

# Structures des ARN

Hélène TOUZET

`touzet@lifl.fr`

# Structure secondaire de l'ARN

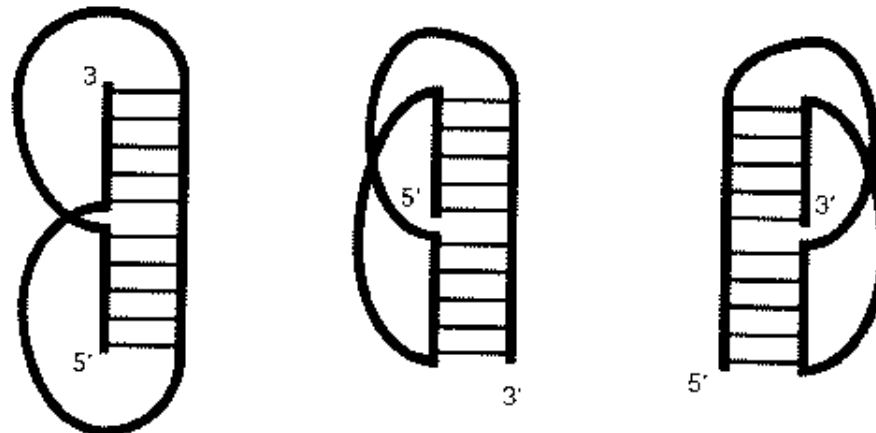
## ▷ Appariements :

*Watson-Crick* : A–U, C–G

*Wobble* : G–U

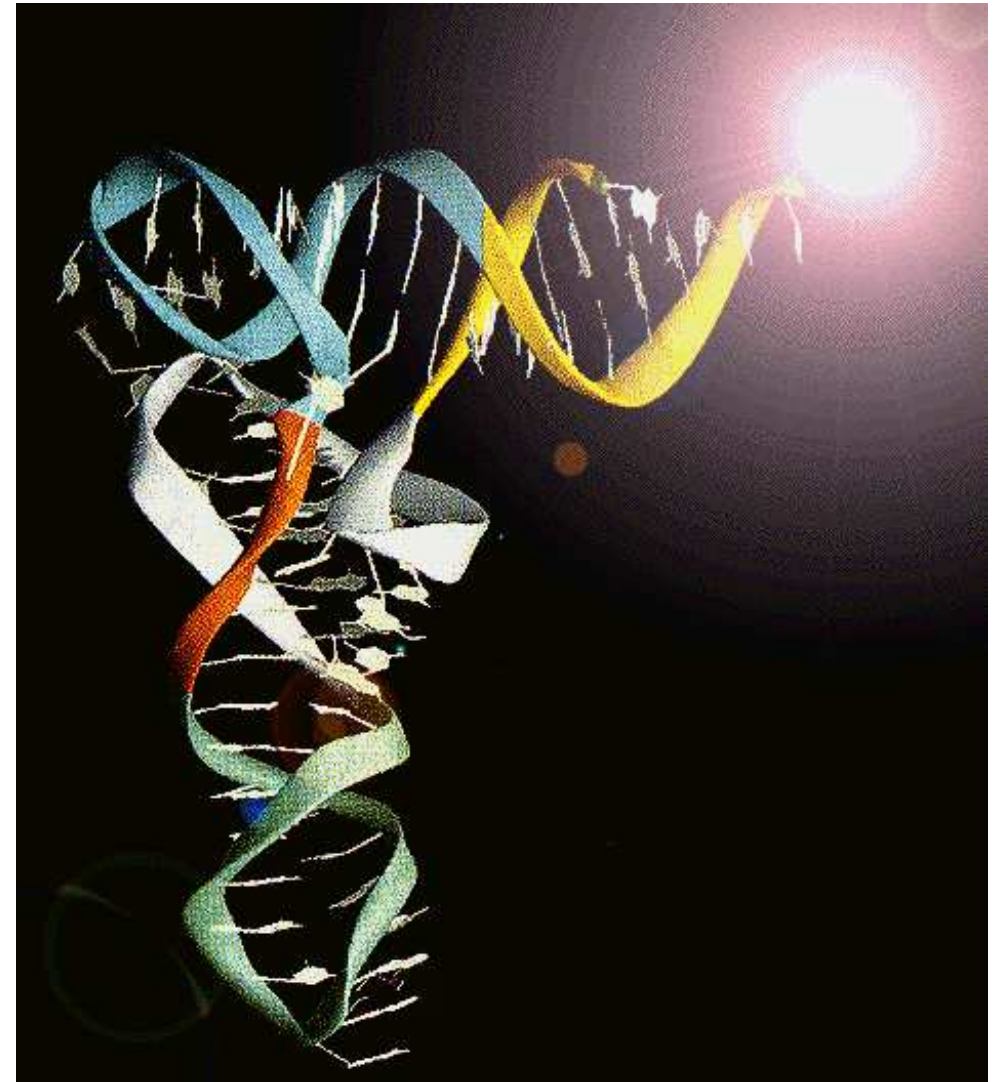
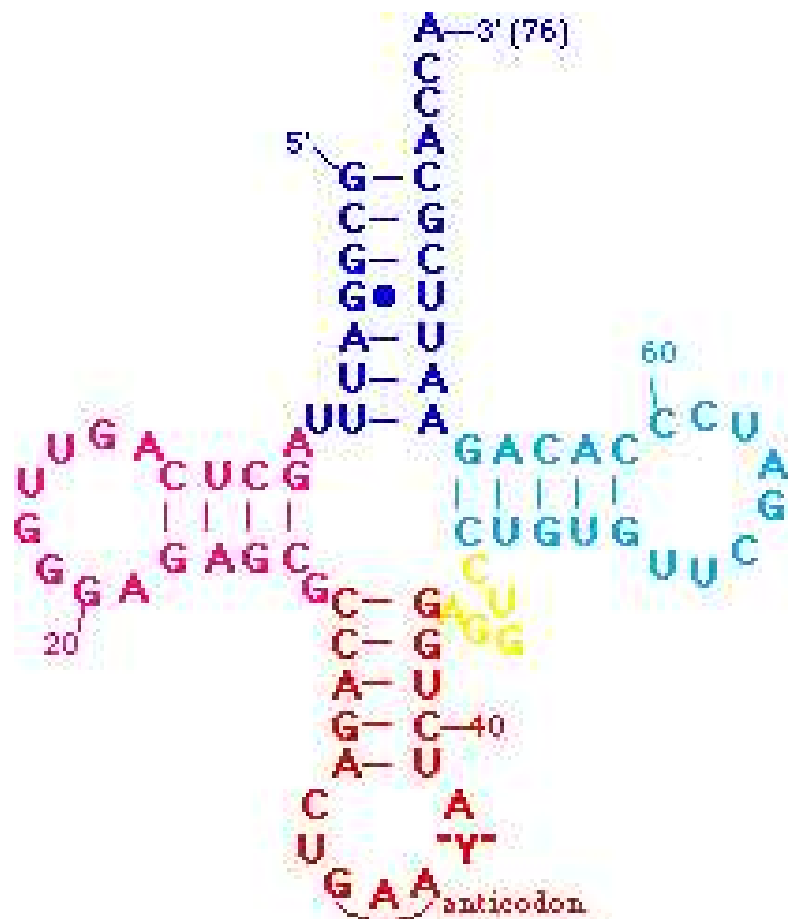
U–C, G – A , . . .

## ▷ Pseudo-nœuds



## ▷ Triplets de bases





# Comment déterminer la structure d'une molécule ?

## ▷ structure primaire

*séquençage*

## ▷ structure secondaire et tertiaire

- de manière exacte, expérimentalement

*Cristallographie par diffraction à rayons X*

*Résonance magnétique nucléaire (RMN)*

 Long, difficile et coûteux

- par extrapolation, à partir de la structure primaire : **Algorithmes de prédiction de structures secondaires**

*Méthode 1* : approche thermodynamique

*Méthode 2* : approche comparative

# Approche thermodynamique

1. À chaque configuration de la molécule correspond une quantité d'énergie libre.
  2. La configuration la plus stable est celle qui minimise l'énergie libre.
  3. 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.*

**Simplification du modèle :** *la structure secondaire exclut les pseudo-noeuds et les triplets*

## Modèle de départ

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  $(r_i, r_j)$

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

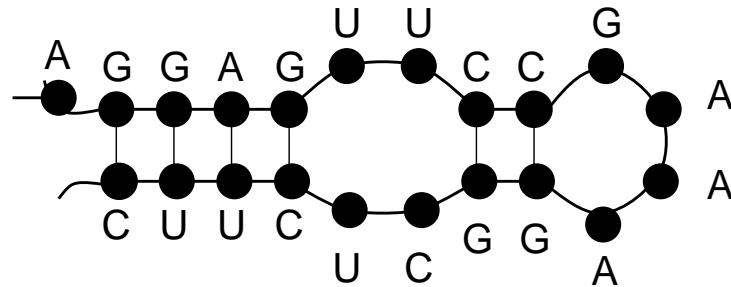
$$\alpha(r_i, r_j) = 0 \quad \text{si } i = j$$

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

▷ 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 :

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

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

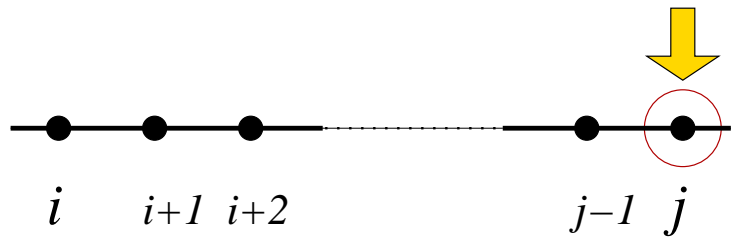
$$\alpha(G, U) = -1$$

*La structure a pour énergie -15*

---

**Algorithme:** 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  $r_i \dots r_{j-1}$  :

$$\mathcal{S}_{i,j} = \mathcal{S}_{i,j-1} \cup \{r_j, r_j\} \Rightarrow E(\mathcal{S}_{i,j}) = E(\mathcal{S}_{i,j-1})$$

2.  $r_j$  s'apparie avec  $r_i$  :

$$\mathcal{S}_{i,j} = \mathcal{S}_{i+1,j-1} \cup \{r_i, r_j\} \Rightarrow E(\mathcal{S}_{i,j}) = E(\mathcal{S}_{i+1,j-1}) + \alpha(r_i, r_j)$$

3.  $r_j$  s'apparie avec  $r_k$  pour un  $i < k < j$  :

$$\mathcal{S}_{i,j} = \mathcal{S}_{i,k-1} \cup \{r_k, r_j\} \cup \mathcal{S}_{k+i,j-1} \Rightarrow$$

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

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

## Etape 2 : construction de la table de programmation dynamique

- Une table  $T$ , de dimension 2:  $T(i, j) := E(S_{i,j})$

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

- Une table  $S$  qui stocke le devenir de  $r_j$

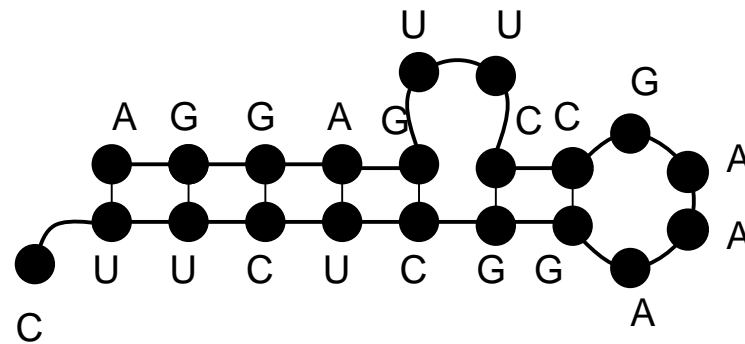
### Complexité

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



**Etape 3:** *construction de la structure secondaire optimale*

On "remonte" la table de programmation dynamique  $S$ .



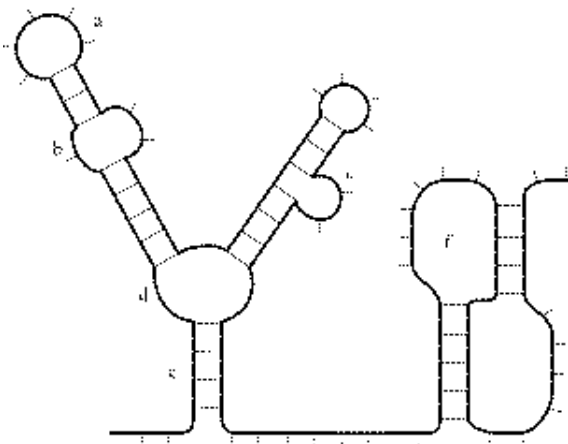
# Mfold

Zuker, Turner *et al.*

▷ Paramètres d'énergie

- liaison hydrogène
- énergie d'empilement

▷ Motifs structuraux



- épinglé à cheveux (*hairpin*)
- boucle interne (*internal loop*)
- renflement (*bulge loop*)
- jonction
- tige (*duplex*)
- Exclus** : pseudo-noeud (*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*

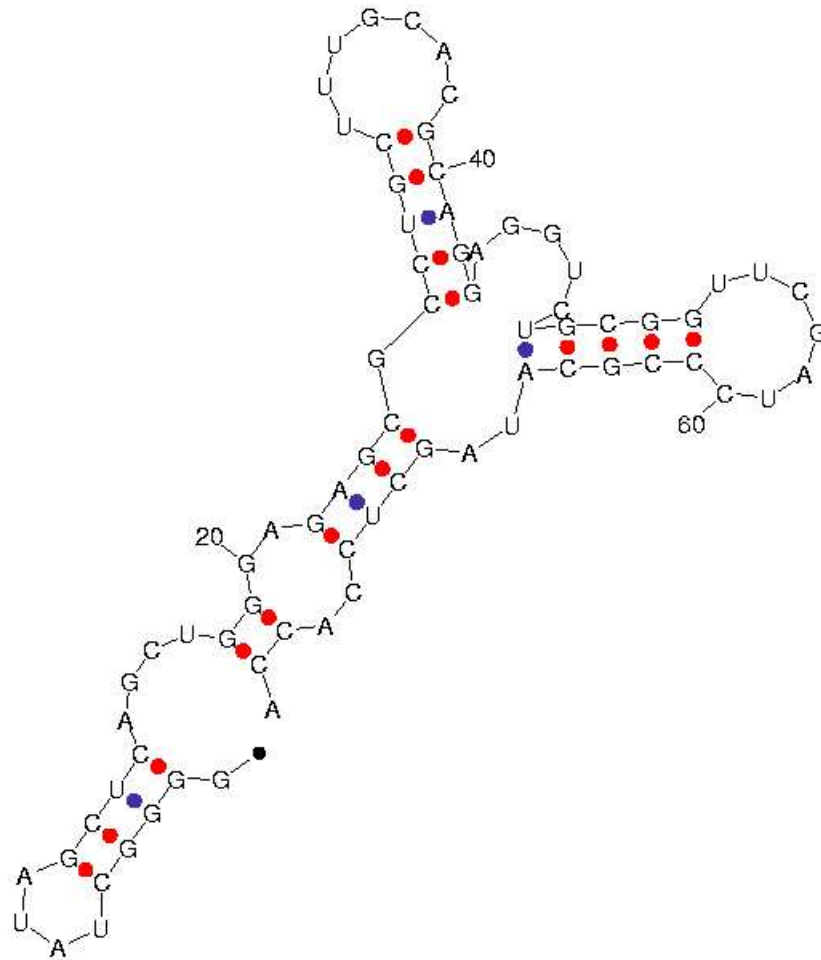
- ▷ Les hypothèses biologiques ne sont pas toutes valides :

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

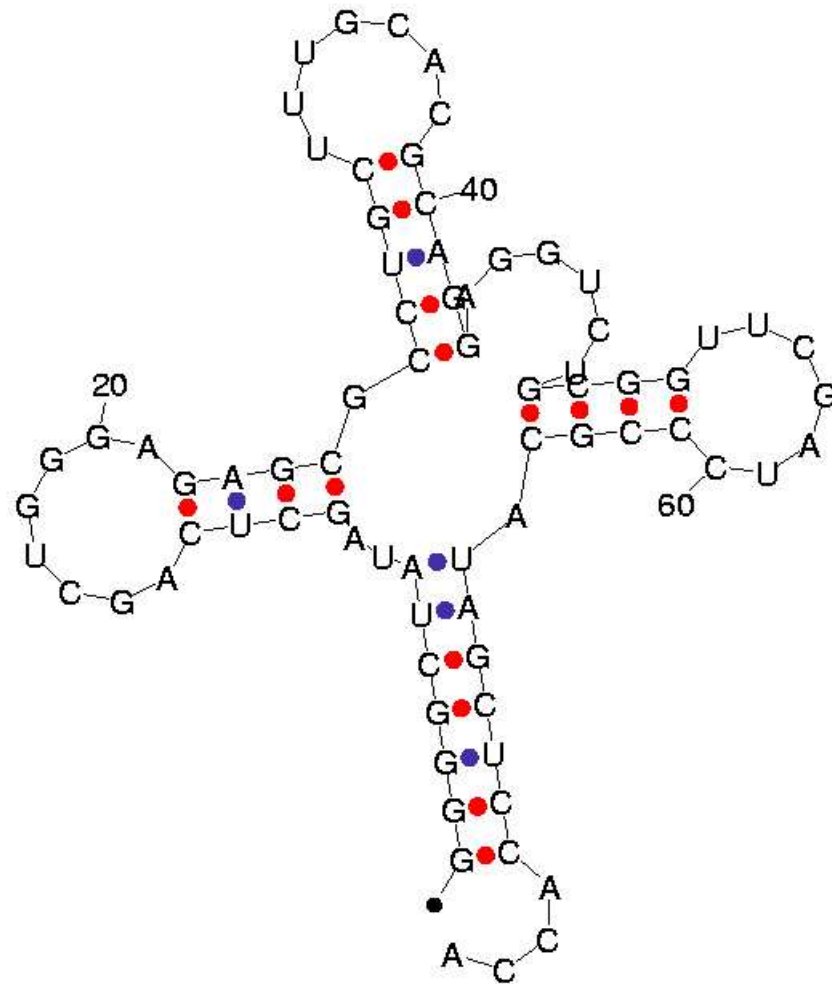
**Solution** : *complexifier les algorithmes*

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

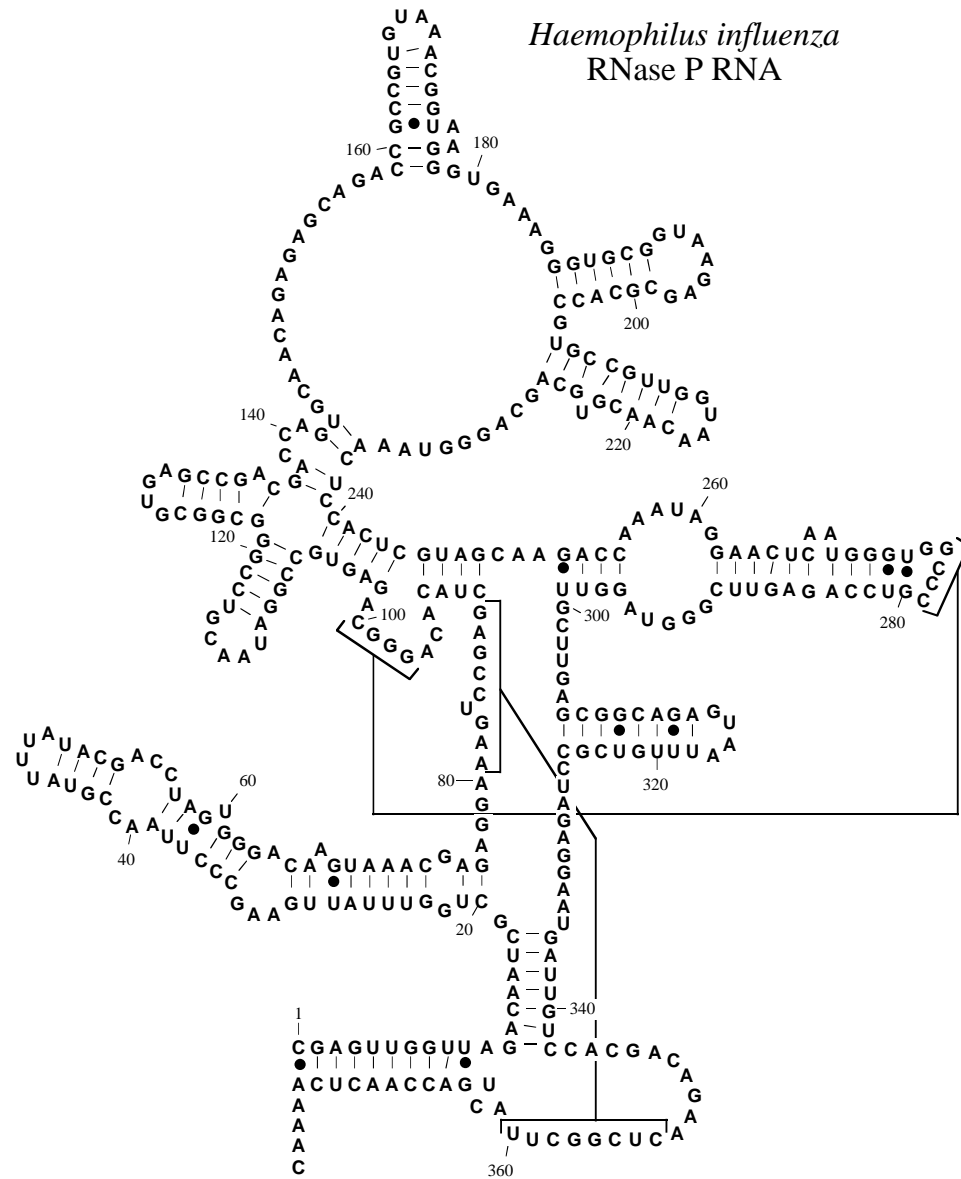
# Exemple : ARN de transfert



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

*Lors de l'évolution, la structure secondaire est mieux préservée que la structure primaire.*



*Des séquences primaires proches, mais différentes, ont la même structure secondaire.*

**Application** Travailler non pas avec une séquence, mais avec deux, ou une famille de séquences.

## Repliement inverse

▷ On cherche une séquence similaire ("homologue") pour laquelle la structure secondaire est connue.

⇒ *succession d'hélices, boucles . . .*

▷ On distribue les motifs le long de la séquence de structure inconnue

- respecter les appariements
- favoriser les insertions et les délétions hors des hélices

▷ **Exemple** : RAGA (C. Notre-Dame, algos génétiques)

## Et si on ne connaît pas la structure

... mais on dispose d'une famille de séquences homologues ?

```
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
```

### Changement de base compensatoire:

*Quand une base impliquée dans un appariement mute, la base complémentaire mute également, pour préserver la paire, et donc, la structure secondaire.*

**Étape 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

**Étape 2 :** *détection des positions corrélées*

présomption d'appariement

**Étape 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

**Étape 2 :** *détection des positions corrélées*

présomption d'appariement

## Comment mesurer la corrélation entre deux colonnes ?

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

$$\mathcal{I}(i, j) = \sum f_{x_i, x_j} \log_2 \frac{f_{x_i, x_j}}{f_{x_i} f_{x_j}}$$

- $f_{x_i}$  fréquence de la base  $x_i$  dans la colonne  $i$
- $f_{x_i} f_{x_j}$  fréquence du couple  $x_i x_j$  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.  $M_{ij}$  est maximale quand  $i$  et  $j$  individuellement paraissent aléatoires ( $f_i = f_j = 0.25$ ) 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.

GGGG**A**ATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTTCAGCGGTTTCGATCCCGCTAT**T**CTCCA---  
GGGG**C**TATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTTCGATCCCGCATA**G**CTCCACCA  
GGGG**C**TATAGCTCAGCT-GGGAGAGCGCTTGCATGGCATGCAAGAGGTCAGCGGTTTCGATCCCGCTTA**G**CTCCACCA  
GGGG**A**ATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTTCAGCGGTTTCGATCCCGCTAT**T**CTCCA---  
GGGG**C**CTTAGCTCAGTC-GGTAGAGCACTGCCTTTGCAAGGCAGATGTCAGGGGTTTCGATTCCCCTAG**G**CTCCA---  
GGGG**G**TATAGCTCAGTT-GGTAGAGCGCTGCCTTTGCAAGGCAGAAGTCAGCGGTTTCGA . TCCGCTTA**C**CCCCA---  
GGGG**C**TATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTTCGATCCCGCATA**G**CTCCACCA  
GGGG**C**TATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTTCGATCCCGCATA**G**CTCCACCA  
GGGG**G**CATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTG**C**CTCCACCA  
GGGG**C**CATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTTCAGGAGTTTCGATCCTCCTTG**G**CTCCACCA  
GGGG**G**CATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTG**C**CTCCACCA  
GGGG**C**CATAGCTCAGCTGGGGAGAGCGCCTGCCTTGCACGCAGGAGGTCAACGGTTCGATCCCGTTTG**G**CTCCA---  
GGGG**G**CATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTG**C**CTCCACCA  
GGGG**C**ATTAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTTCAGCGGTTTCGATCCCGCTAT**T**CTCCACCA  
GGGG**C**CATAGCTCAGTT-GGTAGAGCGCCTGCTTTGCAAGCAGGTGT-CGTCGGTTCGAATCCGTCTG**G**CTCCACCA  
GGGG**C**CGTAGCTCAGCTGGG-AGAGCACCTGCTTTGCAAGCAGGGGGTTCGGAGGTTTCGATCCCGTCCG**G**CTCCACCA  
GGGG**C**CGTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTTCGTCGGTTCGATCCCGTCCG**G**CTCCACCA  
GGGG**C**CGTAGCTCAGCT-GG-AGAGCACCTGCTTTGCAAGCAGGGGGTTCGTCGGTTCGATCCCGTCCG**G**CTCCACCA

ARNt

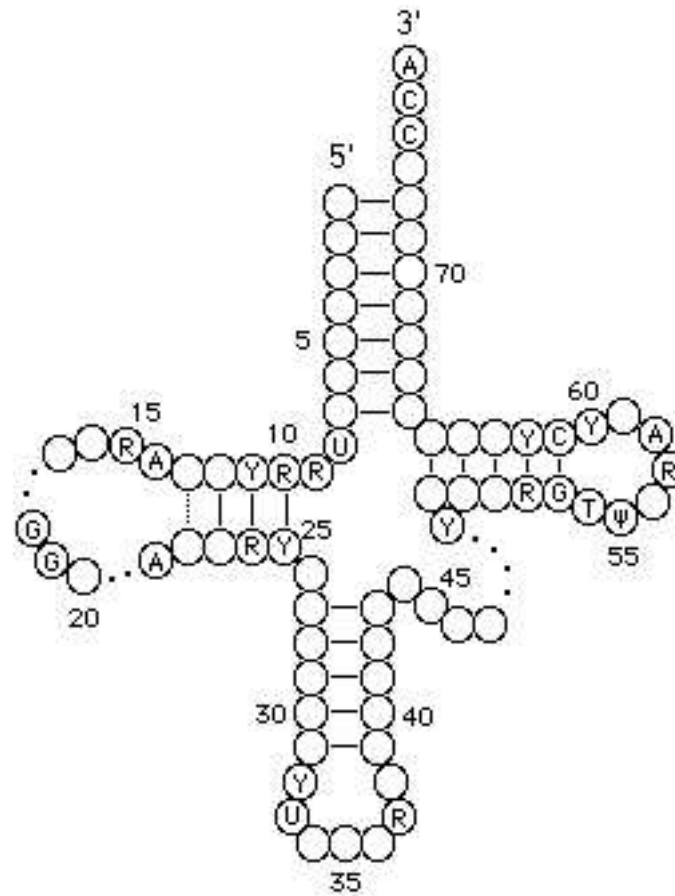
GGGGATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTTCAGCGGTTTCGATCCCGCTATTCTCCA---  
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTTCGATCCCGCATAGCTCCACCA  
GGGGCTATAGCTCAGCT-GGGAGAGCGCTTGCATGGCATGCAAGAGGTCAGCGGTTTCGATCCCGCTTAGCTCCACCA  
GGGGATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTTCAGCGGTTTCGATCCCGCTATTCTCCA---  
GGGGCTTAGCTCAGTC-GGTAGAGCACTGCCTTTGCAAGGCAGATGTCAGGGGTTTCGATTCCCCTAGGCTCCA---  
GGGGGTATAGCTCAGTT-GGTAGAGCGCTGCCTTTGCAAGGCAGAAGTCAGCGGTTTCGA . TCCGCTTACCCCA---  
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTTCGATCCCGCATAGCTCCACCA  
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTTCGATCCCGCATAGCTCCACCA  
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA  
GGGGCCATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTTCAGGAGTTCGATCCTCCTTGGCTCCACCA  
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA  
GGGGCCATAGCTCAGCTGGGGAGAGCGCCTGCCTTGCACGCAGGAGGTCAACGGTTCGATCCCGTTTGGCTCCA---  
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA  
GGGGCATTAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTTCAGCGGTTTCGATCCCGCTATTCTCCACCA  
GGGGCCATAGCTCAGTT-GGTAGAGCGCCTGCTTTGCAAGCAGGTGT-CGTCGGTTCGAATCCGTCTGGCTCCACCA  
GGGGCCGTAGCTCAGCTGGG-AGAGCACCTGCTTTGCAAGCAGGGGGTTCGGAGGTTTCGATCCCGTCCGGCTCCACCA  
GGGGCCGTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTTCGTCGGTTCGATCCCGTCCGGCTCCACCA  
GGGGCCGTAGCTCAGCT-GG-AGAGCACCTGCTTTGCAAGCAGGGGGTTCGTCGGTTCGATCCCGTCCGGCTCCACCA

ARNt

GGGGAATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTTCAGCGGTTTCGATCCCGCTATTCTCCA---  
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTTCGATCCCGCATAGCTCCACCA  
GGGGCTATAGCTCAGCT-GGGAGAGCGCTTGCATGGCATGCAAGAGGTTCAGCGGTTTCGATCCCGCTTAGCTCCACCA  
GGGGAATTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTTCAGCGGTTTCGATCCCGCTATTCTCCA---  
GGGGCTTAGCTCAGTC-GGTAGAGCACTGCCTTTGCAAGGCAGATGTCAGGGGTTTCGATTCCCCTAGGCTCCA---  
GGGGGTATAGCTCAGTT-GGTAGAGCGCTGCCTTTGCAAGGCAGAAGTTCAGCGGTTTCGA . TCCGCTTACCCCA---  
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTTCGATCCCGCATAGCTCCACCA  
GGGGCTATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTTCGATCCCGCATAGCTCCACCA  
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA  
GGGGCCATAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTTCAGGAGTTCGATCCTCCTTGGCTCCACCA  
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA  
GGGGCCATAGCTCAGCTGGGGAGAGCGCCTGCCTTGCACGCAGGAGGTCAACGGTTCGATCCCGTTTGGCTCCA---  
GGGGGCATAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGT-CGTCGGTTCGATCCCGTCTGCCTCCACCA  
GGGGCAATTAGCTCAGCT-GGGAGAGCGCCTGCTTTGCACGCAGGAGGTTCAGCGGTTTCGATCCCGCTATTCTCCACCA  
GGGGCCATAGCTCAGTT-GGTAGAGCGCCTGCTTTGCAAGCAGGTGT-CGTCGGTTCGAATCCGTCTGGCTCCACCA  
GGGGCCGTAGCTCAGCTGGG-AGAGCACCTGCTTTGCAAGCAGGGGGTTCGGAGGTTTCGATCCCGTCCGGCTCCACCA  
GGGGCCGTAGCTCAGCT-GGGAGAGCACCTGCTTTGCAAGCAGGGGGTTCGTCGGTTCGATCCCGTCCGGCTCCACCA  
GGGGCCGTAGCTCAGCT-GG-AGAGCACCTGCTTTGCAAGCAGGGGGTTCGTCGGTTCGATCCCGTCCGGCTCCACCA

ARNt

# Vérification . . .



Structure secondaire de l'ARNt

# Comparaison de molécules d'ARN

## Comparaison des séquences

Algorithmes sur les mots

Alignement, recherche de motifs, etc.

## Comparaison des structures secondaires

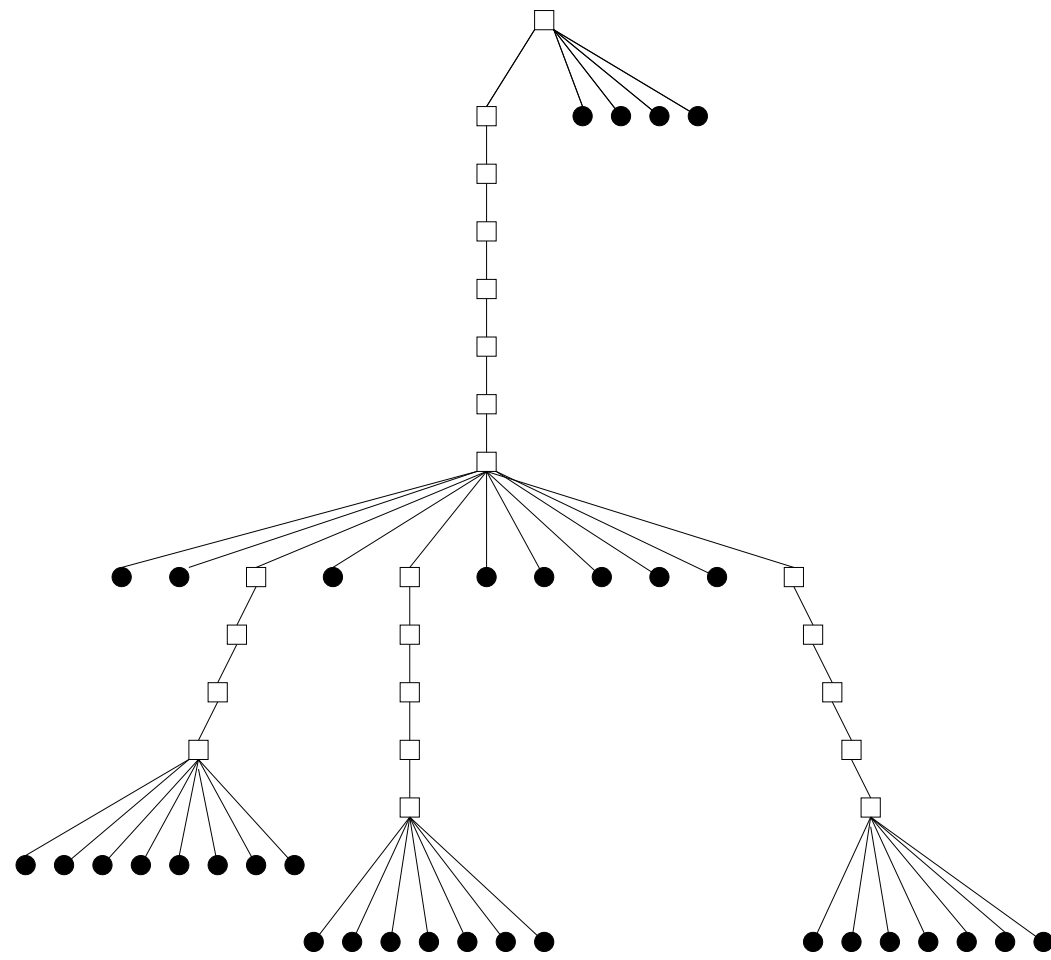
Algorithmes sur les arbres

Alignement, recherche de motifs, etc.

*La comparaison des structures secondaires prend en compte plus d'information, mais est moins courante :*

▷ *peu de structures sont connues*

▷ *complexité des algorithmes*



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

□ : 2 bases appariées (hélice)

● : base libre (boucle)