

Grand prix Nobel pour une petite molécule

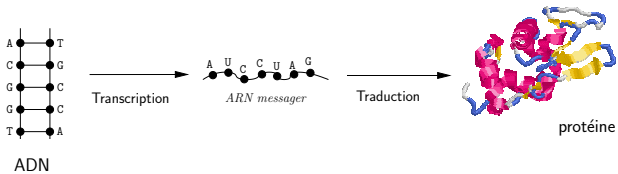
Libération - 2 octobre 2006



Le prestigieux Nobel de Médecine récompense deux chercheurs américains, Andrew Fire et Craig Mello, spécialistes de biologie moléculaire. Ils ont découvert en 1998 l'*ARN interférent*, une molécule fondamentale de la vie

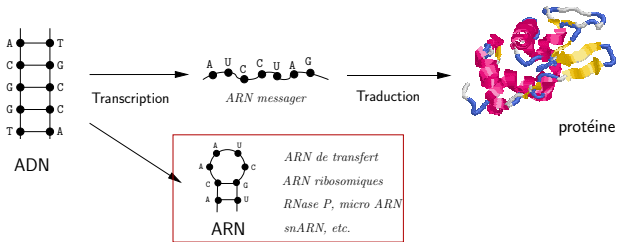
Cette découverte publiée le 19 février 1998 dans Nature a résolu de nombreuses questions que se posaient les scientifiques depuis une décennie. En travaillant sur le ver *Caenorhabditis elegans*, Mello et Fire avaient indirectement répondu aux chercheurs qui s'arrachaient les cheveux sur la couleur des pétunias: en introduisant deux gènes de couleur rouge, ils perdaient toute coloration: le mécanisme d'interférence neutralise l'expression du gène qui produit la couleur du pétunia. Depuis six ans, il ne se passe pas d'année sans que l'on découvre de nouvelles formes de cette petite molécule. En 2002, l'ARN interférent avait été sacré molécule de l'année par la revue Science. Cet été là, le généticien Axel Kahn avait expliqué à Libération les subtilités de la petite molécule, *le plus ancestral moyen de lutte contre les agents infectieux, un des mécanismes fondamentaux de la vie.*

Les gènes à ARN



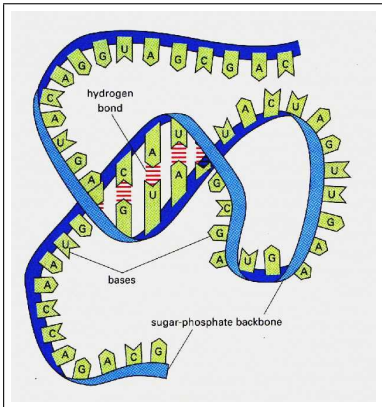
- **Gènes codant pour des protéines**

Les gènes à ARN



- ▶ **Gènes** codant pour des protéines
- ▶ **Gènes à ARN** : synthèse protéique, activité catalytique

Structure des ARN



Séquence: mot sur $\{A, U, C, G\}$
orienté de 5' en 3'

Structure: formation de liaisons
hydrogènes entre deux nucléotides

Watson-Crick : A-U, C-G

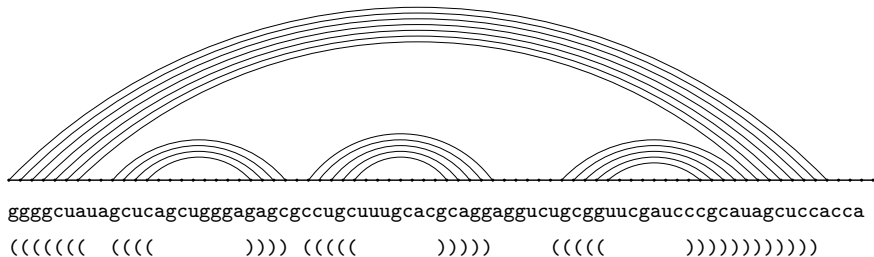
faible : G-U

U-C, G-A , ...

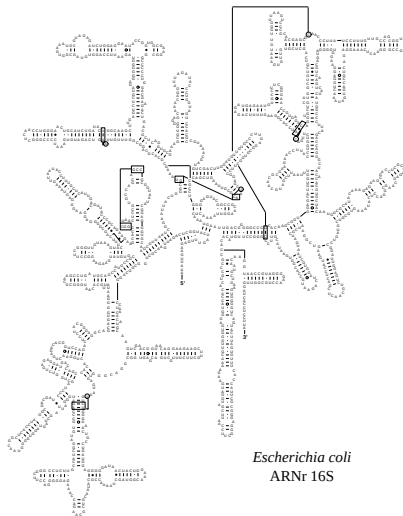
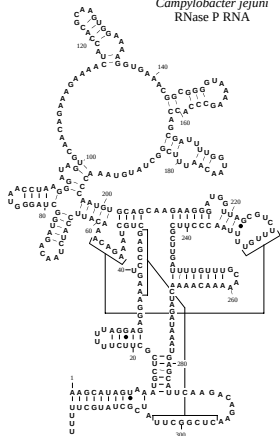
Pas de croisement entre les appariements

**La structure est fonctionnellement
importante**

Exemple : ARN de transfert (Alanine – E. coli)



Campylobacter jejuni
RNase P RNA



Escherichia coli
ARNr 16S

Comment déterminer la structure d'une molécule ?

- ▶ **Structure primaire:** séquençage
- ▶ **Structure secondaire et tertiaire**
 - ▶ Expérimentalement : cristallographie par diffraction à rayons X, résonance magnétique nucléaire (RMN)
Long, difficile et coûteux
 - ▶ Par bio-informatique: algorithmes de prédiction de structures secondaires
 - Méthode 1 : approche thermodynamique*
 - Méthode 2 : approche comparative*

Approche thermodynamique

- ▶ **Trois hypothèses:**
 - ▶ À chaque configuration de la molécule correspond une quantité d'énergie libre.
 - ▶ La configuration la plus stable est celle qui minimise l'énergie libre.
 - ▶ La molécule, en se repliant, adopte la configuration la plus stable.
- ▶ **On s'est ramené à un problème combinatoire : trouver la structure dont l'énergie est optimale.**

Modèle initial (Nussinov - 1978)

- ▶ L'énergie de la molécule est la somme des énergies de chaque paire de bases.
- ▶ $\alpha(r_i, r_j)$: énergie libre de l'appariement (r_i, r_j)

$$\alpha(r_i, r_j) < 0 \quad \text{si } j - i > 3 \text{ et } r_i \leftrightarrow r_j$$

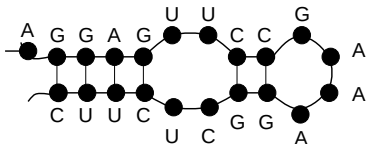
$$\alpha(r_i, r_j) = 0 \quad \text{si } i = j$$

$$\alpha(r_i, r_j) = +\infty \quad \text{sinon}$$

- ▶ Énergie libre de la structure secondaire $\mathcal{S}$

$$E(\mathcal{S}) = \sum_{(r_i, r_j) \in \mathcal{S}} \alpha(r_i, r_j)$$

► **Exemple**



Fonction d'énergie :

$$\alpha(A, U) = -2$$

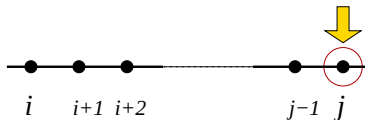
$$\alpha(C, G) = -3$$

$$\alpha(G, U) = -1$$

Energie totale: -15

- **Comment calculer la structure optimale ?** par programmation dynamique

Etape 1: décomposition du problème en instances plus petites



- ▶ $\mathcal{S}_{i,j}$: structure secondaire optimale pour la sous-séquence $i \dots j$
- ▶ Trois possibilités pour r_j :
 1. r_j ne s'apparie pas avec la sous-chaîne $r_i \dots r_{j-1}$: $E(\mathcal{S}_{i,j}) = E(\mathcal{S}_{i,j-1})$
 2. r_j s'apparie avec r_i : $E(\mathcal{S}_{i,j}) = E(\mathcal{S}_{i+1,j-1}) + \alpha(r_i, r_j)$
 3. r_j s'apparie avec r_k pour un $i < k < j$:

$$E(\mathcal{S}_{i,j}) = \min\{E(\mathcal{S}_{i,k-1}) + \alpha(r_k, r_j) + E(\mathcal{S}_{k+1,j-1}), k \in]i, j[\}$$

- ▶ $E(\mathcal{S}_{i,j}) = \min$ cas **1**, **2** et **3**

Etape 2 : construction de la table de programmation dynamique

- Une table T , de dimension 2: $T(i, j) := E(\mathcal{S}_{i,j})$

$$T(i, j) = \min \begin{cases} T(i, j-1) \\ T(i+1, j-1) + \alpha(r_i, r_j) \\ \min\{T(i, k-1) + \alpha(r_k, r_j) + T(k+1, j-1)\} \end{cases}$$

- Une table S qui stocke le devenir de r_j

Etape 3: construction de la structure secondaire optimale, par retour arrière

Complexité

- Chaque case de la table nécessite $O(n)$ calculs.
- Complexité globale en $O(n^3)$.

[illegible]

[illegible]

[illegible]

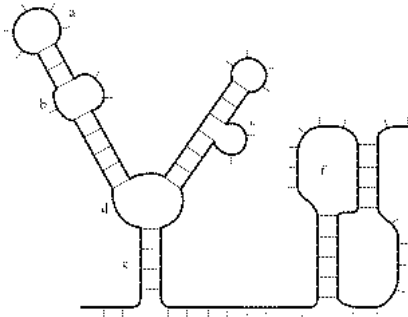
Mfold

Zuker, Turner *et al.*

► Paramètres d'énergie

- liaison hydrogène
- énergie d'empilement

► Motifs structuraux



- a. épingle à cheveux (*hairpin*)
- b. boucle interne (*internal loop*)
- c. renflement (*bulge loop*)
- d. jonction
- e. tige (*duplex*)
- f. **Exclus** : pseudo-nœud (*pseudoknot*)

Limites de l'approche thermodynamique

- ▶ Pertinence de la définition de la fonction d'énergie.

Solution : *transformer l'algorithme pour obtenir un ensemble de configurations sous-optimales*

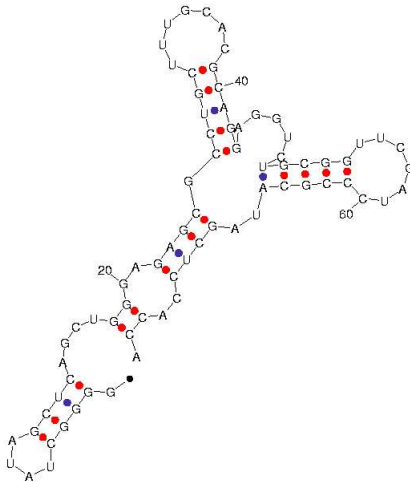
- ▶ Les hypothèses biologiques ne sont pas toutes valides :

Il existe des molécules d'ARN qui se replient en formant des nœuds, ou dans lesquelles un appariement regroupent 3 nucléotides.

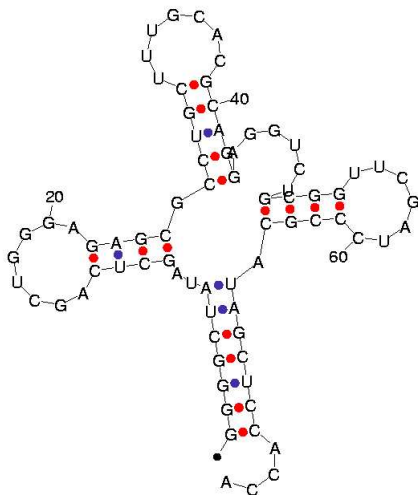
Solution : *complexifier les algorithmes*

- ▶ Pas de prise en compte d'interactions avec des molécules voisines.
- ▶ Pas de prise en compte du sens de la synthèse de l'ARN.

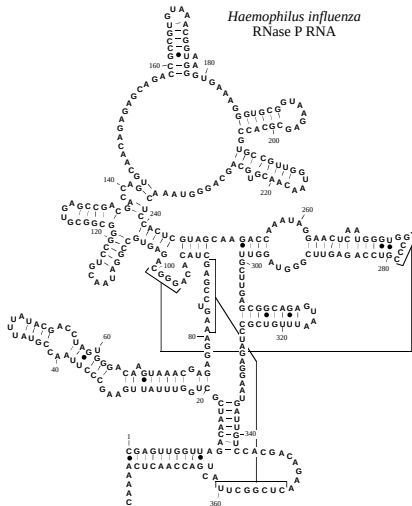
Exemple : ARN de transfert



dG = -27.57 [initially -29.7] tRNA_coli



dG = -29.04 [initially -29.6] tRNA_colI

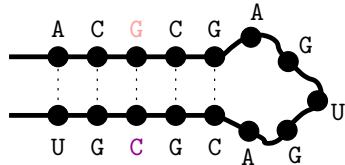
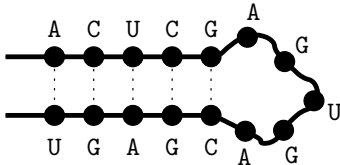


sous-unité ARN de RNase P
(*Hemophilus influenzae*)

Mfold :
21 structures sous-optimales

Approche comparative

- ▶ On dispose d'une famille de séquences homologues
- ▶ Lors de l'évolution, la structure est mieux préservée que la séquence
- ▶ Les séquences homologues partagent la même structure
- ▶ **Phénomène de changement de base compensatoire:**
Quand une base impliquée dans un appariement mute, la base complémentaire mute également, pour préserver la structure



► **Jeu de séquences**

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

► **Jeu de séquences**

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

► **Etape 1 : construction d'un alignement multiple**

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

► **Jeu de séquences**

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

► **Etape 1 : construction d'un alignement multiple**

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

► **Etape 2 : détection des positions corrélées**

► **Jeu de séquences**

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

► **Etape 1 : construction d'un alignement multiple**

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

► **Etape 2 : détection des positions corrélées**

► **Jeu de séquences**

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

► **Etape 1 : construction d'un alignement multiple**

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

► **Etape 2 : détection des positions corrélées**

Comment mesurer la corrélation entre deux colonnes ?

$\mathcal{I}(i, j)$: **information mutuelle** des colonnes i et j de l'alignement multiple

$$\mathcal{I}(i, j) = \sum_{x, y=A, C, G, T} f_{x, y}^{i, j} \log_2 \frac{f_{x, y}^{i, j}}{f_x^i f_y^j}$$

- ▶ 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

$\mathcal{I}(i, j)$ varie entre 0 et 2 bits.

Quantité d'information révélée par la colonne j , la colonne i étant connue. $\mathcal{I}(i, j)$ est maximale quand i et j individuellement paraissent aléatoires ($f_x^i = f_y^j = 0.25$) et que i et j sont parfaitement corrélées.

Exemples

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

$$\mathcal{I} = \frac{\log_2(4/3)}{2} + \frac{\log_2(4/3)}{4} + \frac{\log_2(4)}{4}$$

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

$$\mathcal{I} = \frac{\log_2(4)}{4} + \frac{\log_2(4)}{4} + \frac{\log_2(4)}{4} + \frac{\log_2(4)}{4} = 2$$

La corrélation entre les deux colonnes est maximale.

GGGGAAATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCAGCGGTTTCGATCCCGCTATTCTCCA---
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTTCGATCCCGCATAGCTCCACCA
GGGGCTATAGCTCAGCT-GGGAGAGCGCTTGCATGGCATGCAAGAGGTTCAGCGGTTTCGATCCCGCTTAGCTCCACCA
GGGGAAATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCAGCGGTTTCGATCCCGCTATTCTCCA---
GGGGCCTTAGCTCAGTC-GGTAGAGCACTGCCTTTGCAAGGCAGATGTCAGGGGTTTCGATTCCCCTAGGCTCCA---
GGGGGTATAGCTCAGTT-GGTAGAGCGCTGCCTTTGCAAGGCAGAAGTCAGCGGTTTCGA.TCCGCTTACCCCCA---
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTTCGATCCCGCATAGCTCCACCA
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTTCGATCCCGCATAGCTCCACCA
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTTCGATCCCGTCTGCTCCACCA
GGGGCCATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTTCAGGAGTTTCGATCCTCCTTGCTCCACCA
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTTCGATCCCGTCTGCTCCACCA
GGGGGCATAGCTCAGCTGGGGAGAGCGCCTGCCTTGCACGCAGGAGGTCAACGGTTTCGATCCCGTTTGGCTCCA---
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTTCGATCCCGTCTGCTCCACCA
GGGGCATTAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTTCAGCGGTTTCGATCCCGCTATTCTCCACCA
GGGGCCATAGCTCAGTT-GGTAGAGCGCCTGCTTTGCAAGCAGGTGT-CGTCGGTTTCGAATCCGTCTGGCTCCACCA
GGGGCCGTAGCTCAGCTGGG-AGAGCACCTGCTTTGCAAGCAGGGGGTCGGAGGTTTCGATCCCGTCCGGCTCCACCA
GGGGCCGTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCGTCGGTTTCGATCCCGTCCGGCTCCACCA
GGGGCCGTAGCTCAGCT-GG-AGAGCACCTGCTTTGCAAGCAGGGGGTCGTCGGTTTCGATCCCGTCCGGCTCCACCA

ARN de transfert

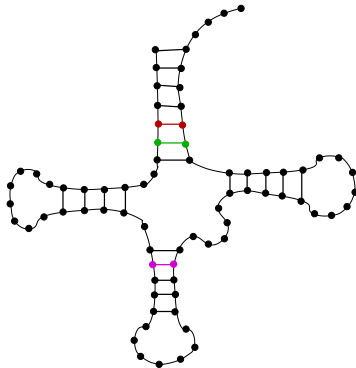
GGGGAAATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCAGCGGTTTCGATCCCGCTATTCTCTCCA---
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTTCGATCCCGCATAGCTCCACCA
GGGGCTATAGCTCAGCT-GGGAGAGCGCTTGCATGGCATGCAAGAGGTTCAGCGGTTTCGATCCCGCTTAGCTCCACCA
GGGGAAATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCAGCGGTTTCGATCCCGCTATTCTCTCCA---
GGGGCCTTAGCTCAGTC-GGTAGAGCACTGCCTTTGCAAGGCAGATGTCAGGGGTTTCGATTCCCCTAGGCTCCA---
GGGGGTATAGCTCAGTT-GGTAGAGCGCTGCCTTTGCAAGGCAGAAAGTCAGCGGTTTCGA . TCCGCTTACCCCCA---
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTTCGATCCCGCATAGCTCCACCA
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTTCGATCCCGCATAGCTCCACCA
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTTCGATCCCGTCTGCCTCCACCA
GGGGCCATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTTCAGGAGTTTCGATCCTCCTTGGCTCCACCA
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTTCGATCCCGTCTGCCTCCACCA
GGGGCCATAGCTCAGCTGGGGAGAGCGCCTGCCTTGCACGCAGGAGGTCAACGGTTTCGATCCCGTTTTGGCTCCA---
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTTCGATCCCGTCTGCCTCCACCA
GGGGCAATTAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTTCAGCGGTTTCGATCCCGCTATTCTCTCCACCA
GGGGCCATAGCTCAGTT-GGTAGAGCGCCTGCTTTGCAAGCAGGTGT-CGTCGGTTTCGAATCCGTCTGGCTCCACCA
GGGGCGTAGCTCAGCTGGG-AGAGCACCTGCTTTGCAAGCAGGGGGTCGGAGGTTTCGATCCCGTCCGGCTCCACCA
GGGGCGTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCGTCGGTTTCGATCCCGTCCGGCTCCACCA
GGGGCGTAGCTCAGCT-GG-AGAGCACCTGCTTTGCAAGCAGGGGGTCGTCGGTTTCGATCCCGTCCGGCTCCACCA

ARN de transfert

GGGGAATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCAGCGGTTTCGATCCCGCTATTCTCTCCA---
 GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTTCGATCCCGCATAGCTCCACCA
 GGGGCTATAGCTCAGCT-GGGAGAGCGCTTGCATGGCATGCAAGAGGTTCAGCGGTTTCGATCCCGCTTAGCTCCACCA
 GGGGAATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCAGCGGTTTCGATCCCGCTATTCTCTCCA---
 GGGGCCCTTAGCTCAGTC-GGTAGAGCACTGCCTTTGCAAGGCAGATGTCAGGGGTTTCGATTCCCCTAGGCTCCA---
 GGGGCTATAGCTCAGTT-GGTAGAGCGCTGCCTTTGCAAGGCAGAAAGTCAGCGGTTTCGA.TCCGCTTACCCCCA---
 GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTTCGATCCCGCATAGCTCCACCA
 GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTTCGATCCCGCATAGCTCCACCA
 GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTTCGGTTTCGATCCCGTCTGCCTCCACCA
 GGGGCCATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTTCAGGAGTTTCGATCCTCCTTGGCTCCACCA
 GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTTCGGTTTCGATCCCGTCTGCCTCCACCA
 GGGGGCATAGCTCAGCTGGGGAGAGCGCCTGCTTTGCACGCAGGAGGTCAACGGTTTCGATCCCGTTTTGGCTCCA---
 GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTTCGGTTTCGATCCCGTCTGCCTCCACCA
 GGGGCATTAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTTCAGCGGTTTCGATCCCGCTATTCTCTCCACCA
 GGGGCCATAGCTCAGTT-GGTAGAGCGCCTGCTTTGCAAGCAGGTGT-CGTTCGGTTTCGAATCCGTCTGGCTCCACCA
 GGGGCCGTAGCTCAGCTGGG-AGAGCACCTGCTTTGCAAGCAGGGGGTCGGAGGTTTCGATCCCGTCCGGCTCCACCA
 GGGGCCGTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTCGTTCGGTTTCGATCCCGTCCGGCTCCACCA
 GGGGCCGTAGCTCAGCT-GG-AGAGCACCTGCTTTGCAAGCAGGGGGTCGTTCGGTTTCGATCCCGTCCGGCTCCACCA

ARN de transfert

Vérification ...



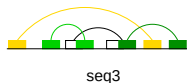
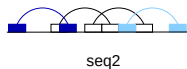
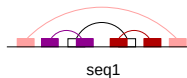
Structure secondaire de l'ARNt

Méthodes hybrides

Exemple : caRNAC (Perriquet Touzet - 2003)

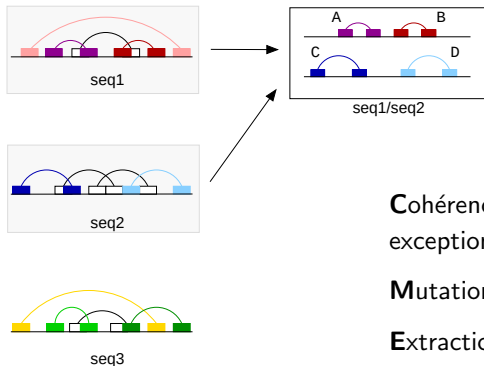
- ▶ **Trois sources d'information :**
 - ▶ énergie libre
 - ▶ mutations compensatoires
 - ▶ conservation éventuelle des séquences
- ▶ **Supporter des séquences et des structures divergentes**
Pas d'alignement préalable des séquences
- ▶ **Permettre la gestion de longs jeux de données**
- ▶ <http://bioinfo.lifl.fr/carnac>

caRNAc : stratégie de prédiction



1. Identification des tiges potentielles, séquence par séquence

caRNAc : stratégie de prédiction



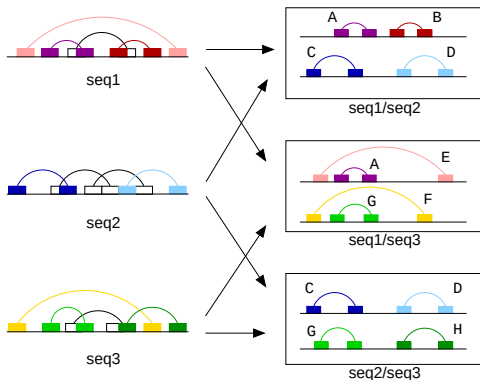
Cohérence avec les motifs
exceptionnellement conservés

Mutations compensatoires

Extraction de la structure commune
par programmation dynamique

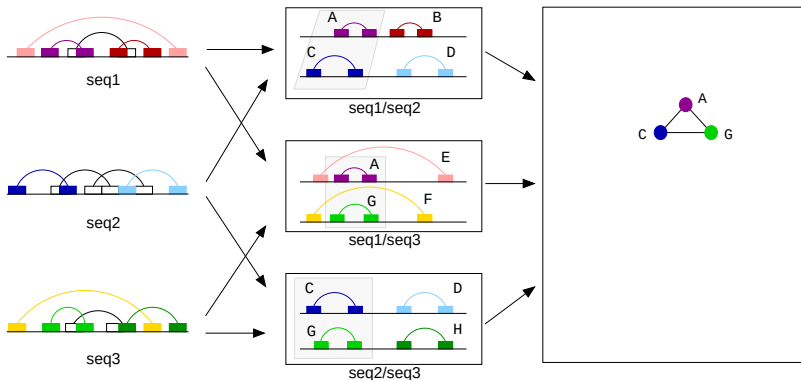
2. Confrontation deux à deux

caRNAc : stratégie de prédiction



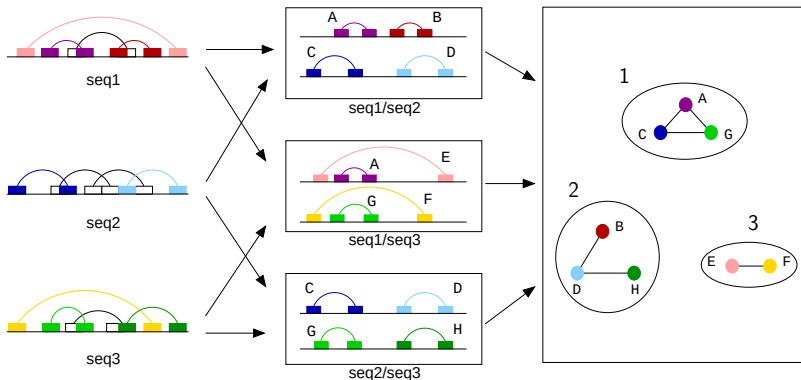
3. Combinaison des prédictions 2 à 2 : graphe des tiges

caRNAc : stratégie de prédiction



3. Combinaison des prédictions 2 à 2 : graphe des tiges

caRNAc : stratégie de prédiction

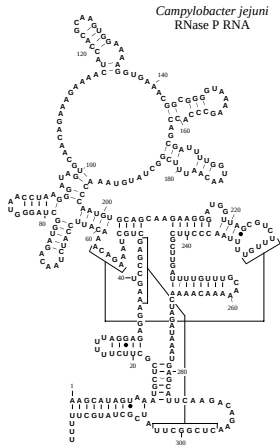


4. Sélection finale des tiges

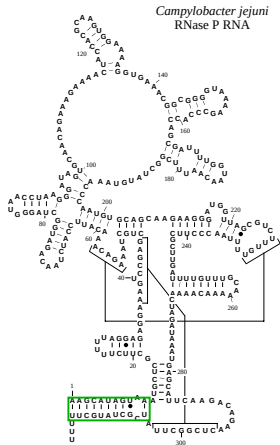
Comparaison de molécules d'ARN

- ▶ **Comparaison des séquences** : algorithmes sur les mots
Alignement, recherche de motifs, etc.
- ▶ **Comparaison des structures secondaires** : algorithmes sur les arbres
Alignement, recherche de motifs, etc.

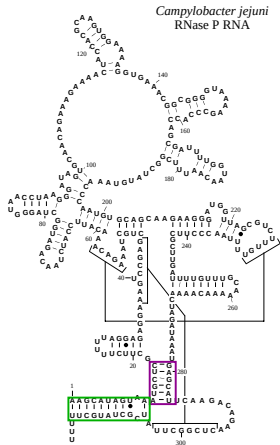
L'ARN vu comme un arbre



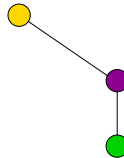
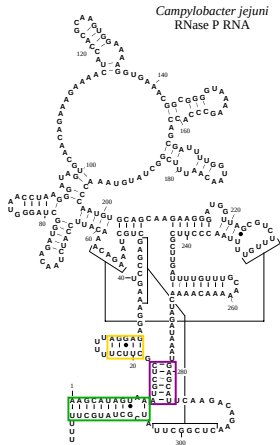
L'ARN vu comme un arbre



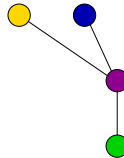
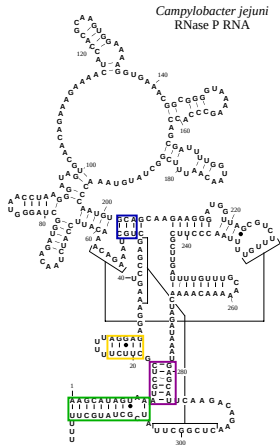
L'ARN vu comme un arbre



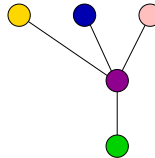
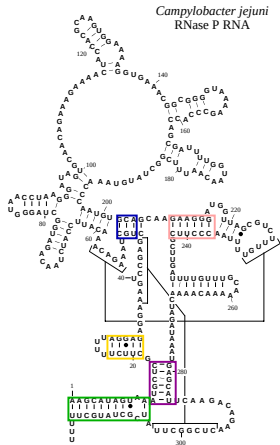
L'ARN vu comme un arbre



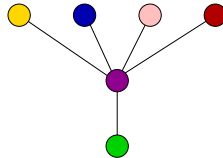
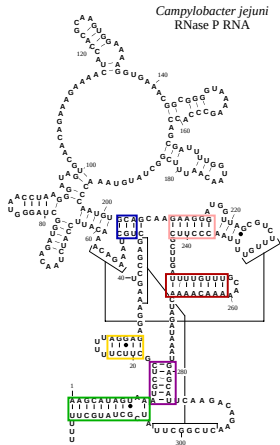
L'ARN vu comme un arbre



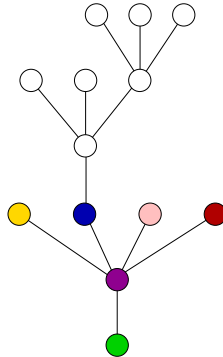
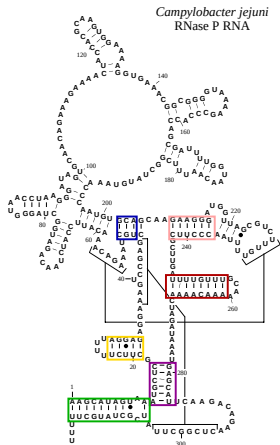
L'ARN vu comme un arbre

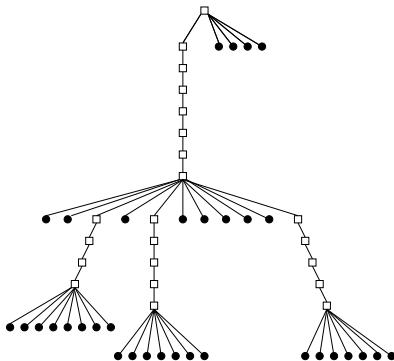
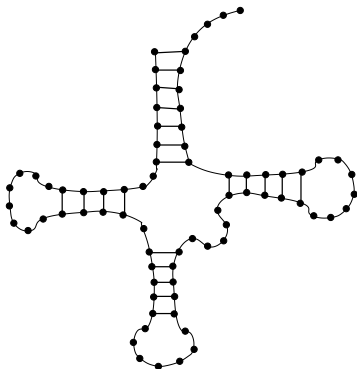


L'ARN vu comme un arbre



L'ARN vu comme un arbre





Représentation arborescente pour la structure secondaire de l'ARN de transfert

□ : 2 bases appariées (hélice)

● : base libre (boucle)