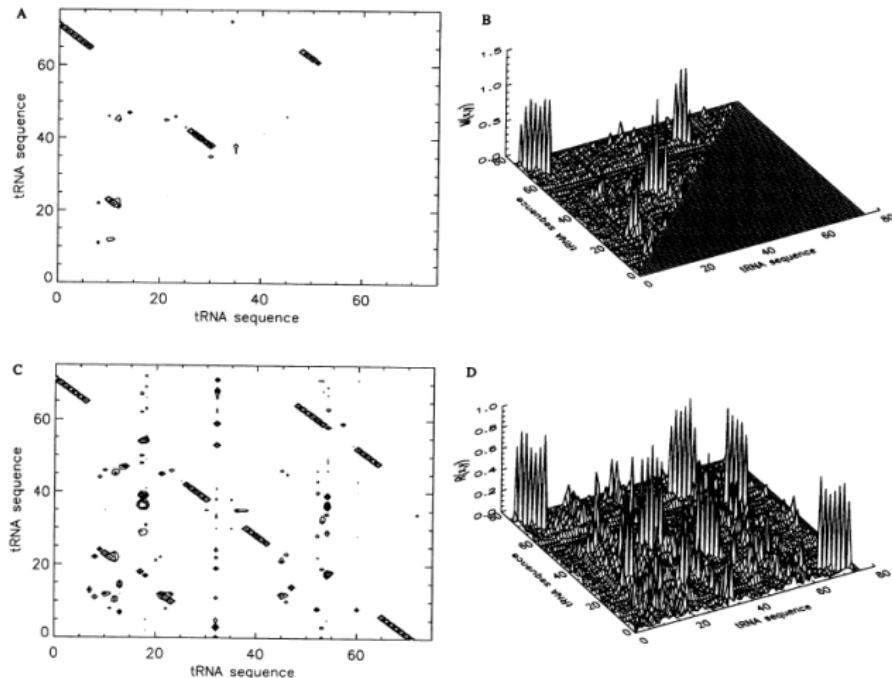


Figure 2. Graphical display of $M(x,y)$ and R values. Only values above 0.2 are displayed on the Contour plots. A: Contour plot of $M(x,y)$ values. B: Surface plot of $M(x,y)$ values. C: Contour plot of R values. The values are determined by taking the values from $M(x,y)$ (shown in part A) and replicating them symmetrically into the other half of the matrix, and then dividing each row by the entropy of the position on the vertical axis. As described in the text, $R_1(x,y) = R_2(y,x)$, so this plot shows both values. If the vertical axis is considered to be position x , then the plot is of $R_1(x,y)$; if the vertical axis is considered to be position y then the plot is of $R_2(x,y)$. Sorting by $R_1(x,y)$ is equivalent to sorting within rows and sorting by $R_2(x,y)$ is equivalent to sorting within columns. D: Surface plot of R values. The values are the same as in part C, but displayed as a 3-D plot.

Atelier : Matériel

- ① Programme Nussinov (<http://baba.sourceforge.net>)
- ② The mfold webserver (<http://unafold.rna.albany.edu>)
- ③ RNAfold webserver
(<http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi>)
- ④ Rfam (<http://rfam.xfam.org>)
- ⑤ mirBase (<http://www.mirbase.org>)

Atelier : Énoncé

- ① Parcourir la base de données Rfam
- ② Chercher la famille mir156 Rfam
- ③ Chercher la famille mir156 mirBase et télécharger la séquence du précurseur ath-mir156a
- ④ Replier la sequence avec les trois outils : nussinov, mfold et rnafold
- ⑤ Parcourir les résultats des trois programmes
- ⑥ Comparer ces résultats

Lecture

