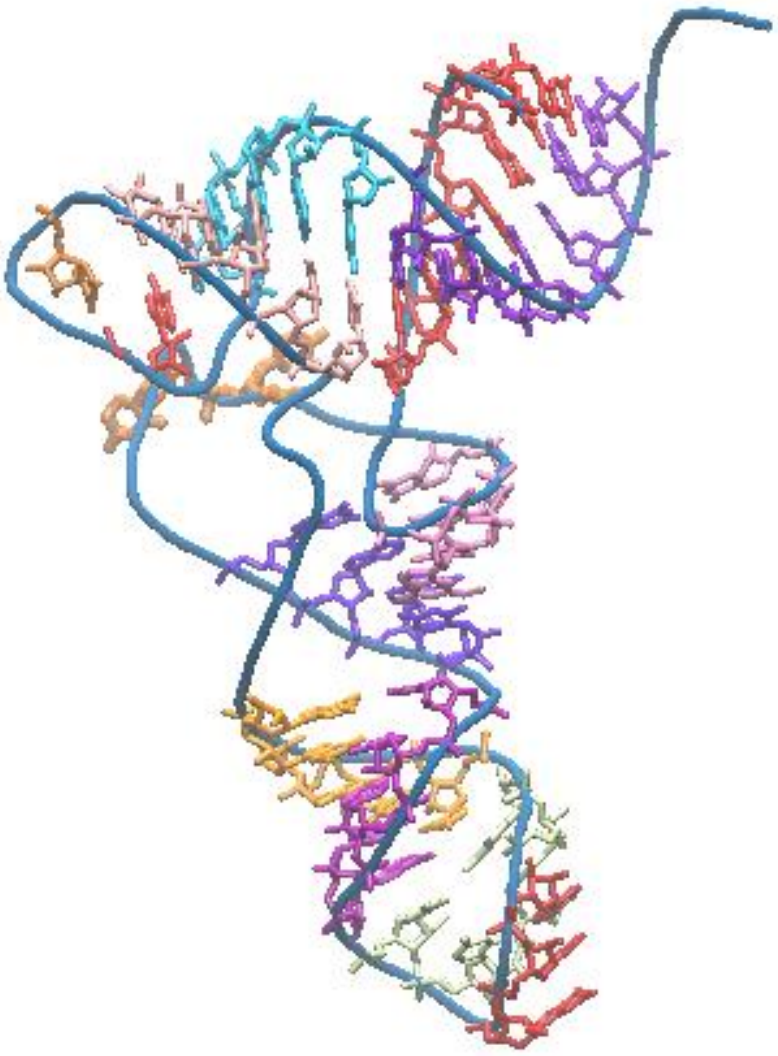
INTRODUCTION À LA PRÉDICTION DE STRUCTURES D'ARN

Manuel Lafond, Université de Sherbrooke

- Vous vous rappelez que l'ADN est transcrit en ARN.
- Structure de l'ADN = double-hélice 
- Structure de l'ARN = ???

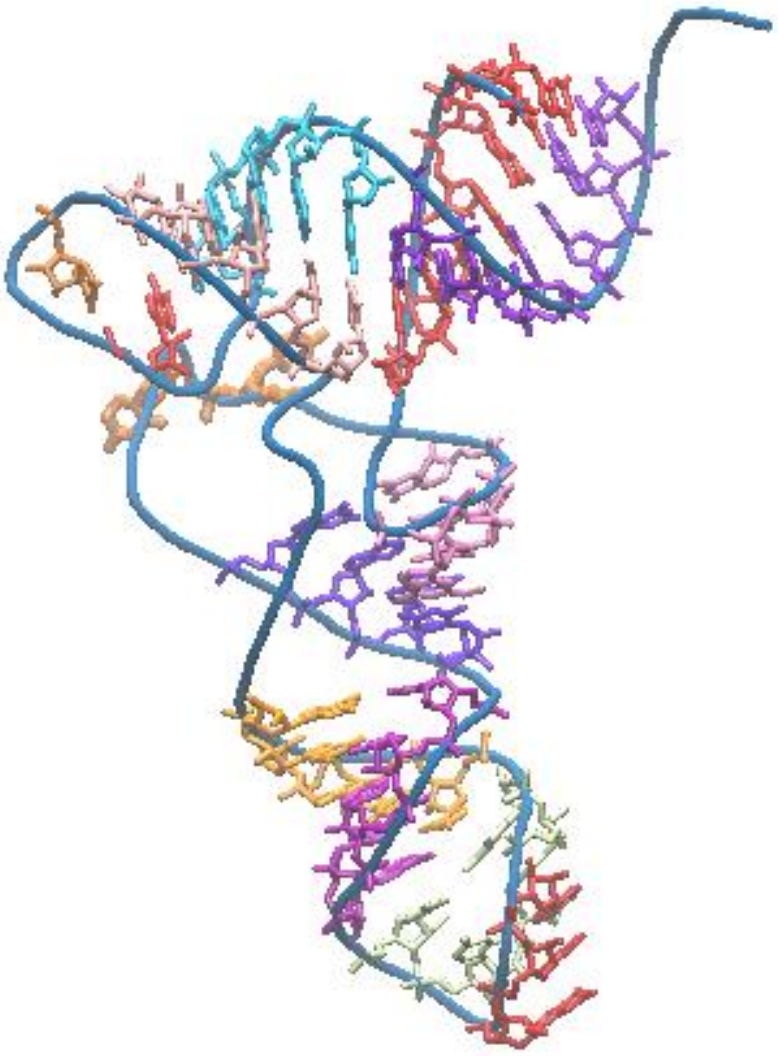
- Vous vous rappelez que l'ADN est transcrit en ARN.
- Structure de l'ADN = double-hélice 
- Structure de l'ARN = ???
  - ADN = alphabet A,C,G,T
  - ARN = alphabet A,C,G,U

- Vous vous rappelez que l'ADN est transcrit en ARN.
- Structure de l'ADN = double-hélice 
- Structure de l'ARN = ???
  - ADN = alphabet A,C,G,T
  - ARN = alphabet A,C,G,U
  - La structure dépend beaucoup de la séquence
  - Chaque nucléotide (A,C,G,U) veut s'apparier.

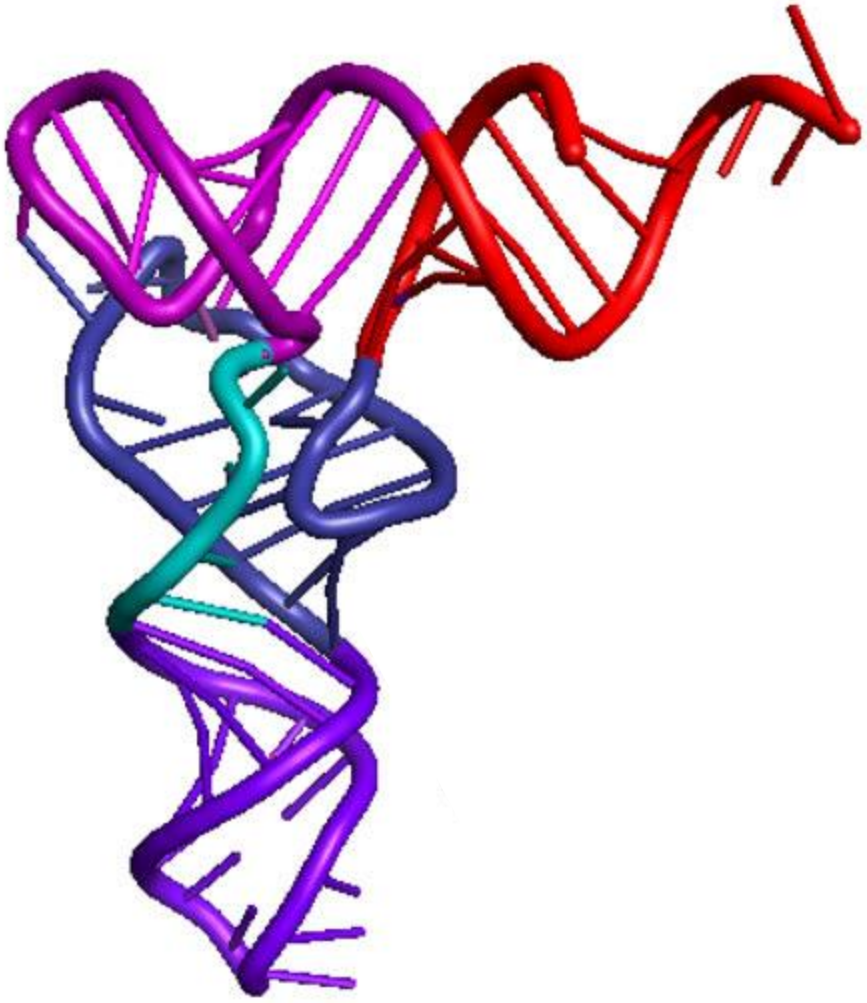


Source: http://lowelab.ucsc.edu/tRNAscan-SE/image/3D-tRNA-lowelab.png

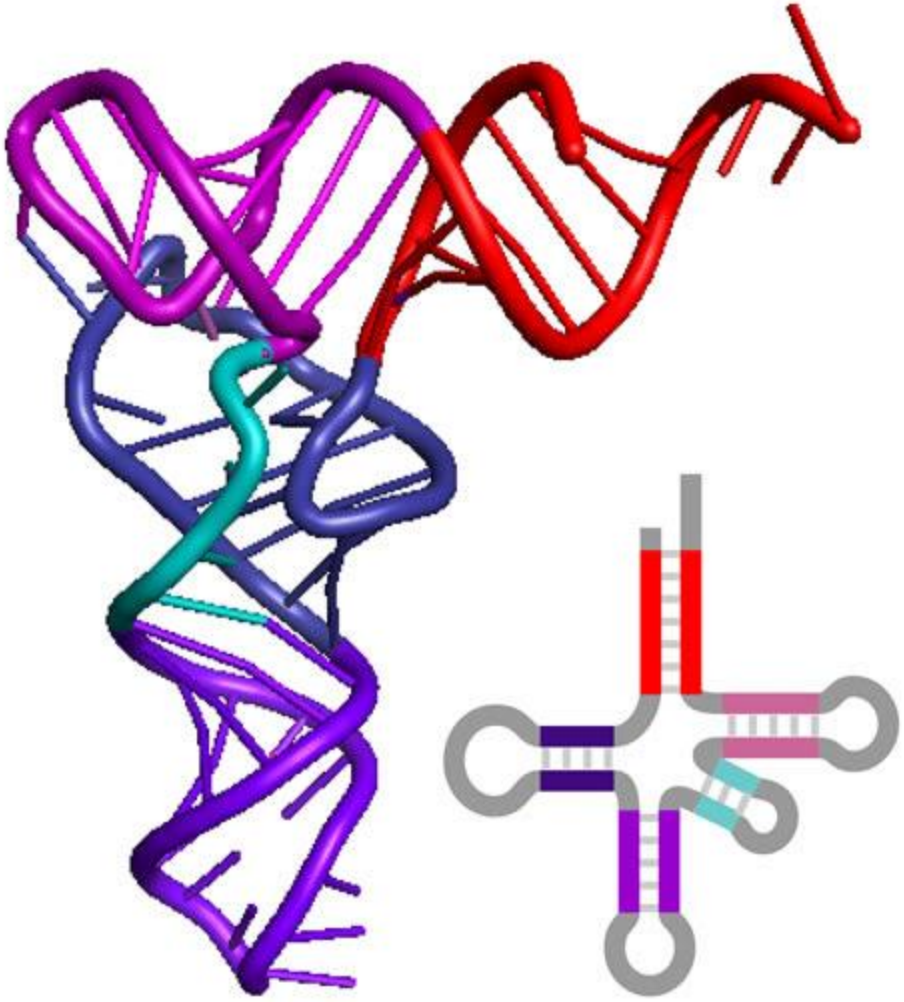
- ADN => 2 brins => chaque nucléotide déjà apparié
- ARN => 1 brin => recherche d'"auto-appariements"



Source: http://lowelab.ucsc.edu/tRNAscan-SE/image/3D-tRNA-lowelab.png

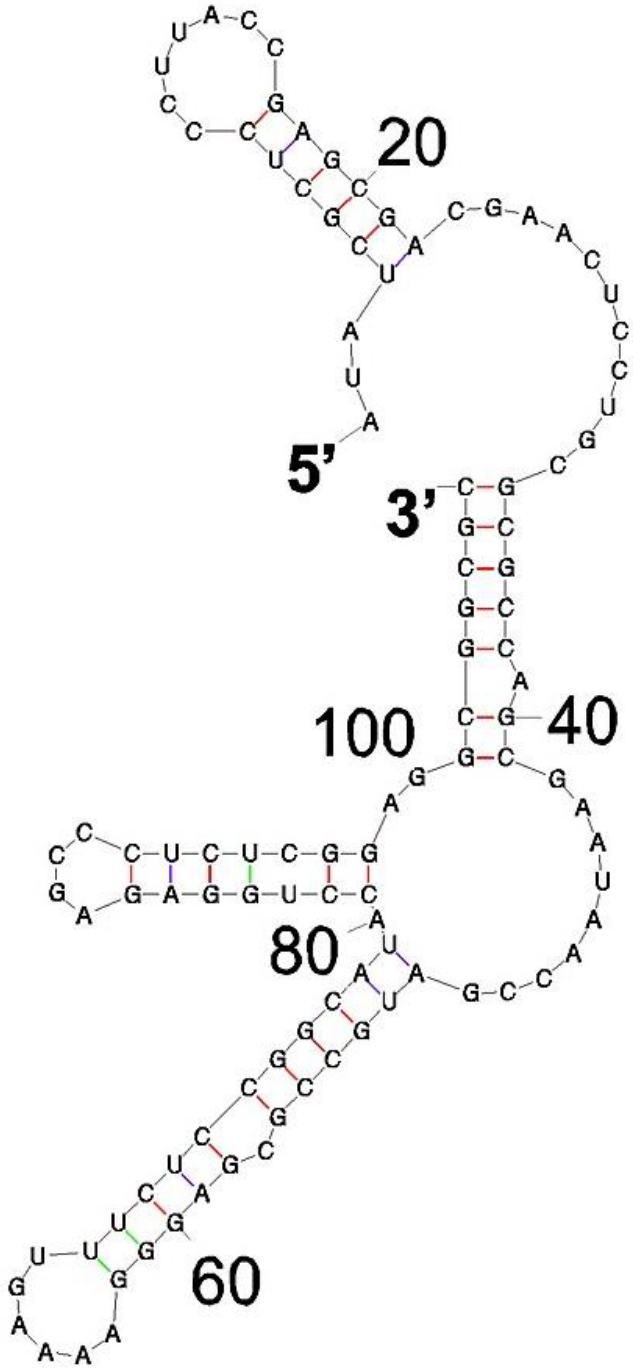


Source: https://news.illinois.edu/files/6367/206368/44394.jpg

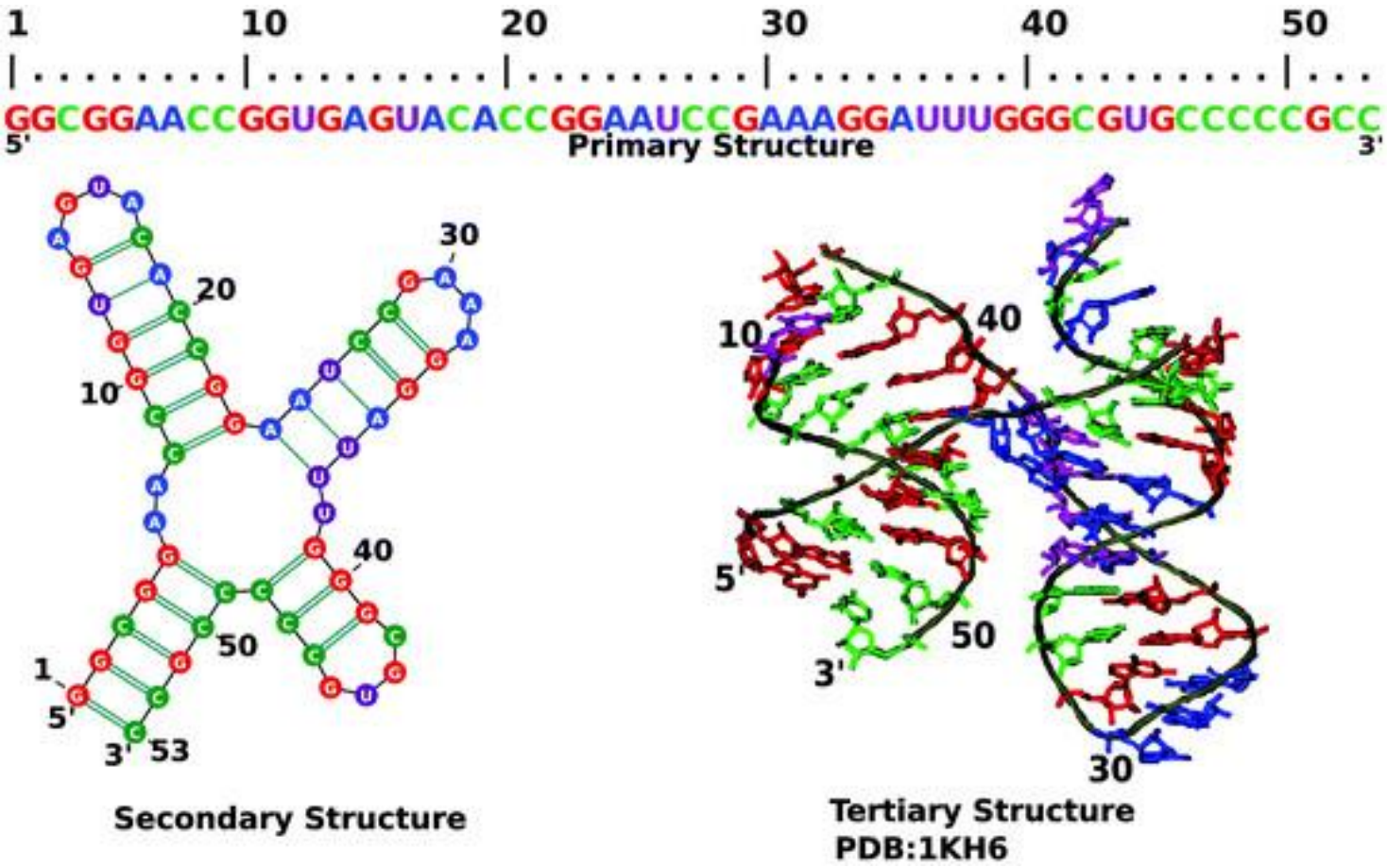


Source: https://news.illinois.edu/files/6367/206368/44394.jpg

- La structure 2D sert à montrer les appariements.
- Les sous-structures de la représentation informent (parfois) sur la fonction de l'ARN.
- Non-appariement = énergie libre (à combler)



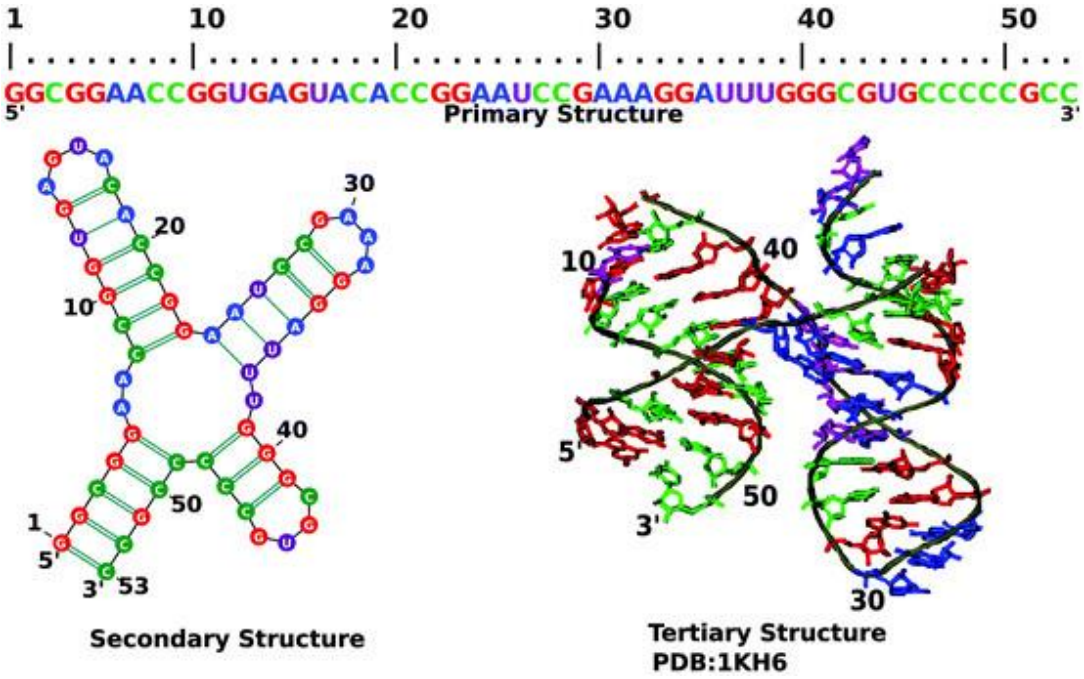
Structures de l'ARN



Source: Kim et al., "Graph Applications to RNA Structure and Function", Biophysics of RNA Folding 2012

Structures de l'ARN

- Structure primaire: la séquence (1D)
- Structure secondaire: paires de bases appariées (2D)
- Structure teritiaire: représentation (3D)

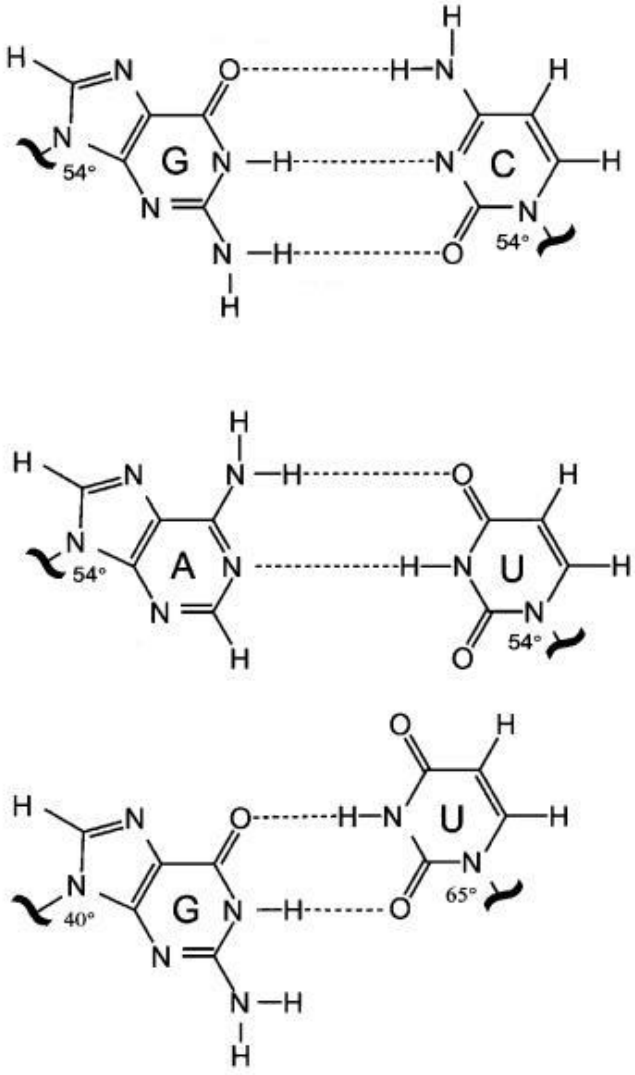


Source: Kim et al., "Graph Applications to RNA Structure and Function", Biophysics of RNA Folding 2012

Qui se ressemble s'apparie

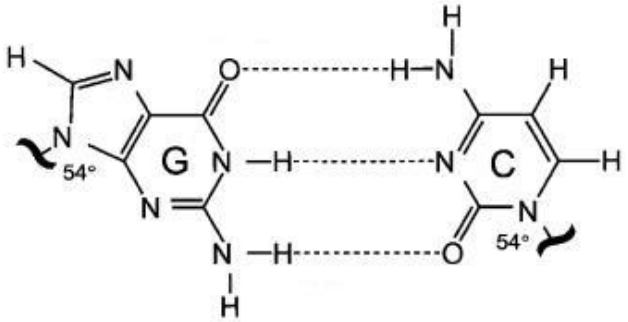
- Paire de bases = deux nucléotides appariés
- Appariements canoniques
  - G - C
  - A - U
- Appariement "wobble"
  - G - U

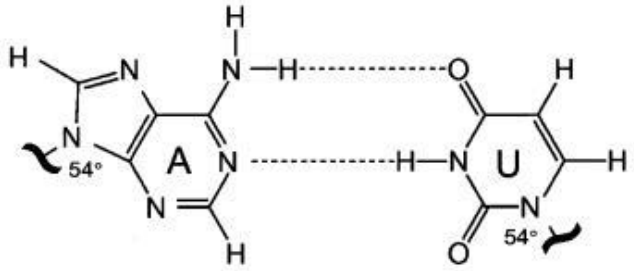
- Autres appariements possibles, mais rares.



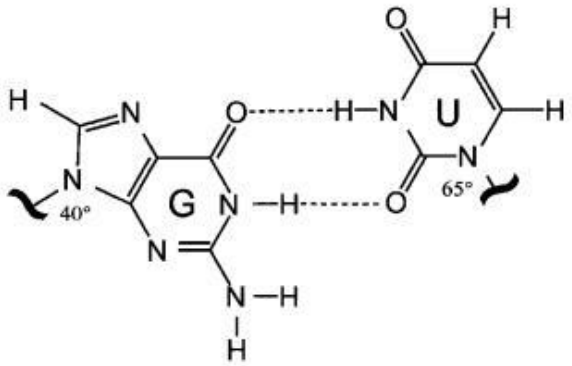
Qui se ressemble s'apparie

- Pour nos fins, appariements possibles
  - G - C
  - A - U
  - G - U





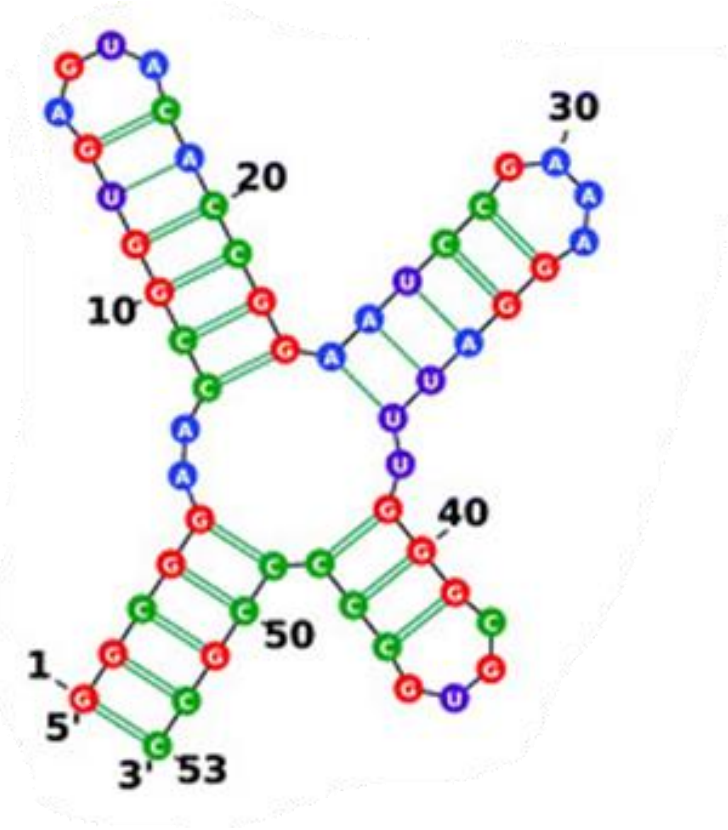
- Possiblement avec poids différent.



Qui se ressemble s'apparie

- Pour nos fins, appariements possibles
  - G - C
  - A - U
  - G - U

- Possiblement avec poids différent.



Modélisation du problème

- Entrée: séquence S d'ARN
- Sortie: liste des paires de positions

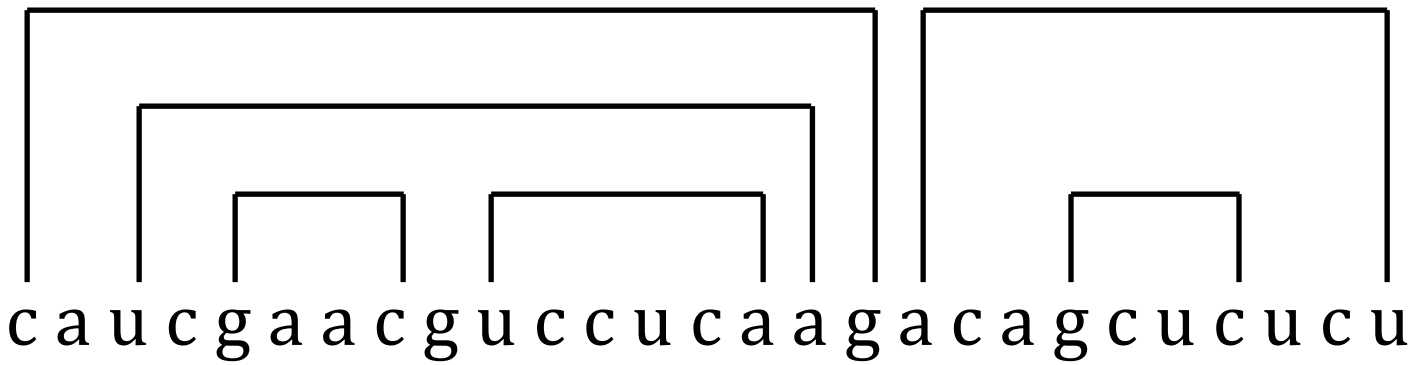
  $(s_1, t_1), (s_2, t_2), \ldots, (s_k, t_k)$ des paires de bases.

c a u c g a a c g u c c u c a a g a c a g c u c u c u

Modélisation du problème

- Entrée: séquence S d'ARN
- Sortie: liste des paires de positions

  $(s_1, t_1), (s_2, t_2), ..., (s_k, t_k)$ des paires de bases.



Modélisation du problème

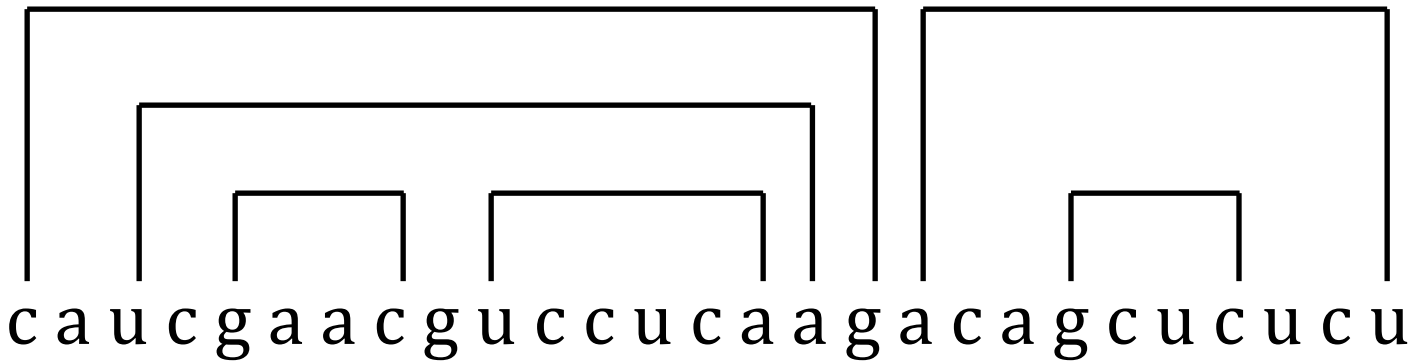
- Entrée: séquence S d'ARN
- Sortie: liste des paires de positions

  $(s_1, t_1), (s_2, t_2), ..., (s_k, t_k)$ des paires de bases.

- Nucléotide non-apparié = énergie libre
- **Hypothèse 0**: l'ARN va se replier de façon à minimiser l'énergie libre.
  - => l'ARN a tendance à maximiser les appariements

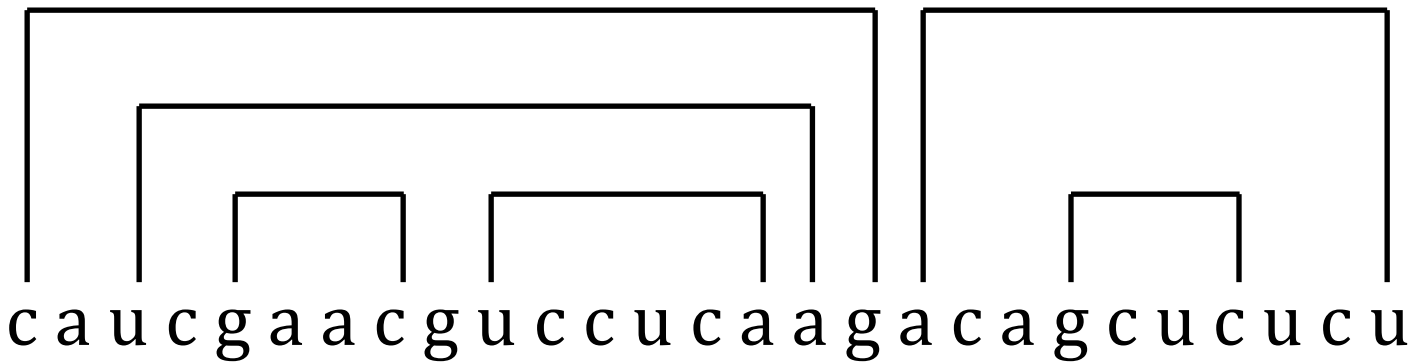
Modélisation du problème

- Nucléotide non-apparié = énergie libre
- **Hypothèse 0**: l'ARN va se replier de façon à minimiser l'énergie libre.
  - => l'ARN a tendance à maximiser les appariements



Modélisation du problème

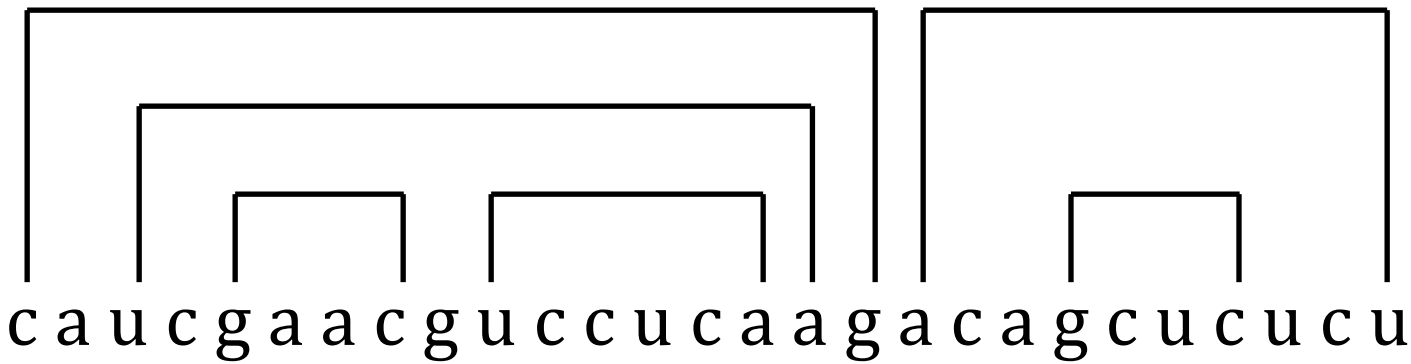
- **Hypothèse 1**: chaque nucléotide est apparié 0 ou 1 fois.



Modélisation du problème

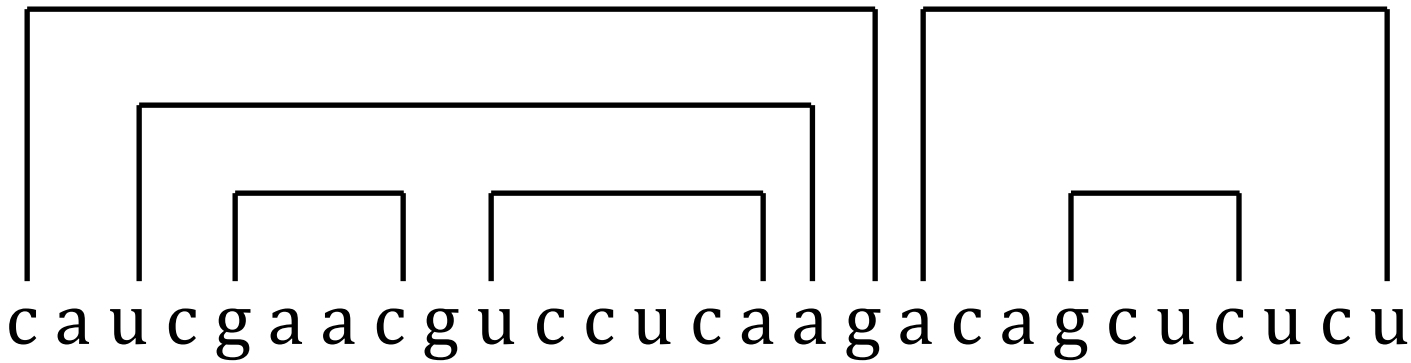
- **Hypothèse 1**: chaque nucléotide est apparié 0 ou 1 fois.

- **Hypothèse 2**: les appariements se font de façon "planaire" dans l'espace
  - Pas de croisement entre les appariements



Modélisation du problème

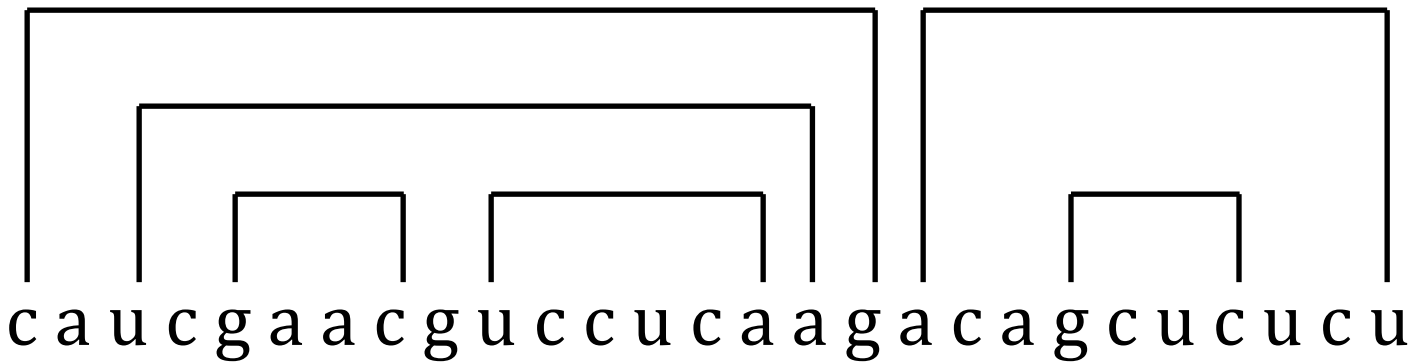
- **Hypothèse 1**: chaque nucléotide est apparié 0 ou 1 fois.
  - *Pas toujours vrai (triplets de bases existent)*
- **Hypothèse 2**: les appariements se font de façon "planaire" dans l'espace
  - Pas de croisement entre les appariements
  - *Pas toujours vrai (croisement = pseudo-noeud)*



Modélisation du problème

- **Hypothèse 1**: chaque nucléotide est apparié 0 ou 1 fois.

- **Hypothèse 2**: les appariements se font de façon "planaire" dans l'espace
  - Pas de croisement entre les appariements



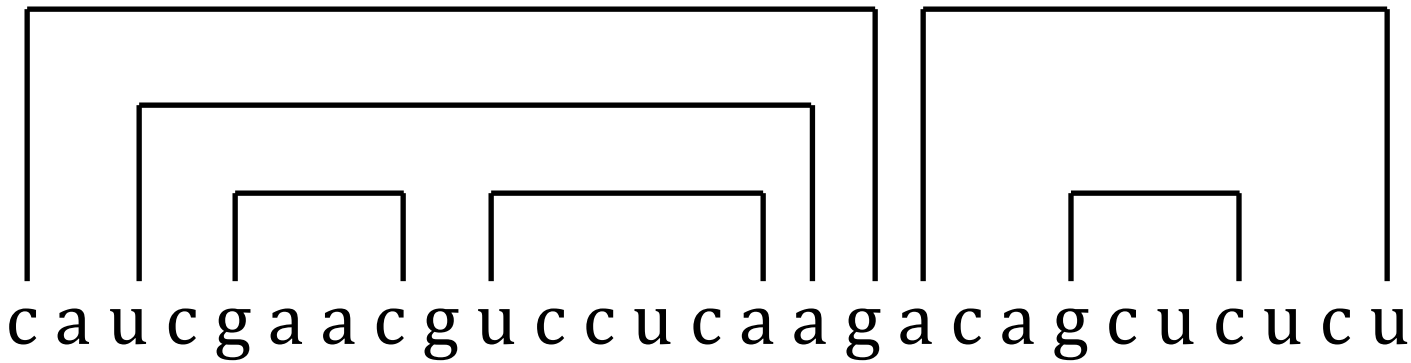
Modélisation du problème

- L'hypothèse du non-croisement donne une condition d'emboîtement sur les appariements

  $A = \{ (s_1, t_1), (s_2, t_2), ..., (s_k, t_k) \}$

- Pour tout (s_i, t_i) et (s_j, t_j) dans A, on a

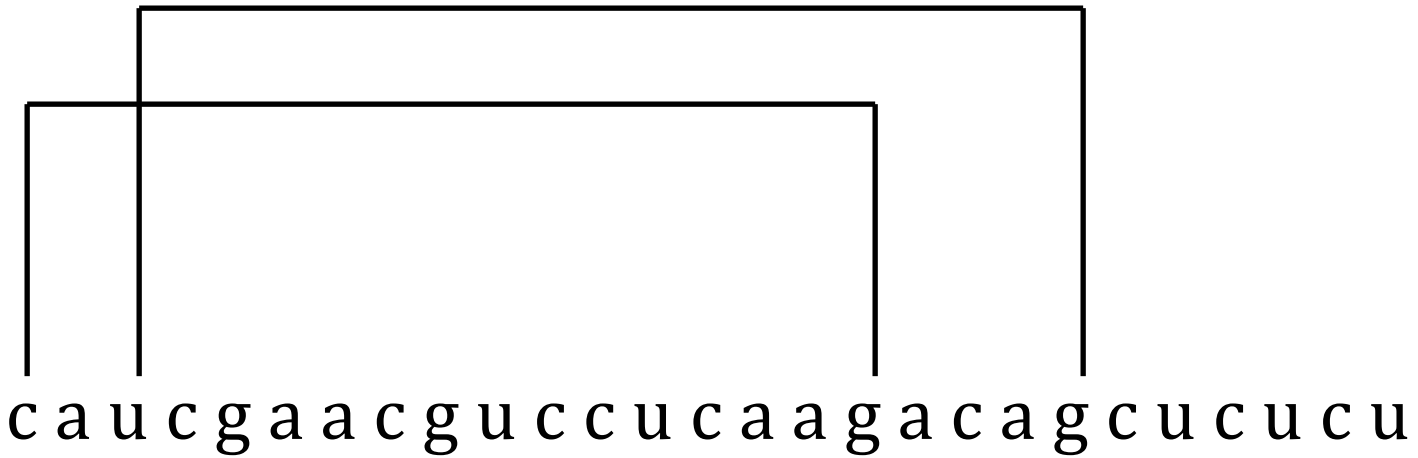
  $s_i < s_j < t_i$ si et seulement si $s_i < t_j < t_i$

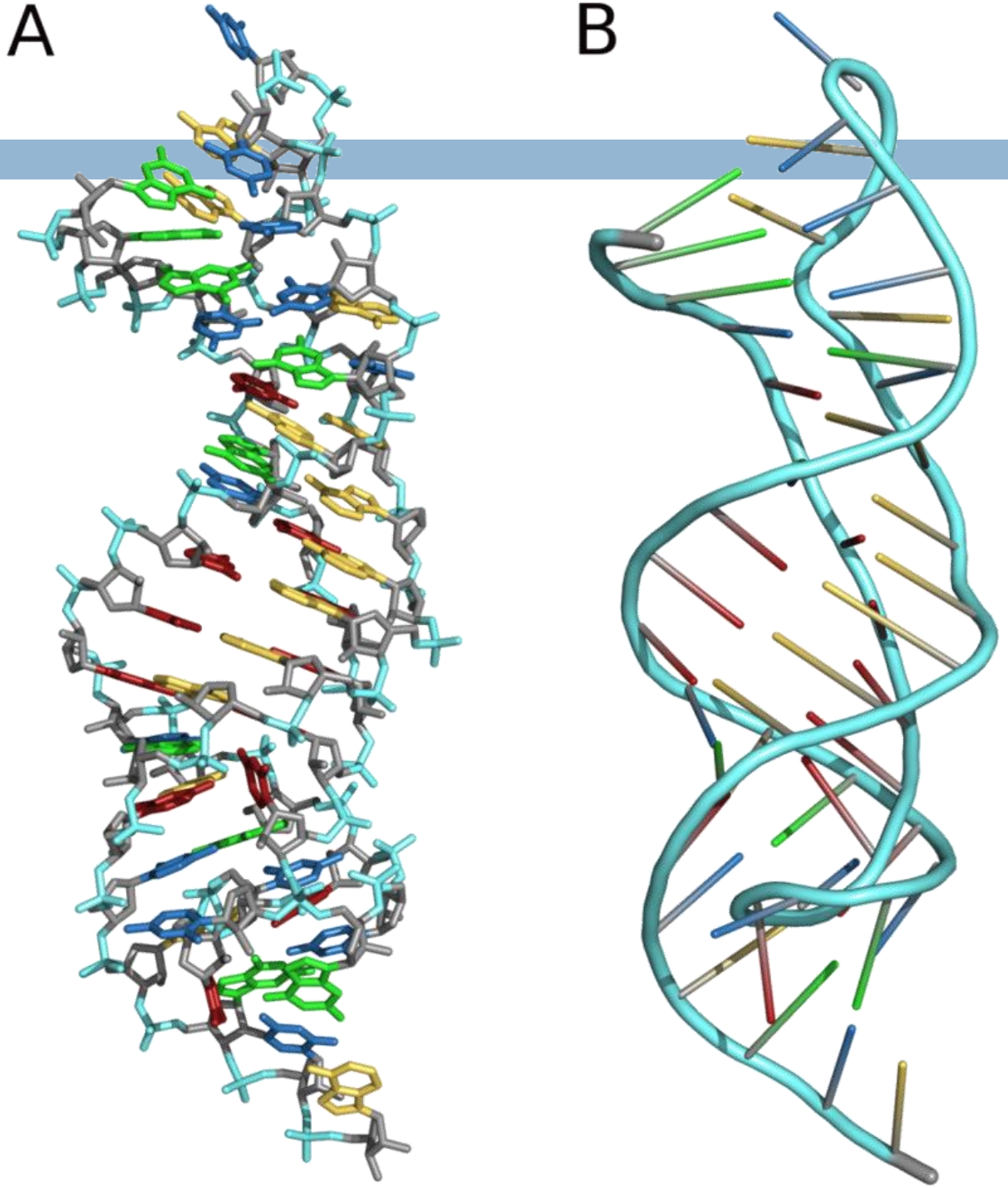


Modélisation du problème

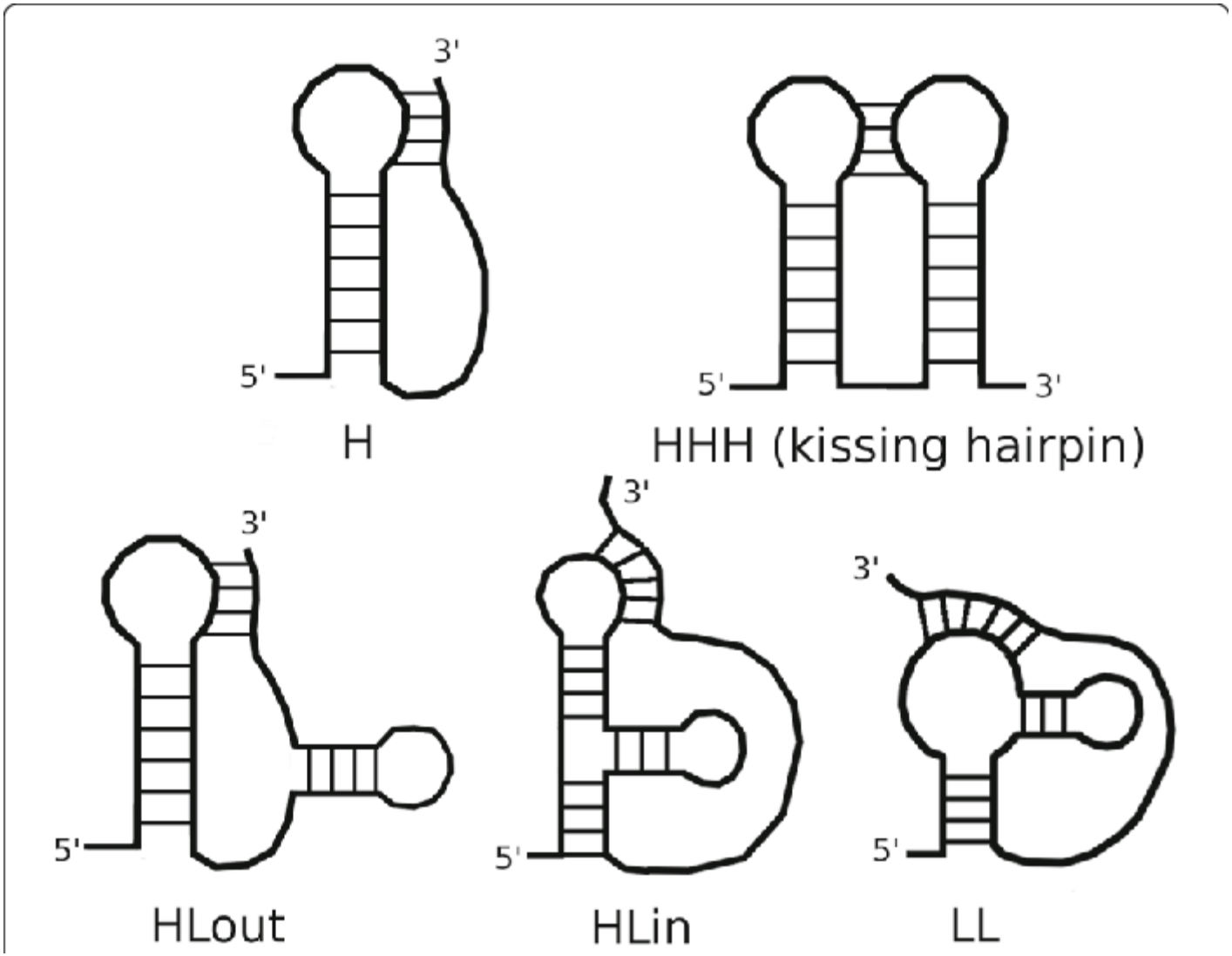
- Pour tout (s_i, t_i) et (s_j, t_j) dans A, on a

  $s_i < s_j < t_i$ si et seulement si $s_i < t_j < t_i$
- Si cette condition n'est pas respectée, on a ce qu'on appelle un *pseudo-noeud.*




A
B

Pseudo-noeuds



Modélisation du problème

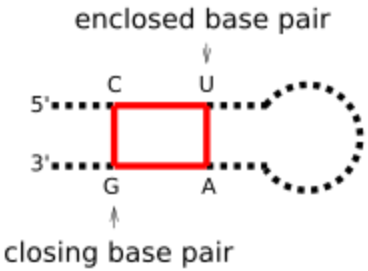
- Pour tout (s_i, t_i) et (s_j, t_j) dans A, on a

  $s_i < s_j < t_i$ si et seulement si $s_i < t_j < t_i$
- Si cette condition n'est pas respectée, on a ce qu'on appelle un *pseudo-noeud.*

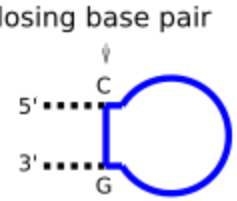
- L'hypothèse 2 implique "pas de pseudo-noeuds".
  - Relativement rare (mais bien existant), et compliqué!

Modélisation du problème

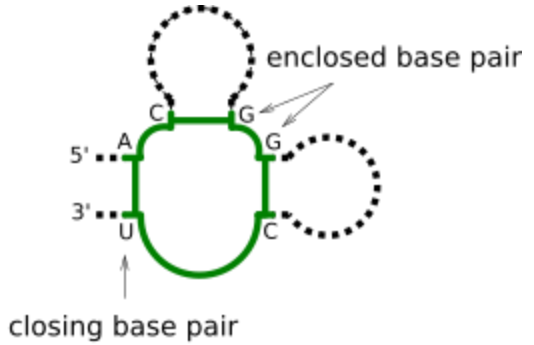
- Même sans pseudo-noeuds, il y a plusieurs sous-structures, chacune avec ses implications en énergie libre.



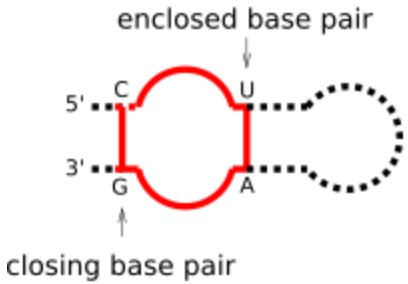
stacking pair
(empilement)



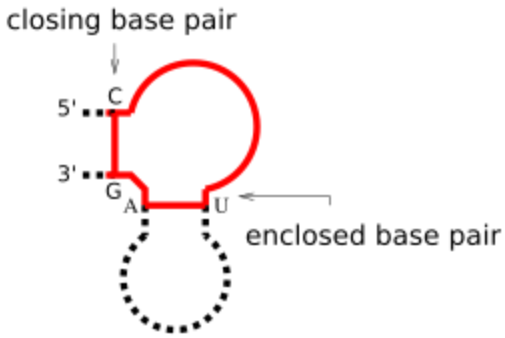
hairpin loop
(tige-boucle/hélice)



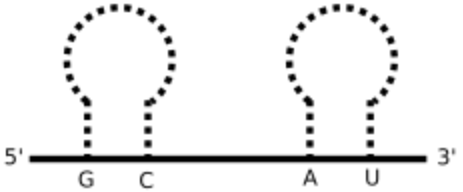
multi loop
(boucle multiple)



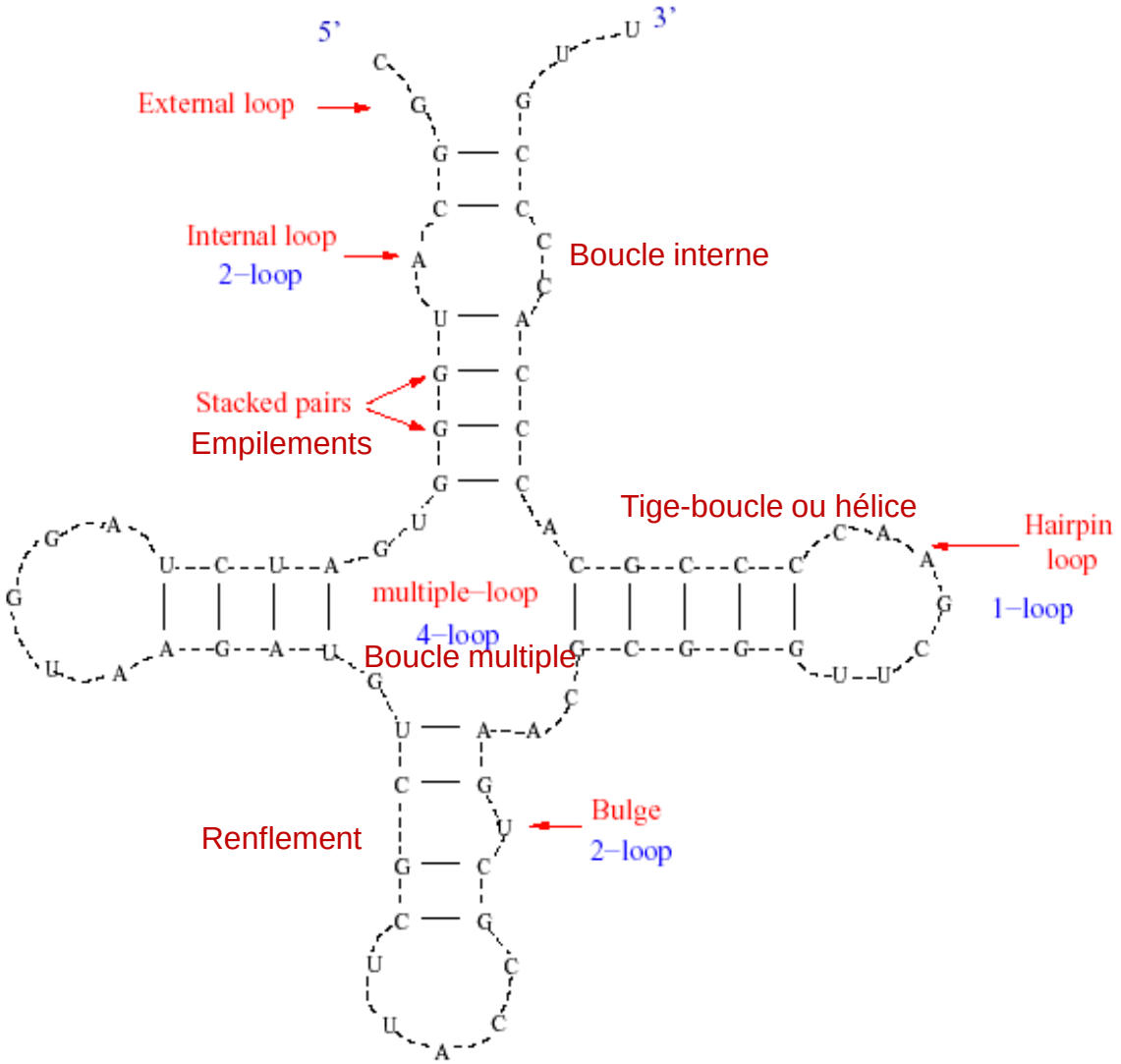
interior loop
(boucle intérieure)



bulge
(renflement)



exterior loop
(boucle externe)


5'
3'
External loop
Internal loop
2–loop
Boucle interne
Stacked pairs
Empilements
Tige-boucle ou hélice
Hairpin
loop
1–loop
multiple–loop
4–loop
Boucle multiple
Renflement
Bulge
2–loop

Modélisation du problème

- Pour l'instant, version simple du problème: maximiser les paires de bases sous nos hypothèse.

- Entrée: séquence S d'ARN
- Sortie: nombre **maximum** de paires de positions $(s_1, t_1), (s_2, t_2), ..., (s_k, t_k)$ de paires de bases avec condition $s_i < s_j < t_i$ si et seulement si $s_i < t_j < t_i$

Modélisation du problème

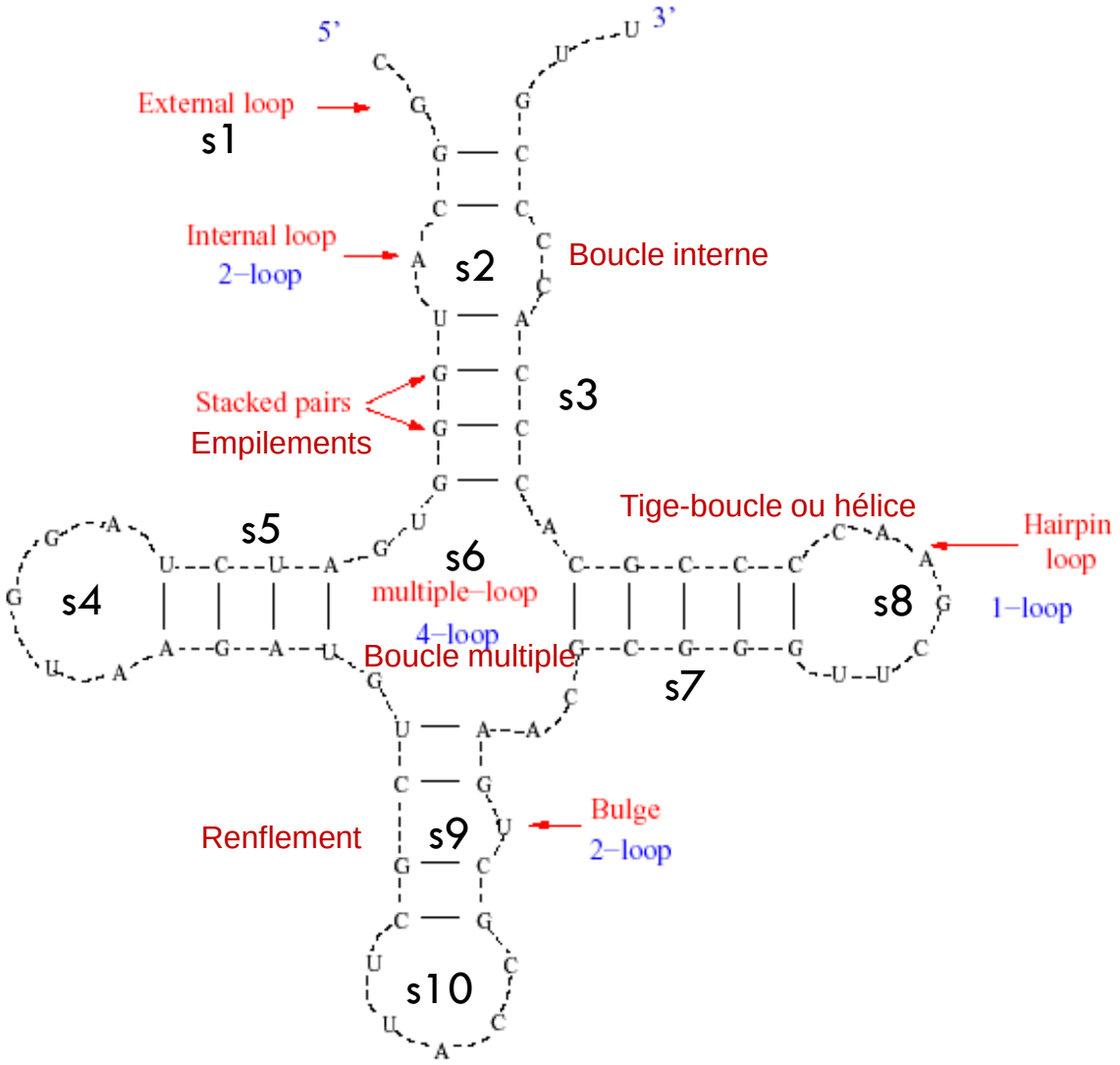
- Pour l'instant, version simple du problème: maximiser les paires de bases sous nos hypothèse.

- Entrée: séquence S d'ARN
- Sortie: nombre **maximum** de paires de positions $(s_1, t_1), (s_2, t_2), ..., (s_k, t_k)$ de paires de bases avec condition $s_i < s_j < t_i$ si et seulement si $s_i < t_j < t_i$

Algorithme de Nussinov (1978)

Modélisation du problème V2

- Formulation simpliste: ignore le fait que chaque sous-structure a son propre apport en énergie libre.


5'
3'
External loop
s1
Internal loop
2–loop
s2
Boucle interne
Stacked pairs
Empilements
s3
Tige-boucle ou hélice
Hairpin
loop
s5
s6
multiple–loop
4–loop
Boucle multiple
s4
s8
1–loop
s7
Renflement
s9
Bulge
2–loop
s10

Modélisation du problème V2

- Formulation simpliste: ignore le fait que chaque sous-structure a son propre apport en énergie libre.
- Soient s_1, s_2, ..., s_k les sous-structures de l'ARN. Son énergie libre est

$$E(S) = e(s_1) + e(s_2) + \ldots + e(s_k)$$

Modélisation du problème V2

- Formulation simpliste: ignore le fait que chaque sous-structure a son propre apport en énergie libre.
- Soient s_1, s_2, ..., s_k les sous-structures de l'ARN. Son énergie libre est

$$E(S) = e(s_1) + e(s_2) + \ldots + e(s_k)$$

- Fonction e évaluée expérimentalement sur de petits ARN synthétiques.

Modélisation du problème V2

Énergie d'empilement de paires de base (kcal/mole à 37degC):

	A/U	C/G	G/C	U/A	G/U	U/G
A/U	-0.9	-1.8	-2.3	-1.1	-1.1	-0.8
C/G	-1.7	-2.9	-3.4	-2.3	-2.1	-1.4
G/C	-2.1	-2.0	-2.9	-1.8	-1.9	-1.2
U/A	-0.9	-1.7	-2.1	-0.9	-1.0	-0.5
G/U	-0.5	-1.2	-1.4	-0.8	-0.4	-0.2
U/G	-1.0	-1.9	-2.1	-1.1	-1.5	-0.4

Turner, Sugimoto (1988)

Énergie de destabilisation pour les autres boucles:

Nombre de bases	**1**	**5**	**10**	**20**	**30**
Boucle interne	-	5.3	6.6	7.0	7.4
Bulge	3.9	4.8	5.5	6.3	6.7
Épingle à cheveux	-	4.4	5.3	6.1	6.5

Serra, Turner (1995)

- Désolé pour les amateurs de slides - elles s'arrêtent ici.