

Analyse comparative pour l'étude des gènes d'ARN

Hélène Touzet

Équipe Bioinfo — LIFL — USTL

<http://www.lifl.fr/BIOINFO>



modèle
thermodynamique

MFOLD RNAfold

analyse
comparative
pure

mutations
compensatoires

**prédiction
de
structures**

RNAalifold
caRNAc

**prédiction
de
gènes**

QRNA
RNAz

Approche thermodynamique

- ▶ **P**ostulat fondateur: à l'équilibre, la molécule se replie dans une configuration qui minimise son énergie libre
- ▶ **A**pplication: prédiction de structure (secondaire)
- ▶ **C**oeur algorithmique: dénombrement des appariements
- ▶ **D**éclinaisons: Mfold, RNAfold

How do RNA folding algorithms work?

Eddy SR Nat Biotechnol. 2004 Nov ; 22(11): 1457-8

Modèle école

- ▶ **Modèle énergétique additif**

On néglige les contributions énergétiques entre appartements voisins

- ▶ **Plusieurs déclinaisons**

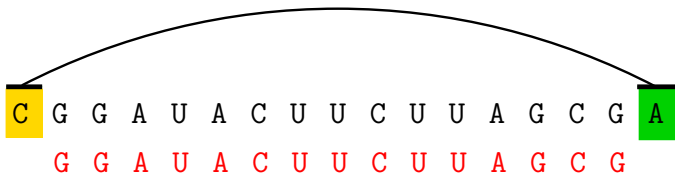
- ▶ Optimisation par programmation dynamique (Nussinov - 1978)
- ▶ Modélisation par des grammaires stochastiques hors-contextes
- ▶ Modélisation en terme de filtrage maximal sans croisement dans un cercle

C G G A U A C U U C U U A G C G A

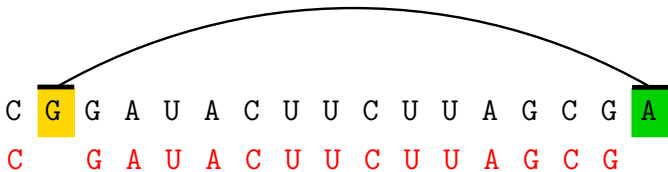
C G G A U A C U U C U U A G C G A

C G G A U A C U U C U U A G C G A
C G G A U A C U U C U U A G C G

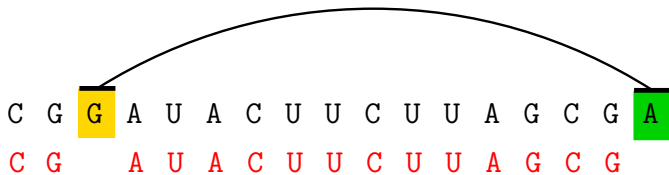
► **Cas 1** : A est libre



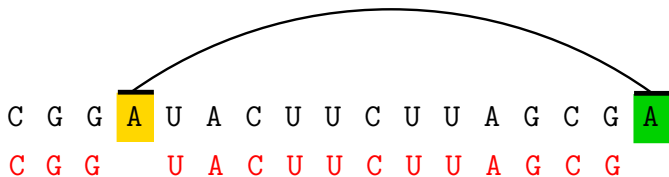
- ▶ **Cas 1** : A est libre
- ▶ **Cas 2** : A est apparié avec C



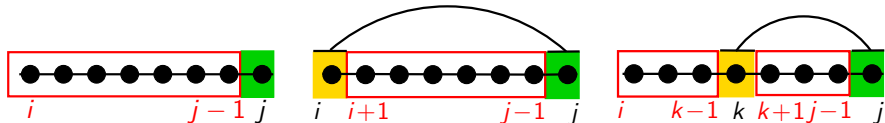
- ▶ **Cas 1** : A est libre
- ▶ **Cas 2** : A est apparié avec C
- ▶ **Cas 3** : A est apparié avec une autre base en 5'



- ▶ **Cas 1** : A est libre
- ▶ **Cas 2** : A est apparié avec C
- ▶ **Cas 3** : A est apparié avec une autre base en 5'



- ▶ **Cas 1** : A est libre
- ▶ **Cas 2** : A est apparié avec C
- ▶ **Cas 3** : A est apparié avec une autre base en 5'



$$E(i, j) = \min \begin{cases} E(i, j-1) \\ E(i+1, j-1) + \alpha(i, j) \\ \min\{E(i, k-1) + \alpha(k, j) + E(k+1, j-1), i+1 \leq k \leq j-1\} \end{cases}$$

- ▶ $E(i, j)$: énergie optimale pour la sous-séquence entre les positions i et j
- ▶ $\alpha(k, j)$: paramètre d'énergie pour la formation de l'appariement entre la base k et la base j
- ▶ Mise en œuvre par programmation dynamique

	C	G	G	A	U	A	C	U	U	C	U	U	A	G	A	C	G	A
C	0	0	0	0	0	0	3	4	4	6	6	6	6	9	9	11	14	14
G		0	0	0	0	0	3	4	4	6	6	6	6	7	9	11	11	11
G			0	0	0	0	3	3	3	5	5	5	5	7	8	10	10	10
A				0	0	0	0	2	2	2	2	4	4	5	7	7	8	10
U					0	0	0	0	0	0	2	2	4	5	7	7	8	10
A						0	0	0	0	0	2	2	2	5	5	5	8	8
C							0	0	0	0	0	0	2	5	5	5	8	8
U								0	0	0	0	0	2	3	5	5	6	7
U									0	0	0	0	2	3	5	5	5	7
C										0	0	0	0	3	3	3	5	5
U											0	0	0	0	2	2	2	3
U												0	0	0	0	0	1	1
A													0	0	0	0	0	0
G														0	0	0	0	0
A															0	0	0	0
C																0	0	0
G																	0	0
A																		0

A=U: 2 G-U:1 G≡C:3

	C	G	G	A	U	A	C	U	U	C	U	U	A	G	A	C	G	A
C	0	0	0	0	0	0	3	4	4	6	6	6	6	9	9	11	14	14
G		0	0	0	0	0	3	4	4	6	6	6	6	7	9	11	11	11
G			0	0	0	0	3	3	3	5	5	5	5	7	8	10	10	10
A				0	0	0	0	2	2	2	2	4	4	5	7	7	8	10
U					0	0	0	0	0	0	2	2	4	5	7	7	8	10
A						0	0	0	0	0	2	2	2	5	5	5	8	8
C							0	0	0	0	0	0	2	5	5	5	8	8
U								0	0	0	0	0	2	3	5	5	6	7
U									0	0	0	0	2	3	5	5	5	7
C										0	0	0	0	3	3	3	5	5
U											0	0	0	0	2	2	2	3
U												0	0	0	0	0	1	1
A													0	0	0	0	0	0
G														0	0	0	0	0
A															0	0	0	0
C																0	0	0
G																	0	0
A																		0

A=U: 2 G-U:1 G≡C:3

	C	G	G	A	U	A	C	U	U	C	U	U	A	G	A	C	G	A
C	0	0	0	0	0	0	3	4	4	6	6	6	6	9	9	11	14	14
G		0	0	0	0	0	3	4	4	6	6	6	6	7	9	11	11	11
G			0	0	0	0	3	3	3	5	5	5	5	7	8	10	10	10
A				0	0	0	0	2	2	2	2	4	4	5	7	7	8	10
U					0	0	0	0	0	0	2	2	4	5	7	7	8	10
A						0	0	0	0	0	2	2	2	5	5	5	8	8
C							0	0	0	0	0	0	2	5	5	5	8	8
U								0	0	0	0	0	2	3	5	5	6	7
U									0	0	0	0	2	3	5	5	5	7
C										0	0	0	0	3	3	3	5	5
U											0	0	0	0	2	2	2	3
U												0	0	0	0	0	1	1
A													0	0	0	0	0	0
G														0	0	0	0	0
A															0	0	0	0
C																0	0	0
G																	0	0
A																		0

A=U: 2 G=U:1 G=C:3

	C	G	G	A	U	A	C	U	U	C	U	U	A	G	A	C	G	A
C	0	0	0	0	0	0	3	4	4	6	6	6	6	9	9	11	14	14
G		0	0	0	0	0	3	4	4	6	6	6	6	7	9	11	11	11
G			0	0	0	0	3	3	3	5	5	5	5	7	8	10	10	10
A				0	0	0	0	2	2	2	2	4	4	5	7	7	8	10
U					0	0	0	0	0	0	2	2	4	5	7	7	8	10
A						0	0	0	0	0	2	2	2	5	5	5	8	8
C							0	0	0	0	0	0	2	5	5	5	8	8
U								0	0	0	0	0	2	3	5	5	6	7
U									0	0	0	0	2	3	5	5	5	7
C										0	0	0	0	3	3	3	5	5
U											0	0	0	0	2	2	2	3
U												0	0	0	0	0	1	1
A													0	0	0	0	0	0
G														0	0	0	0	0
A															0	0	0	0
C																0	0	0
G																	0	0
A																		0

A=U: 2 G-U:1 G=C:3

	C	G	G	A	U	A	C	U	U	C	U	U	A	G	A	C	G	A
C	0	0	0	0	0	0	3	4	4	6	6	6	6	9	9	11	14	14
G		0	0	0	0	0	3	4	4	6	6	6	6	7	9	11	11	11
G			0	0	0	0	3	3	3	5	5	5	5	7	8	10	10	10
A				0	0	0	0	2	2	2	2	4	4	5	7	7	8	10
U					0	0	0	0	0	0	2	2	4	5	7	7	8	10
A						0	0	0	0	0	2	2	2	5	5	5	8	8
C							0	0	0	0	0	0	2	5	5	5	8	8
U								0	0	0	0	0	2	3	5	5	6	7
U									0	0	0	0	2	3	5	5	5	7
C										0	0	0	0	3	3	3	5	5
U											0	0	0	0	2	2	2	3
U												0	0	0	0	0	1	1
A													0	0	0	0	0	0
G														0	0	0	0	0
A															0	0	0	0
C																0	0	0
G																	0	0
A																		0

A=U: 2 G-U:1 G=C:3

	(C	G	G	A	U	A	C	U	U	C	U	U	A	G	A	C)	.
	C	G	G	A	U	A	C	U	U	C	U	U	A	G	A	C	G	A	
C	0	0	0	0	0	0	3	4	4	6	6	6	6	9	9	11	14	14	
G		0	0	0	0	0	3	4	4	6	6	6	6	7	9	11	11	11	
G			0	0	0	0	3	3	3	5	5	5	5	7	8	10	10	10	
A				0	0	0	0	2	2	2	4	4	5	7	7	8	10		
U					0	0	0	0	0	2	2	4	5	7	7	8	10		
A						0	0	0	0	2	2	2	5	5	5	8	8		
C							0	0	0	0	0	2	5	5	5	8	8		
U								0	0	0	0	2	3	5	5	6	7		
U									0	0	0	2	3	5	5	5	7		
C										0	0	0	3	3	3	5	5		
U											0	0	0	2	2	2	3		
U												0	0	0	0	1	1		
A													0	0	0	0	0		
G														0	0	0	0		
A															0	0	0		
C																0	0		
G																	0	0	
A																		0	

A=U: 2 G-U:1 G=C:3

	(())	.	
	C	G	G	A	U	A	C	U	U	C	U	U	A	G	A	C	G	A
C	0	0	0	0	0	0	3	4	4	6	6	6	6	9	9	11	14	14
G		0	0	0	0	0	3	4	4	6	6	6	6	7	9	11	11	11
G			0	0	0	0	3	3	3	5	5	5	5	7	8	10	10	10
A				0	0	0	0	2	2	2	2	4	4	5	7	7	8	10
U					0	0	0	0	0	0	2	2	4	5	7	7	8	10
A						0	0	0	0	0	2	2	2	5	5	5	8	8
C							0	0	0	0	0	0	2	5	5	5	8	8
U								0	0	0	0	0	2	3	5	5	6	7
U									0	0	0	0	2	3	5	5	5	7
C										0	0	0	0	3	3	3	5	5
U											0	0	0	0	2	2	2	3
U												0	0	0	0	0	1	1
A													0	0	0	0	0	0
G														0	0	0	0	0
A															0	0	0	0
C																0	0	0
G																	0	0
A																		0

A=U: 2 G-U:1 G=C:3

	(())	.		
	C	G	G	A	U	A	C	U	U	C	U	U	A	G	A	C	G	A
C	0	0	0	0	0	0	3	4	4	6	6	6	6	9	9	11	14	14
G		0	0	0	0	0	3	4	4	6	6	6	6	7	9	11	11	11
G			0	0	0	0	3	3	3	5	5	5	5	7	8	10	10	10
A				0	0	0	0	2	2	2	2	4	4	5	7	7	8	10
U					0	0	0	0	0	0	2	2	4	5	7	7	8	10
A						0	0	0	0	0	2	2	2	5	5	5	8	8
C							0	0	0	0	0	0	2	5	5	5	8	8
U								0	0	0	0	0	2	3	5	5	6	7
U									0	0	0	0	2	3	5	5	5	7
C										0	0	0	0	3	3	3	5	5
U											0	0	0	0	2	2	2	3
U												0	0	0	0	0	1	1
A													0	0	0	0	0	0
G														0	0	0	0	0
A															0	0	0	0
C																0	0	0
G																	0	0
A																		0

A=U: 2 G-U:1 G≡C:3

	((G	A	U	A	C	U	U	C	U	U	A	G)))	.
	C	G	G	A	U	A	C	U	U	C	U	U	A	G	A	C	G	A
C	0	0	0	0	0	0	3	4	4	6	6	6	6	9	9	11	14	14
G		0	0	0	0	0	3	4	4	6	6	6	6	7	9	11	11	11
G			0	0	0	0	3	3	3	5	5	5	5	7	8	10	10	10
A				0	0	0	0	2	2	2	2	4	4	5	7	7	8	10
U					0	0	0	0	0	0	2	2	4	5	7	7	8	10
A						0	0	0	0	0	2	2	2	5	5	5	8	8
C							0	0	0	0	0	0	2	5	5	5	8	8
U								0	0	0	0	0	2	3	5	5	6	7
U									0	0	0	0	2	3	5	5	5	7
C										0	0	0	0	3	3	3	5	5
U											0	0	0	0	2	2	2	3
U												0	0	0	0	0	1	1
A													0	0	0	0	0	0
G														0	0	0	0	0
A															0	0	0	0
C																0	0	0
G																	0	0
A																		0

A=U: 2 G-U:1 G=C:3

	((G	A	U	A	C	(U	U	C	U	U	A	G)))	.
	C	G						U	U	C	U	U	A	G	A	C	G	A	
C	0	0	0	0	0	0	3	4	4	6	6	6	6	9	9	11	14	14	
G		0	0	0	0	0	3	4	4	6	6	6	6	7	9	11	11	11	
G			0	0	0	0	3	3	3	5	5	5	5	7	8	10	10	10	
A				0	0	0	0	2	2	2	2	4	4	5	7	7	8	10	
U					0	0	0	0	0	2	2	4	5	7	7	8	10		
A						0	0	0	0	2	2	2	5	5	5	8	8		
C							0	0	0	0	0	2	5	5	5	8	8		
U								0	0	0	0	2	3	5	5	6	7		
U									0	0	0	2	3	5	5	5	7		
C										0	0	0	3	3	3	5	5		
U											0	0	0	2	2	2	3		
U												0	0	0	0	1	1		
A													0	0	0	0	0		
G														0	0	0	0		
A															0	0	0		
C																0	0		
G																	0	0	
A																		0	

A=U: 2 G-U:1 G=C:3

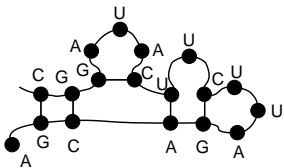
	(C	(G	(G	. A	. U	. A) C	(U	U	C	U	U	A	G) A) C) G	. A
C	0	0	0	0	0	0	3	4	4	6	6	6	6	9	9	11	14	14
G		0	0	0	0	0	3	4	4	6	6	6	6	7	9	11	11	11
G			0	0	0	0	3	3	3	5	5	5	5	7	8	10	10	10
A				0	0	0	0	2	2	2	2	4	4	5	7	7	8	10
U					0	0	0	0	0	0	2	2	4	5	7	7	8	10
A						0	0	0	0	0	2	2	2	5	5	5	8	8
C							0	0	0	0	0	0	2	5	5	5	8	8
U								0	0	0	0	0	2	3	5	5	6	7
U									0	0	0	0	2	3	5	5	5	7
C										0	0	0	0	3	3	3	5	5
U											0	0	0	0	2	2	2	3
U												0	0	0	0	0	1	1
A													0	0	0	0	0	0
G														0	0	0	0	0
A															0	0	0	0
C																0	0	0
G																	0	0
A																		0

A=U: 2 G-U:1 G=C:3

	(((.	.	.)	(.	(.	.	.))))	.
	C	G	G	A	U	A	C	U	U	C	U	U	A	G	A	C	G	A
C	0	0	0	0	0	0	3	4	4	6	6	6	6	9	9	11	14	14
G		0	0	0	0	0	3	4	4	6	6	6	6	7	9	11	11	11
G			0	0	0	0	3	3	3	5	5	5	5	7	8	10	10	10
A				0	0	0	0	2	2	2	2	4	4	5	7	7	8	10
U					0	0	0	0	0	2	2	4	5	7	7	8	10	
A						0	0	0	0	2	2	2	5	5	5	8	8	
C							0	0	0	0	0	2	5	5	5	8	8	
U								0	0	0	0	2	3	5	5	6	7	
U									0	0	0	0	2	3	5	5	5	7
C										0	0	0	0	3	3	3	5	5
U											0	0	0	0	2	2	2	3
U												0	0	0	0	0	1	1
A													0	0	0	0	0	0
G														0	0	0	0	0
A															0	0	0	0
C																0	0	0
G																	0	0
A																		0

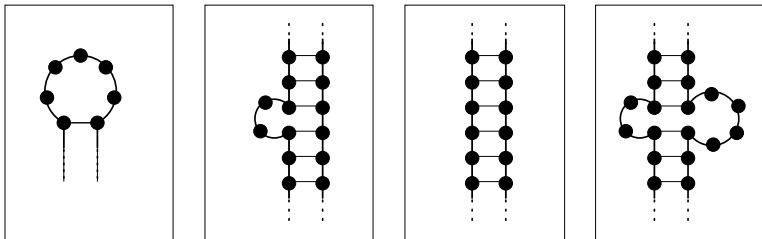
A=U: 2 G-U:1 G=C:3

	(((.	.	.)	(.	(.	.	.))))	.
	C	G	G	A	U	A	C	U	U	C	U	U	A	G	A	C	G	A
C	0	0	0	0	0	0	3	4	4	6	6	6	6	9	9	11	14	14
G		0	0	0	0	0	3	4	4	6	6	6	6	7	9	11	11	11
G			0	0	0	0	3	3	3	5	5	5	5	7	8	10	10	10
A				0	0	0	0	2	2	2	2	4	4	5	7	7	8	10
U					0	0	0	0	0	2	2	4	5	7	7	8	10	
A						0	0	0	0	2	2	2	5	5	5	8	8	
C							0	0	0	0	0	2	5	5	5	8	8	
U								0	0	0	0	2	3	5	5	6	7	
U									0	0	0	0	2	3	5	5	5	7
C										0	0	0	0	3	3	3	5	5
U											0	0	0	0	2	2	2	3
U												0	0	0	0	0	1	1
A													0	0	0	0	0	0
G														0	0	0	0	0
A															0	0	0	0
C																0	0	0
G																	0	0
A																		0

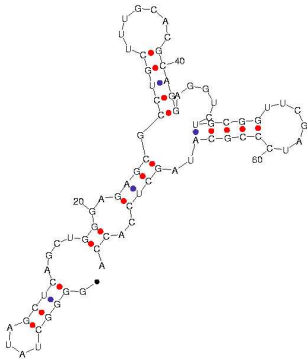


A=U: 2 G-U:1 G≡C:3

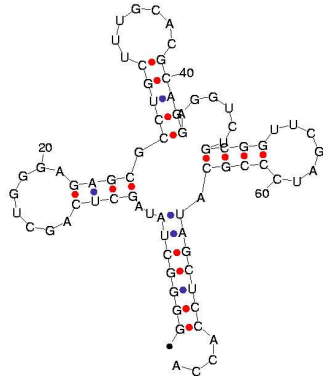
Modèle complet



- ▶ **Motifs structuraux**
- ▶ **Energie d'empilement (Turner)**
- ▶ **MFOLD - Zuker**
<http://www.bioinfo.rpi.edu/applications/mfold/>
- ▶ **RNAfold - Vienna Package - Hofacker**
<http://http://www.tbi.univie.ac.at/~ivo/RNA/>



dG - -27.57 [initially -29.7] tRNA_coi

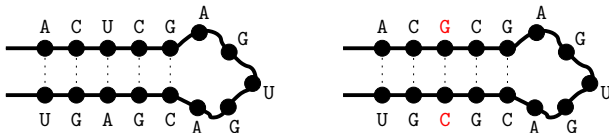


dG - -29.04 [initially -29.5] tRNA_coi


- ▶ **Nombreuses structures potentielles dont l'énergie libre est proche de l'optimum**
- ▶ **Ecueil des pseudo-nœuds**

Approche comparative

- ▶ Famille de séquences homologues alignées
- ▶ Phénomène de changement de base compensatoire



- ▶ Résolution de structure pour l'ARN de transfert, les ARN ribosomiques,...



U	C	A	A	C	G	G	U	G
-	C	G	A	C	G	G	C	A
U	C	G	A	C	G	G	C	A
U	U	C	A	C	G	U	G	G
U	A	U	A	-	G	G	A	G
U	A	G	A	C	U	G	C	G
U	G	A	A	C	G	G	U	G

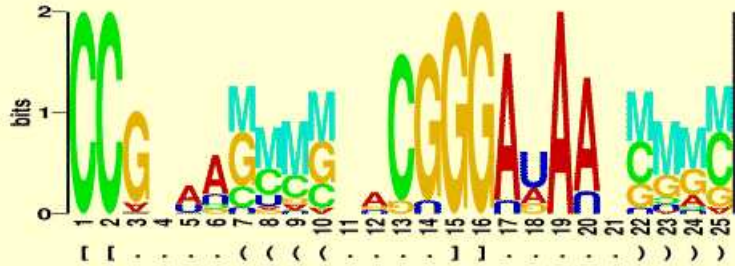
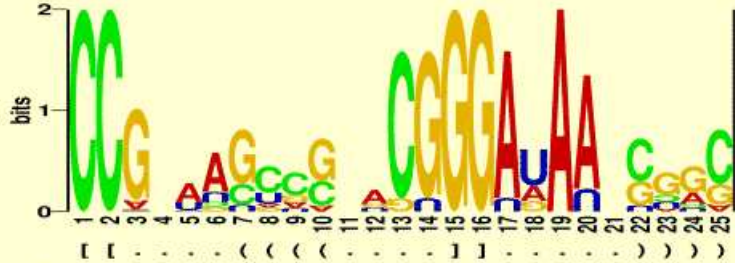
$$\mathcal{I}(i, j) = \sum f_{x,y}^{i,j} \log_2 \frac{f_{x,y}^{i,j}}{f_x^i f_x^j}$$

f_x^i : fréquence de la base x dans la colonne i

$f_{x,y}^{i,j}$: fréquence du couple (x, y) dans les colonnes i et j

Information mutuelle moyenne entre deux colonnes i et j :

- ▶ Quantité d'information révélée par la colonne j , la colonne i étant connue (valeur comprise entre 0 et 2)
- ▶ $\mathcal{I}(i, j)$ est maximale quand les colonnes i et j individuellement paraissent aléatoires et que les positions sont parfaitement corrélées
- ▶ Plus l'information mutuelle est élevée, plus la présomption d'appariement est forte



**Approche
thermodynamique**

Une séquence

Information
intrinsèque

+

**Approche
comparative**

Une famille de
séquences alignées

Informations
transverses



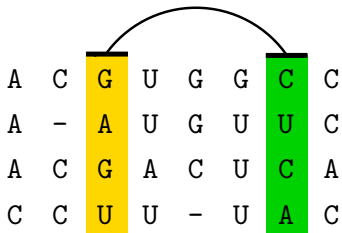
=

**Méthodes
hybrides**

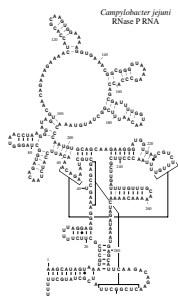
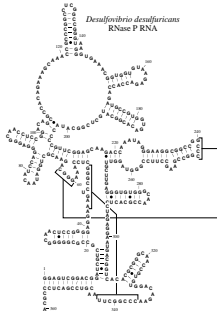
Quelques
séquences

Informations
transverses et
intrinsèques

À partir d'un alignement multiple



- ▶ **A**ppariement: couple de colonnes
- ▶ **R**NAalifold (Hofacker et al. - 2002)
 - ▶ algorithme de Nussinov
 - ▶ modèle d'énergie modifié:
moyenne des énergies d'appariement avec un bonus pour les mutations compensatoires
- ▶ **P**fold (Knudsen et al. - 2003)
 - ▶ grammaires stochastiques hors contexte
 - ▶ poids obtenus par apprentissage



```

<<<<<<<<<[[[[[[[[ << <<<< >>>> >> <<<< <<<<[[[[[[ ]]] ]]] [[
GGAGUCGGACGGAUUCGUCGCCGGGGGCAACUCC.....GGGGAGAAAGUCCGGGCUCCAAAGGGCAGAACCGUGGAUAAACUC...CAG..GG
AAGCAUAGU.....AAAUUCGUCGUUCUUUUAGGAGAGGAAAGUCCGAGCUCGUAAGAGACA.AAC...AUUCCAUCUAACAGAUUG
<<<<<<<<< <[[[[[[ <<<<< >>>>> <<<< << <<<[[[[[[ ]]]]]]]

```

```

[[ [ ] ] ]<< << [ [[[[[[ [ ]]]]]] ] [[[[[[ [ ]]]]]] [[[[[[ [
..AGGGCAACCU.CCGGACAGCGCCACAGAAAGCAAAC....CGCCCGGCCUCGGCCGGGUAAGGGUGAAACGGUGGUGUAAGAGACCACCAGAUGCCG...U
CUAGGGUAACCUAAGGGAUAGUGCAACAGAAAGAAAACUACCACGAAG.....UGGAAAAGGUGAAACGGCGGGGUAAGAGCCACCAG...CGAUUUU
[[[[ [ ]]]] ]<< << [[ [[[[ [ ] ]]] ]]] [[ [[ [[ [ ]]] ]]] [[ [[ [[

```

```

[ ] ] ]]] ] >>>>>>>>> <<<< <<<<<< <<<< >>>> >>>>> >>>
GGUGACA...CGGC.AUGCUCGGCAUACCCCGUUCGGAGCAAGACCAAUAGGGAAGG..CGGCCGGCCCGCCG.....AAGCCUUCGGGUGAGGUUG
GGUAAACAAUUCGGCUAU...GUAACCCAAUGUGCAGCAAGA.....AGGGAUGGUUAG.....CGUCUUUUGUUUUAACCCUUC.....G
[ ] ] ]]] >>>>>>>>> << <<<< <<<< >>>>>>>>

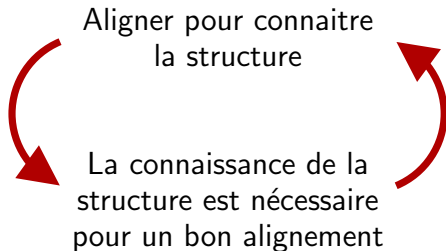
```

```

<<<<<<< >>>>>>> ]]]]]]] <<<<< >>>>> > >>>>>>>>
CUUGAGGGUGUGGGCAACCGCAUCUUCAGAGGAAUGACGGUCACACGCGGGCAACCGUGUGGACAGAACC CGGCUUACAGUCCGACUCC...GCA
CUUGAUUUUUGUUUGCAAAAACAAAACUAGAUAAAUGA.....GCAUUCAA.....GACAGAACUCGGCUUA.....UCGCUAUGCUUUUU
<<<<<<< >>>>>>> ] ]]]]] > >>>>>>>>

```

Inférence de structure sans alignement



Inférence de structure sans alignement



Algorithme de Sankoff:

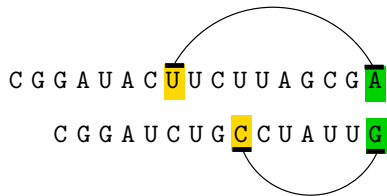
- ▶ **A** aligner et replier un couple de séquences simultanément
- ▶ **E**xtension de l'approche Nussinov à deux séquences, en ajoutant des possibilités d'insertion, de délétion
- ▶ **L**es poids peuvent tenir compte de l'énergie et des informations phylogénétiques

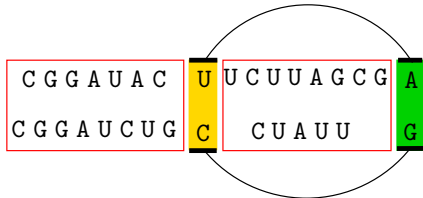
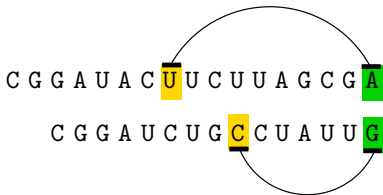
Simultaneous solution of the RNA folding, alignment and protosequence problems

D. Sankoff, SIAM J. Appl. Math., 45-5, p. 810-825, 1985

C G G A U A C U U C U U A G C G A

C G G A U C U G C C U A U U G





- ▶ **Problème de coût de calcul**
Sensible à la longueur et au nombre de séquences
- ▶ **Recours à des stratégies heuristiques**

Mise en pratique 1

- ▶ **Dynalign**

Implémentation de l'algorithme de Sankoff pour 2 séquences

- ▶ **Foldalign**

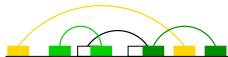
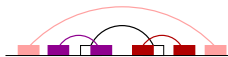
Recherche uniquement de la structure sans ramification la plus stable

Dynalign: An Algorithm for Finding the Secondary Structure Common to Two RNA Sequences
D.H. Mathews & D.H. Turner, Journal of Molecular Biology, 317, p 191-203, 2002

Finding the most significant common sequence and structure motifs in a set of RNA sequences
J. Gorodkin and L.J. Heyer and G.D. Stormo, Nucleic Acids Research, 25, 3724-3732, 1997

Mise en pratique 2

- ▶ **Modélisation macroscopique: tiges**
- ▶ **caRNAc**
 - ▶ algorithme de Sankoff + théorie des graphes
 - ▶ longues séquences
- ▶ **comRNA**
 - ▶ théorie des graphes
 - ▶ pseudo-noeuds
 - ▶ courtes séquences



Carnac : folding families of related RNAs

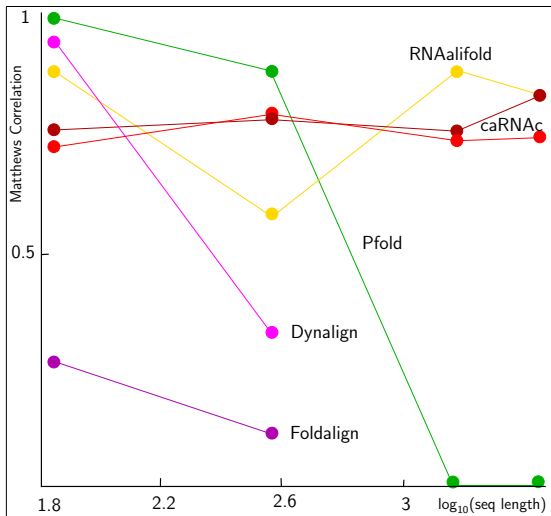
Hélène Touzet, Olivier Perriquet, Nucleic Acid Research 2004, 32, pages 142-145

A graph theoretical approach for predicting common RNA secondary structure motifs including pseudoknots in unaligned sequences

Y. Ji and X. Xu and G.D. Stormo Bioinformatics 2004, 20(10), 1591-1602

Comparaison des performances : BRAliBase

- ▶ **Sensibilité - sélectivité - coefficient de Matthews**
- ▶ **Jeux de données**
 - tRNA-phe (*S. cerevisiae*) – 73 nt
 - RNase P (*E. coli*) – 377 nt
 - SSU rRNA (*E. coli*) – 1542 nt
 - LSU rRNA (*E. coli*) – 2904 nt
- ▶ **Similitude : de 60% à 80%**
- ▶ **Alignements multiples corrects**
- ▶ www.binf.ku.dk/~pgardner/bralibase/bralibase1.html



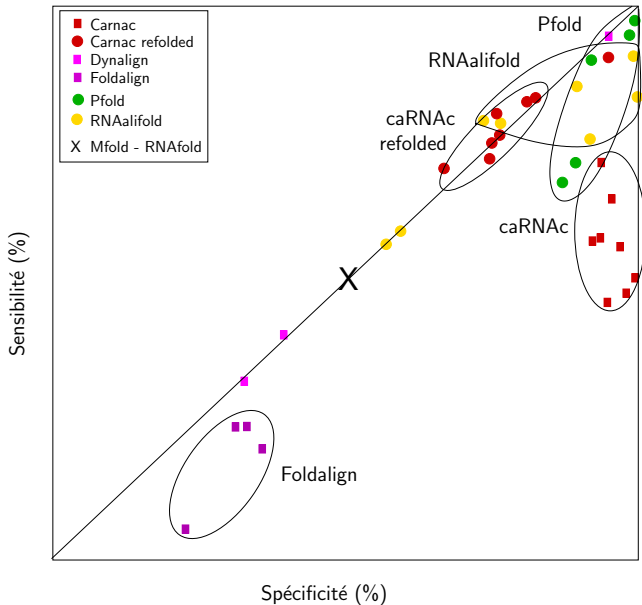
Méthodes avec alignement

- RNAalifold
- Pfold

Sans alignement préalable

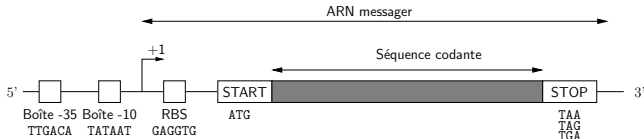
- caRNAc
- caRNAc refolded
- Dynalign
- Foldalign

Sensibilité versus spécificité



Prédiction de gène d'ARN

▶ Gènes à protéine



▶ Gènes d'ARN

- ▶ Niveau de l'énergie libre (Clote 2005)
 - ▶ Biais de composition en di-nucléotides (Schattner 2002)
 - ▶ Unicité du repliement (S.Y. Le & K. Zhang 2002)
- ▶ Variabilité suivant le type d'ARN, suivant l'organisme
- ▶ Signal insuffisant pour une détection fiable ?

A comparison of RNA folding measures

E. Freyhult, P. Gardner & V. Moulton, BMC Bioinformatics, octobre 2005

Prédiction de gènes pour une famille de séquences

- ▶ Est-ce que les séquences font partie d'une même famille d'ARN non-codants ?
- ▶ Schéma général
 - ▶ recherche d'une **structure commune** conservée informations thermo-dynamiques et évolutives
 - ▶ critères pour évaluer la **significativité de la structure**
- ▶ Deux exemples:
 - ▶ QRNA (Rivas et al. - 2001)
 - ▶ RNAz (Hofacker et al. - 2004)

QRNA

G G U C A G A A A G U A C U U
| | | | | | | | | |
G G A C A G A A G G U U C U C

U U G U U C G A A A G A A C G
| | | | | | | | | |
U U G A C C G A A A G G U C G

QRNA

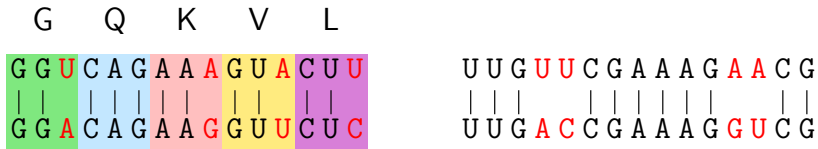
GGUCAGAAAGUACUU
| | | | | | | | | |
GGACAGAAAGGUUCUC

UUGUUCGAAAGAACG
| | | | | | | | | |
UUGACCAGAAAGGUCG

Noncoding RNA genes detection using comparative sequence analysis

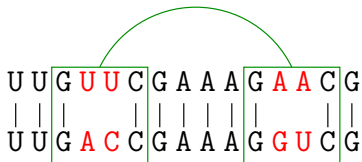
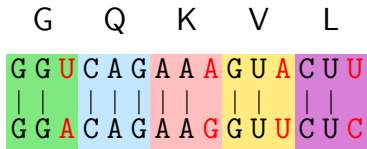
E. Rivas and S. Eddy - BMC Bioinformatics 2001, 2:8

QRNA



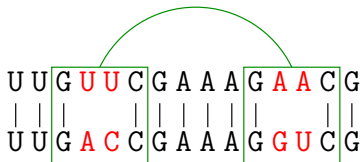
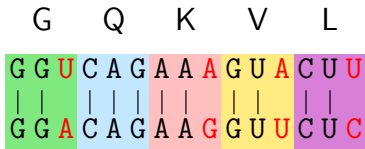
- ▶ Modèle codant
favorise les mutations silencieuses pour le code génétique

QRNA



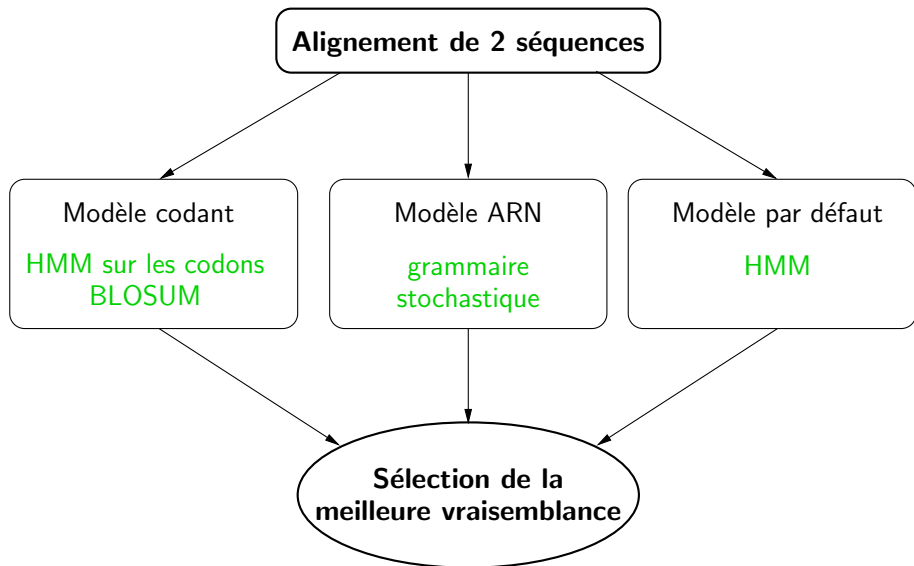
- ▶ Modèle codant
favorise les mutations silencieuses pour le code génétique
- ▶ Modèle ARN
favorise les mutations compensatoires

QRNA

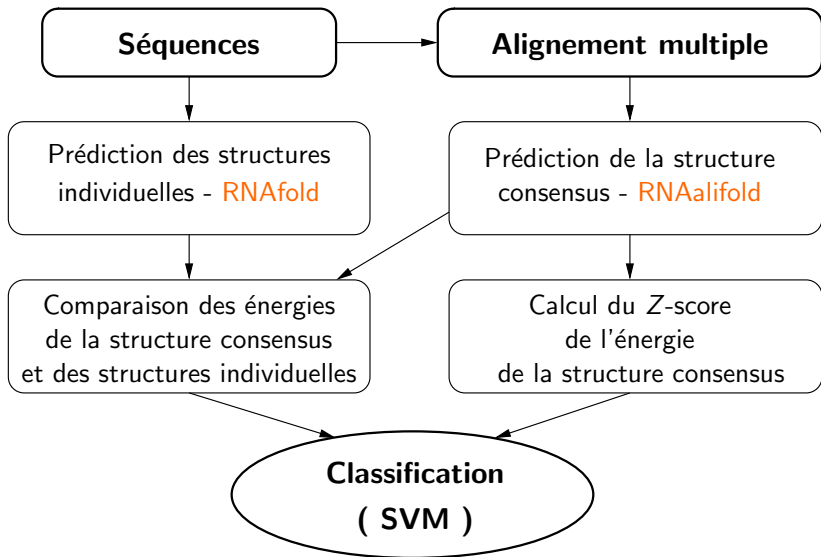


- ▶ Modèle codant
favorise les mutations silencieuses pour le code génétique
- ▶ Modèle ARN
favorise les mutations compensatoires
- ▶ Modèle par défaut

QRNA



RNAz



Fast and reliable prediction of noncoding RNAs

Washietl S., Hofacker I.L., Stadler P.F. Proc. Natl. Acad. Sci. U.S.A. 102, 2454-2459, 2005

Évaluation des performances

Stage de Master Recherche de A. Fontaine

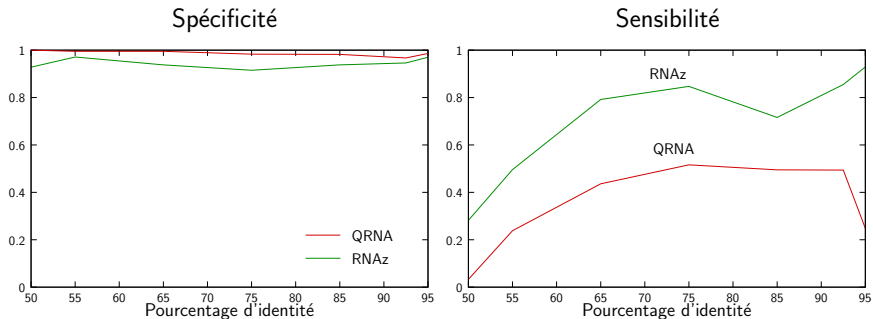
- ▶ **Propriétés évaluées**
 - ▶ influence de la conservation entre les séquences
 - ▶ influence de la méthode d'alignement: Blast, Needle, ClustalW, Dialign2, T-coffee
 - ▶ influence du nombre de séquences: 2, 3, 5 et 10

- ▶ **Evaluation selon trois critères**
 - ▶ Sensibilité: proportion de gènes d'ARN bien classés
 - ▶ Spécificité: proportion de données négatives bien classées
 - ▶ Coefficient de corrélation de Matthews

- ▶ **Résultats disponibles à <http://bioinfo.lifl.fr/rna>**

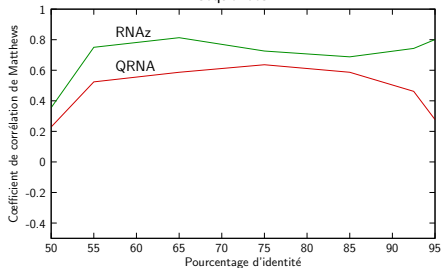
Jeux de données

- ▶ **Positifs:** 21 familles d'ARN non-codants
 - ▶ sources: RFAM, MicroRNA Registry
 - ▶ 10 à 13 séquences par famille
 - ▶ longueur moyenne entre 20 et 370nt
 - ▶ identité moyenne entre 44 et 97%
- ▶ **Négatifs:** 15 familles d'ARN messagers (sans intron)
 - ▶ sources: HomoloGene, TIGR, Ensembl
 - ▶ 4 à 13 séquences par famille
 - ▶ longueur moyenne entre 63 et 300nt
 - ▶ identité moyenne entre 40 et 90%
- ▶ **Négatifs:** “shuffles” des alignements positifs
 - ▶ permutations de positions qui détruisent la structure mais préservent l'identité locale et globale

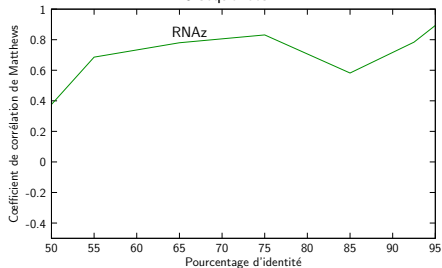


- ▶ **Avantage à Clustalw** parmi toutes les méthodes d'alignement
- ▶ **Bonne spécificité (> 90%)**: prédictions positives fiables
- ▶ **Sensibilité variable**: prédictions négatives moins fiables

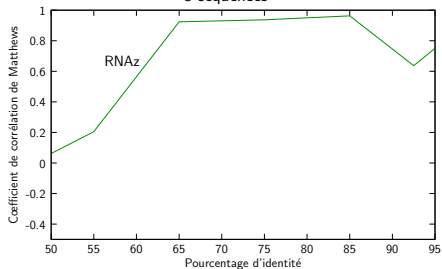
2 séquences



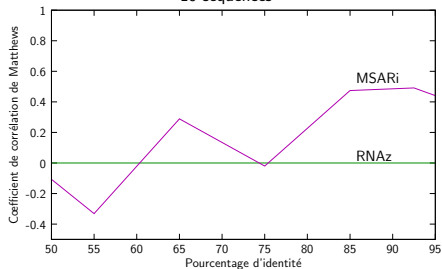
3 séquences



5 séquences

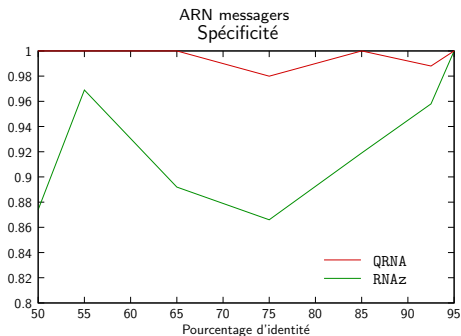
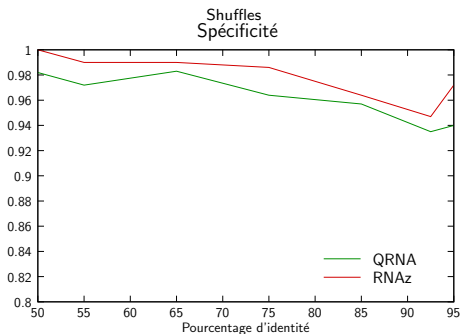


10 séquences



- ▶ Dégradation des résultats avec le nombre de séquences: conséquence de la difficulté à produire un alignement correct
- ▶ Msari sait traiter plus de 10 séquences, car l'alignement est corrigé

Spécificité (ARN messagers vs shuffles)



- ▶ **Excellente spécificité de QRNA sur les ARN messagers: avantage du modèle codant**

Annotation à grande échelle

- ▶ **Cible:** génome humain
- ▶ **Régions conservées entre vertébrés :** 4,81 %
UCSC Genome Browser + PhastCons
- ▶ **Analyse exhaustive avec RNAz :** 30 000 ARN non-codants potentiels

	PhastCons	RNAz
miRNA	▷ 75%	▷ 97% 73%
H/ACA box	▷ 75%	▷ 86 % 65%
C/D box	▷ 50%	▷ 40% 20%

Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome

Washietl S, Hofacker IL, Lukasser M, Huttenhofer A, Stadler PF. Nat Biotechnol. 2005:1383-90.