Prédiction de structures secondaires d'ARN

Nadia El-Mabrouk

I. Introduction - ARN

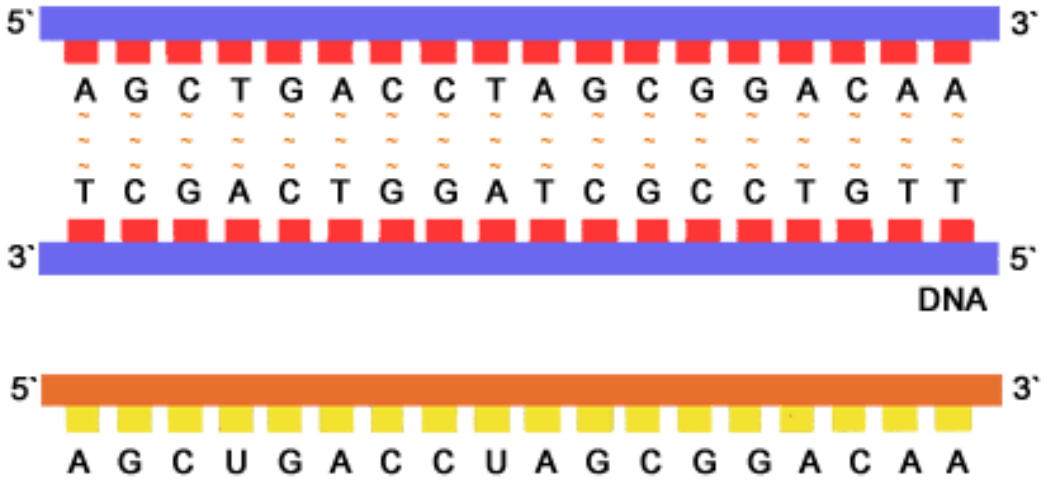
- L'ARN (acide ribonucléique) est une macromolécule qui remplissent un grand nombre de fonctions.
  - Au départ, vus exclusivement comme l'intermédiaire entre l'ADN et les protéines. Rôle exclusif dans la transcription/traduction du message contenu dans l'ADN.
  - Depuis 1980: découverte d'ARN non codants (non transcrits en protéines), jouant de multiples rôles dans la cellule: fonction catalytique, régulation de la transcription, épissage des introns, expression des gènes… A ravivé l'hypothèse du « RNA world », i.e. le fait que l'ARN a précédé l'ADN.
- De nouveaux ARN sont découverts continuellement.
- Conception d'ARN artificiels, en particulier miARN, permettant de contrôler l'expression de certains gènes impliqués dans des maladies comme le cancer.

I. Introduction - ARN non-codant

- Un ARN non-codant (ncRNA) est un ARN fonctionnel qui n'est pas traduit en protéine: tous les ARN autres que les ARNm.
- Gène d'ARN: Séquence d'ADN transcrite en un ARN non-codant.
- ARN non-codants incluent:
  - Les familles d'ARN ayant un rôle fondamental dans la synthèse des protéines:
    - ARN de transfert (ARNt ou tRNA)
    - ARN ribosomique (ARNr)
  - Beaucoup d'autres familles découvertes plus récemment, dont les fonctions ne sont pas toujours connues: snoRNAs, microRNAs, siRNAs, piRNAs, RNase P…

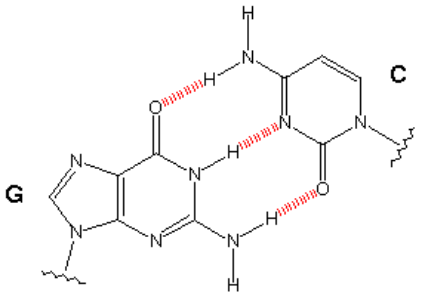
I. Introduction - Séquence d'ARN

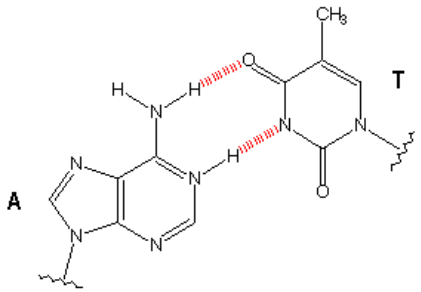
- Structure primaire: Séquence linéaire de 4 acides ribonucléiques: A(dénine), C(ytosine), G(uanine), U(racile)
- Un gène d'ARN est transcrit en une séquence primaire d'ARN:



I. Introduction - Structure secondaire

- En se repliant, la molécule d'ARN forme une structure stabilisée par la force des ponts hydrogènes entre certaines paires de bases (A-U, C-G, G-U) et l'empilement de paires de bases voisines. C'est la structure secondaire de la molécule.
- Paires de bases Watson-Crick: G-C et A-U
- Paire de bases Wobble: G-U





La paire G-C est celle qui donne le plus de stabilité, puis A-U, puis G-U.

http://e-sante.futura-sciences.com/_actualites/base-adn.html

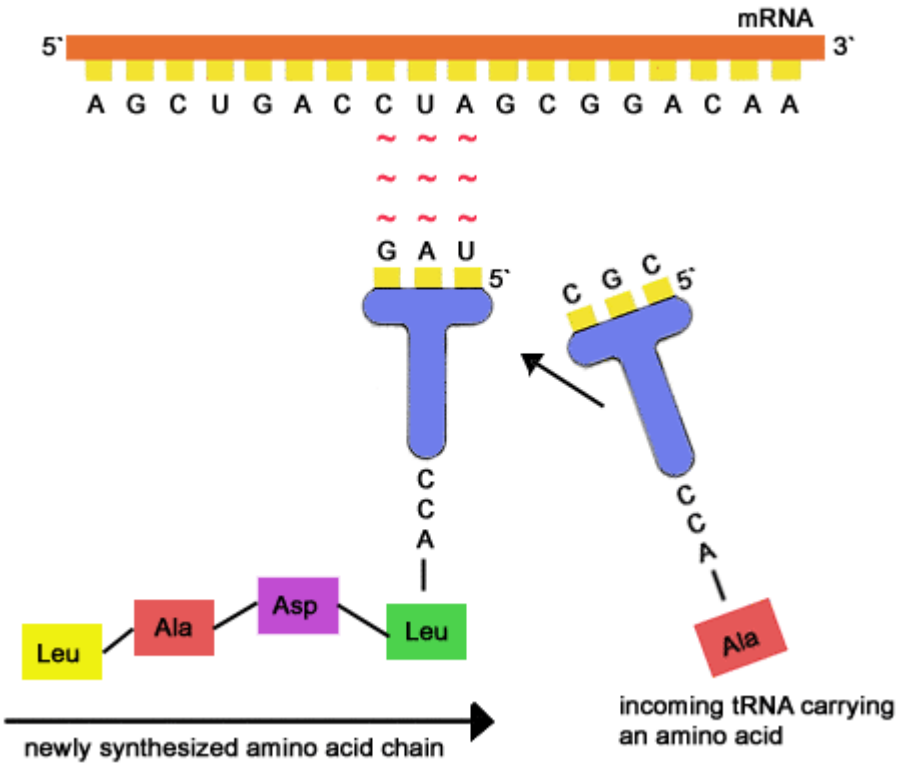
I. Introduction

Exemple: l'ARN de transfert

- L'ARN de transfert est l'« interpréteur » des codons de l'ARNm.
- Le rôle de l'ARN de transfert est de transférer les acides aminés du cytoplasme au ribosome.
- Chaque ARN de transfert a un triplet spécifique appelé « anticodon », qui s'apparie au codon complémentaire sur l'ARNm.
- Des enzymes appelées ARN Synthétases, chargent chaque ARNt avec son AA propre.

I. Introduction

Exemple: l'ARN de transfert



http://library.thinkquest.org/C004535/rna_translation.html

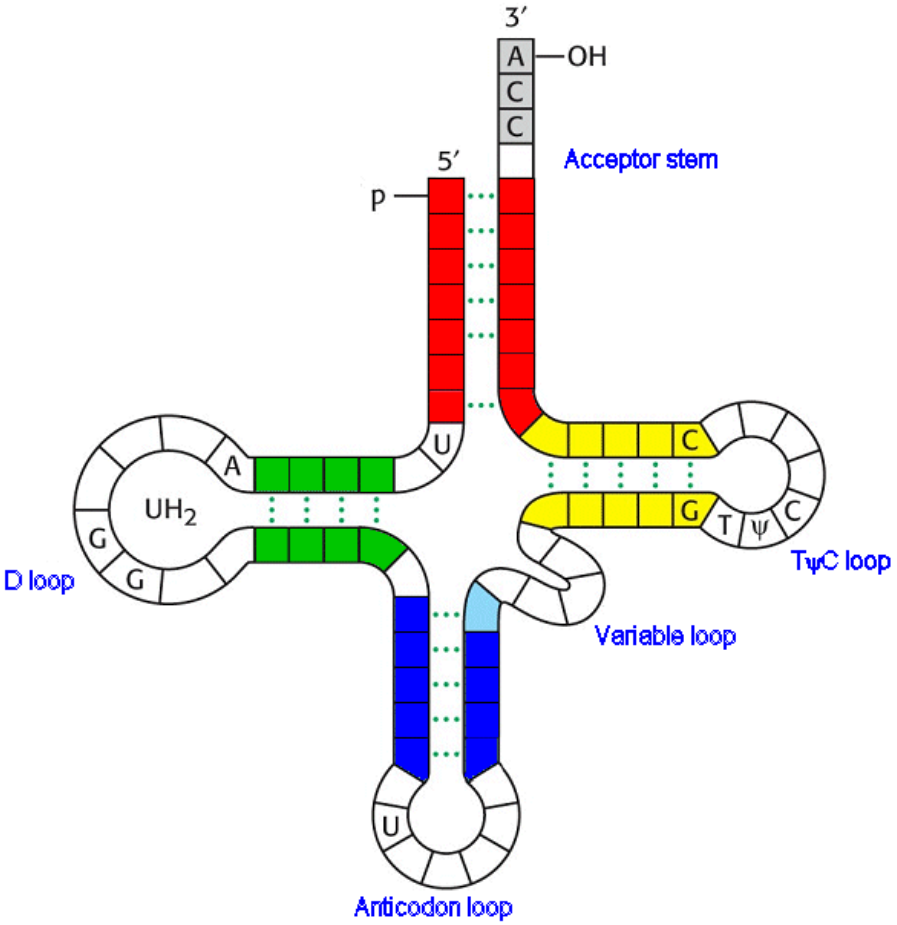
I. Introduction

Exemple: l'ARN de transfert

Molécule d'environ 70-80 nucléotides.

Structure secondaire de l'ARNt
En « feuille de trèfle » - Structure consensus.

L'AA correspondant à l'anticodon s'attache au bras accepteur.

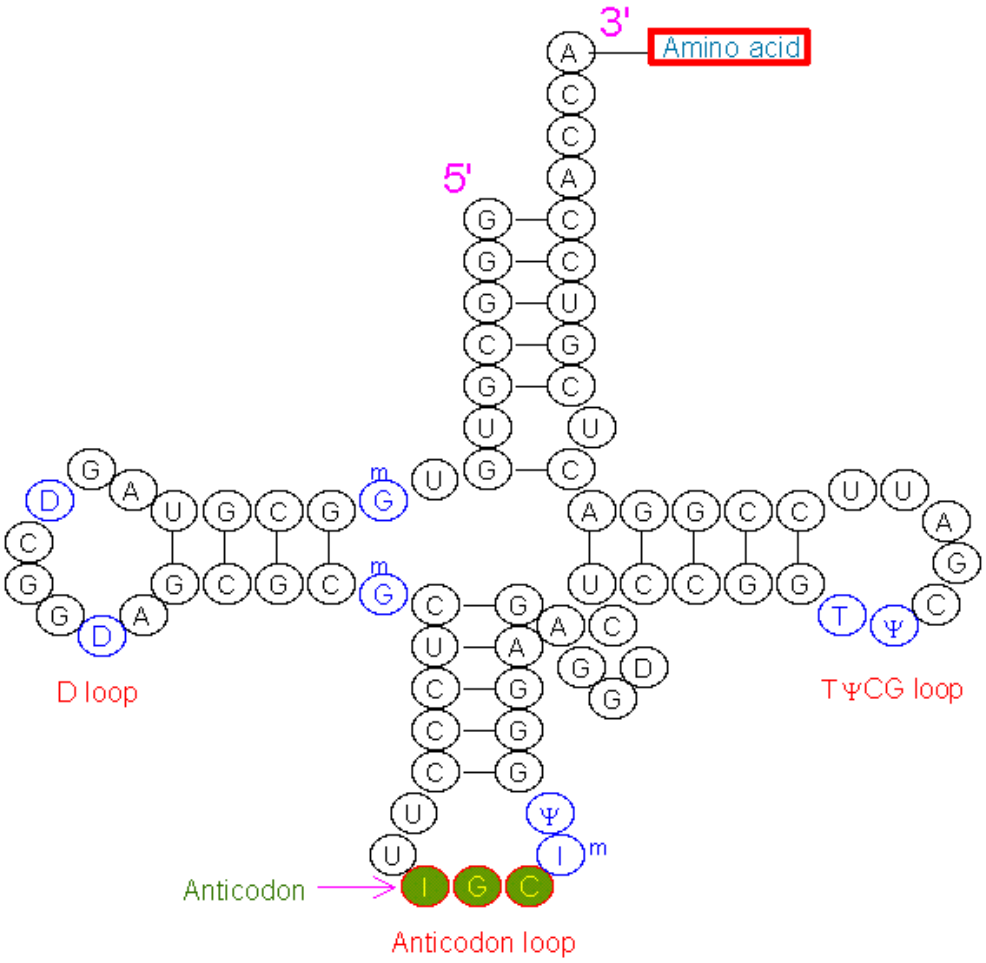


http://www.daviddarling.info/encyclopedia/T/tRNA.html

I. Introduction

Exemple: l'ARN de transfert

Some nucleotides in tRNA have been modified, such as **dihydrouridine (D),pseudouridine (Y)**, and **inosine (I)**. In dihydrouridine, a hydrogen atom is added to each C5 and C6 of uracil. In pseudouridine, the ribose is attached to C5, instead of the normal N1. Inosine plays an important role in codon recognition. In addition to these modifications, a few nucleosides are methylated.

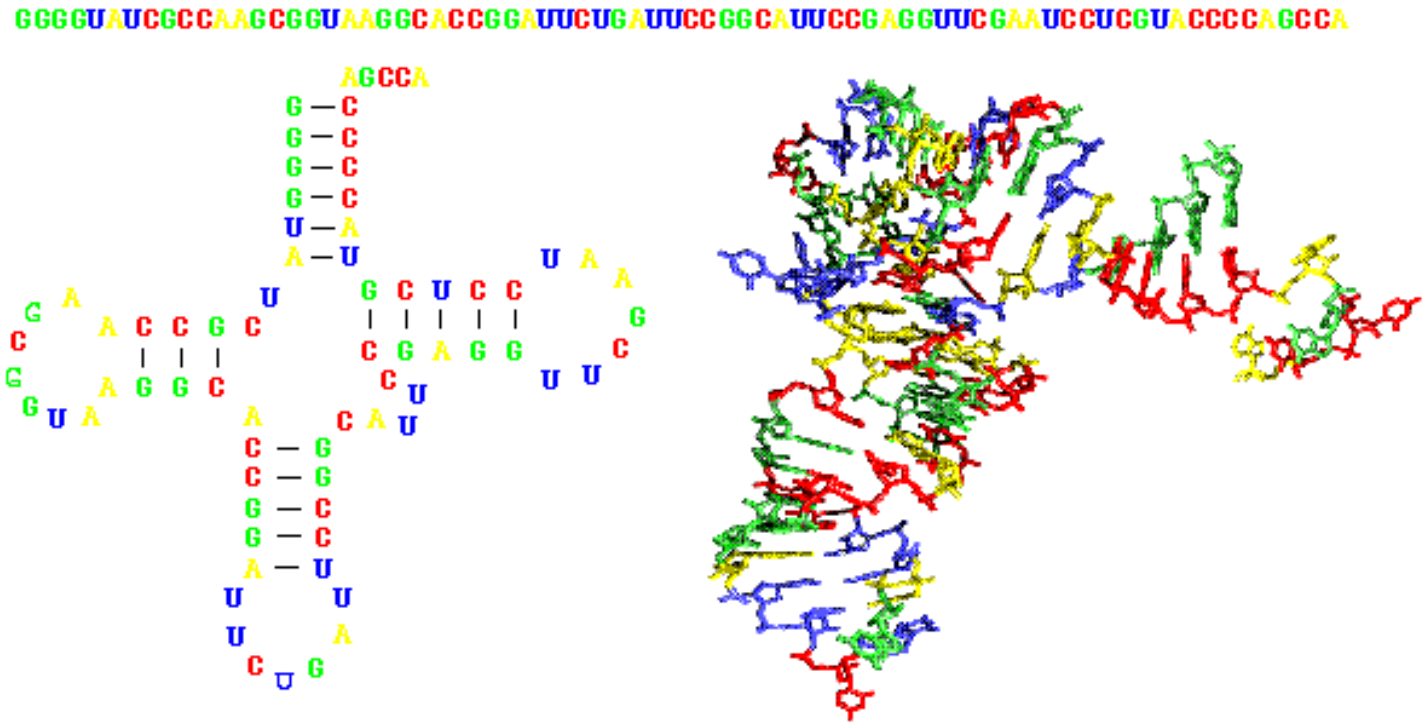


http://www.web-books.com/MoBio/Free/Ch3C2.htm

I. Introduction

Structure tertiaire

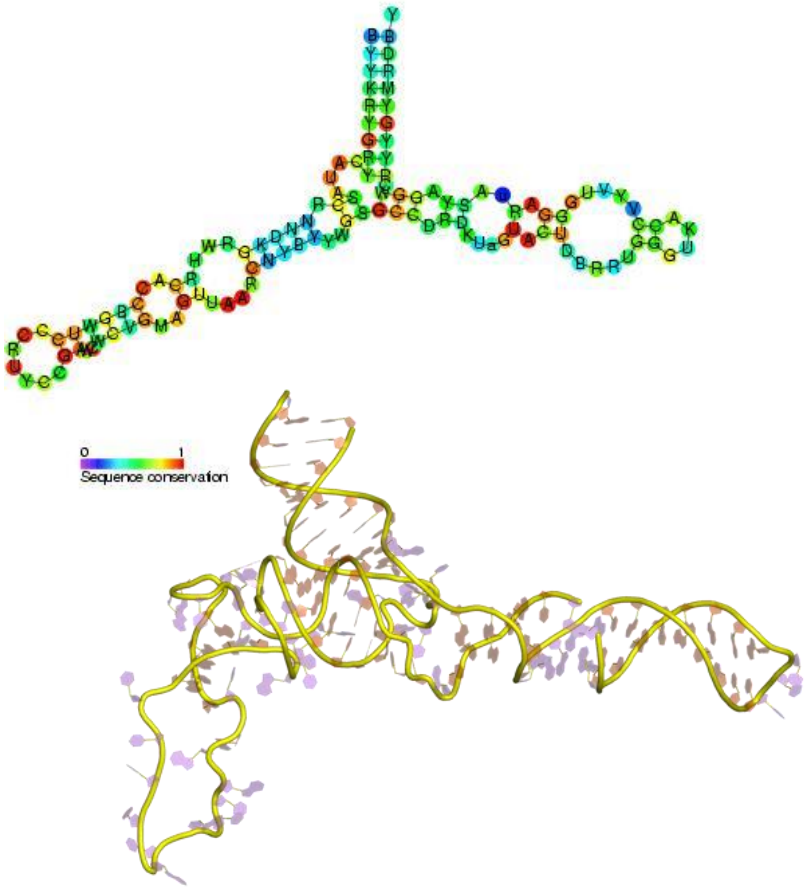
- La structure secondaire peut être vue comme une structure « intermédiaire » entre la structure primaire et la structure tertiaire (3D, coordonnées des atomes dans l'espace)

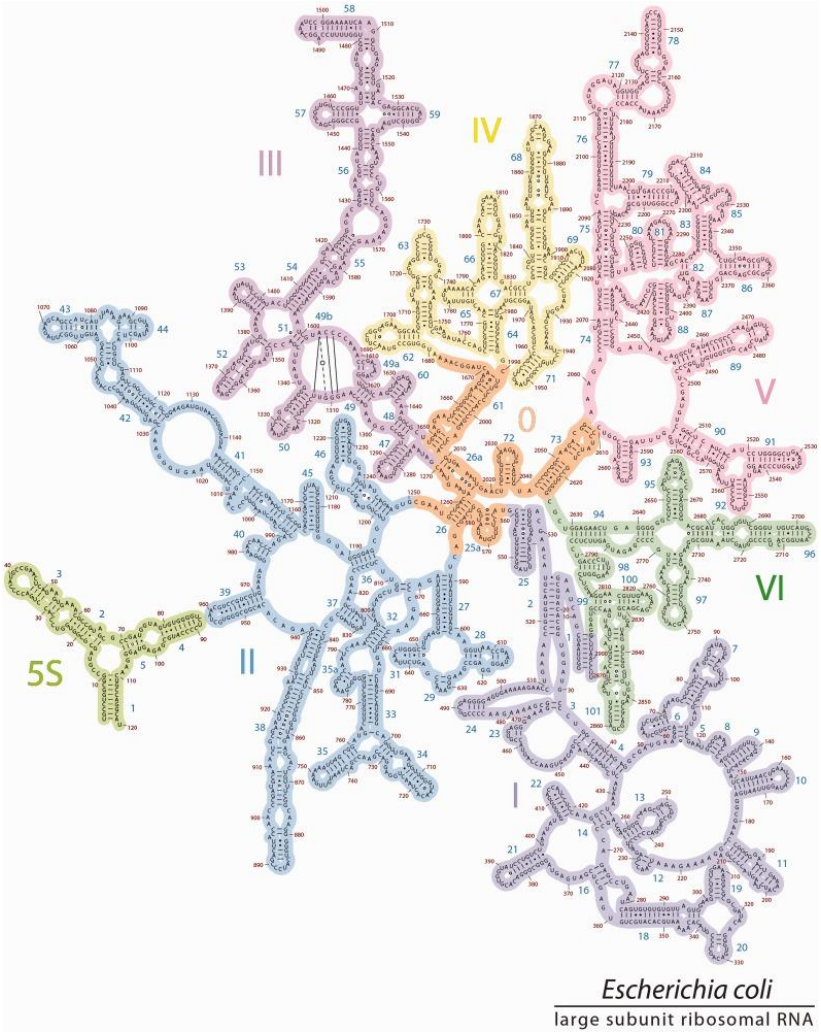


http://www.mpi-inf.mpg.de/departments/d1/projects/CompBio/align.html

D’autres structures secondaires d’ARN

- ARN ribosomique 5S / 23S





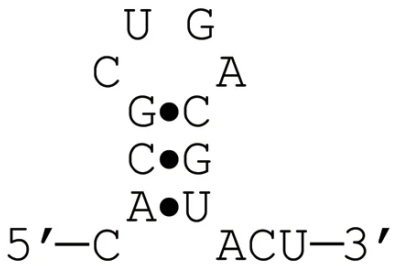
https://fr.wikipedia.org/wiki/ARN_ribosomique_5S

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3753638/

II. Représentation formelle

Définition « théorique » d'une structure sec. d'ARN

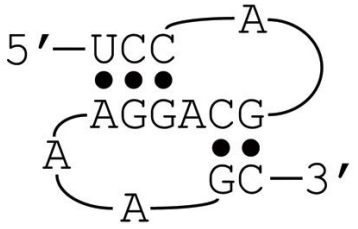
- Pairages canoniques: Watson-Crick (G-C, A-U); non-canonique le plus fréquent: Wobble (G-U).
- Chaque base ne peut être appariée qu'à une seule autre base;
- Pas de « sharp turn »: au moins 3 bases par boucle;



(a) hairpin loop

(b) arc representation of (a)

- Pas de pseudo-nœuds

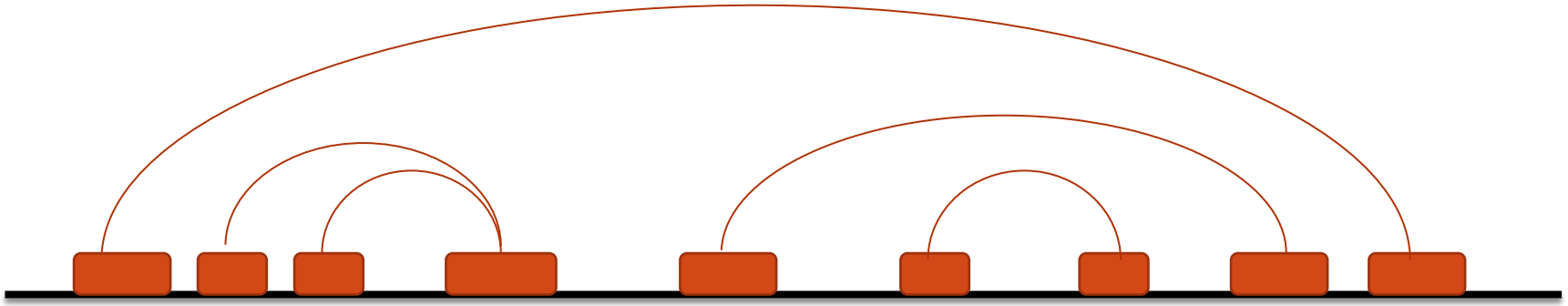


(c) pseudoknot

(d) arc representation of (c)

https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-10-S1-S38/figures/1

II. Représentation formelle d'une structure secondaire d'ARN



Expression bien parenthésée:

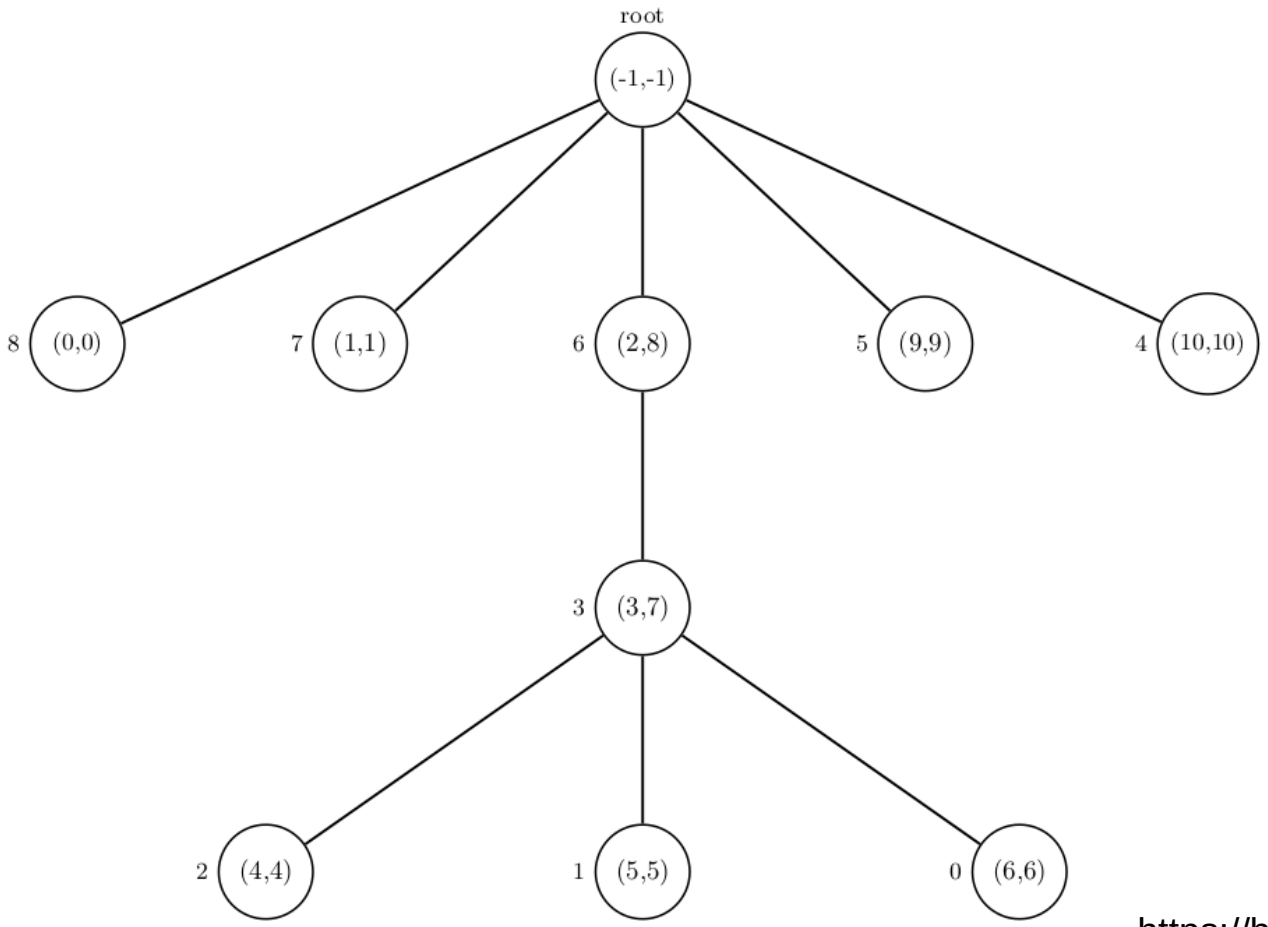
ACGGAGUGCUAAUGGACGUAUUGCCUCAGCUACUGCAAGGUAUACUGU
.(((.((.((....)))).....(((......((.....)).....))))).)))..

Input à des algos permettant de dessiner les structures 2D

VARNA : Interactive drawing and editing of the RNA secondary structure,
Darty, Kévin and Denise, Alain and Ponty,Yann, Bioinformatics, 2009

Représentation sous forme d'arbre

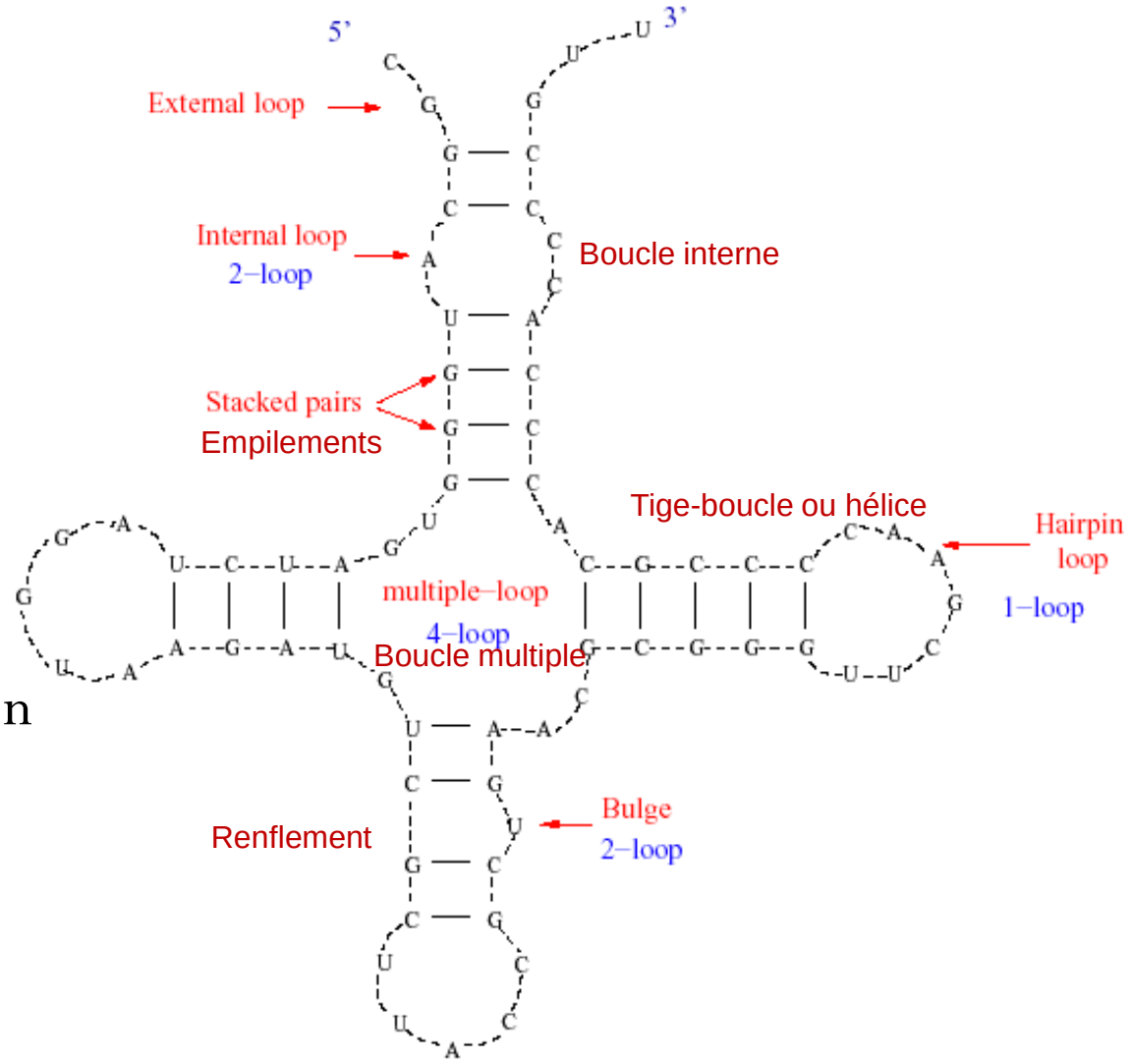
Représentation de la structure ..((…))..



https://bioinfo-fr.net/introduction-a-la-structure-secondaire-des-arn#VARNA

II. Représentation formelle d'une structure secondaire d'ARN

- 1-boucle: Boucle
- 2-boucle:
  - Empilement
  - Boucle interne
  - Bulge
- n-boucle: Boucle multiple, i.e. fermée par n appariements.

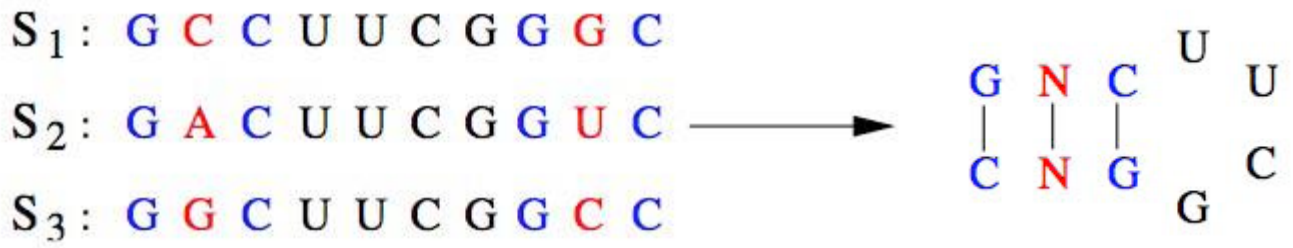


IV. Prédiction de structures secondaires.

1. Co-variance

- Une séquence peut se replier d'une multitude de façons. Par exemple un ARN de 200 bases peut se replier de plus de 10^{50} façons différentes.

- Alignement multiple de séquences homologues: permet de retrouver les sites de co-variance (Woese & Pace, 1993)



IV. Prédiction de structures secondaires.

1. Co-variance

- Cercle vicieux:
  - Inférer la structure correcte nécessite un alignement multiple correct;
  - Inférer un alignement multiple correct nécessite de connaître le repliement correct.
- Solution pour inférer la structure correcte: Processus itératif:
  - Deviner la structure en fonction du meilleur alignement multiple courant
  - Améliorer l'alignement multiple en fonction des nouvelles contraintes de structure prédites.
- Les séquences comparées doivent être suffisamment semblables pour pouvoir être alignées, mais suffisamment divergentes pour pouvoir identifier des contraintes de covariance.

IV. Prédiction

2. Considérations thermodynamiques

- Les ARN ne se replient pas dans des structures aléatoires
- En général, ils préfèrent les conformations de basses énergies
  - En mécanique statistique, les conformations de basses énergies correspondent à celles qui sont souvent observées.
- Le calcul exact de l'énergie E pour une structure arbitraire S est impossible. En pratique:
  - Évaluation expérimentale de l'énergie libre associée à de petits ARN synthétiques (boucles élémentaires);
  - Utiliser l'hypothèse de Tinoco-Uhlenbeck pour évaluer l'énergie libre d'une structure complexe S formée par une suite de boucles.

IV. Prédiction

2. Considérations thermodynamiques

Hypothèse de Tinoco-Uhlenbeck

- Soit S une structure, et s_1, s_2, … s_n la suite des boucles formant S. Alors:

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

$e(s_i)$, pour $1 \leq i \leq n$, estimées expérimentalement.

Dépend du type de boucle.

2. Caractéristiques générales de l'énergie libre

- e(s) est négatif si et seulement si s est un empilement: seules boucles qui contribuent à la stabilité de la molécule.
- Un empilement (G-C)(G-C) plus stable qu'un empilement (A-U)(A-U).
- Les zones externes non appariées ne font partie d'aucun cycle → score nul.
- Si E(S)>0, alors S ne peut pas être stable.

Énergie d'empilement de paires de base - Liaisons de van der Waals (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)

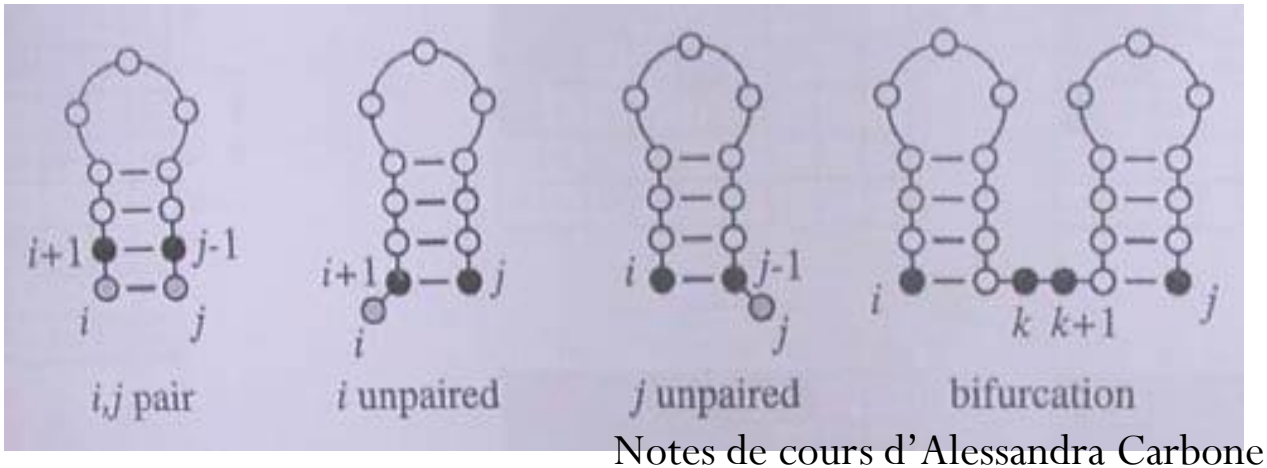
IV. Prédiction de structures secondaires par minimisation de l'énergie libre

- Algorithmes combinatoires:
  - Pipas & McMahon 1975
- Algorithmes récursifs (Prog. dynamique):
  - Nussinov 1978
  - Zuker & Steigler 1981,
  - Zuker & Sankoff 1984,
  - Sankoff 1985
  - … → $O(n^3)$ en espace, $O(n^2)$ en temps.
  - Logiciels:
    - ViennaRNA software: Schuster et al. 1994,
    - MFOLD software: Zuker et al. 1989

IV Prédiction

Algorithme de Nussinov (1978)

- Premier algorithme de programmation dynamique pour calculer le repliement d'une séquence d'ARN qui maximise le nombre de paires de bases créées.
- Problématique: Étant donnée une séquence S[1,n], trouver le repliement de S qui maximise le nombre de paires de bases.
- Idée: Il y a 4 façons de calculer la meilleure structure de S[i,j], pour $1 \leq i \leq j \leq n$, à partir de facteurs de S[i,j]:

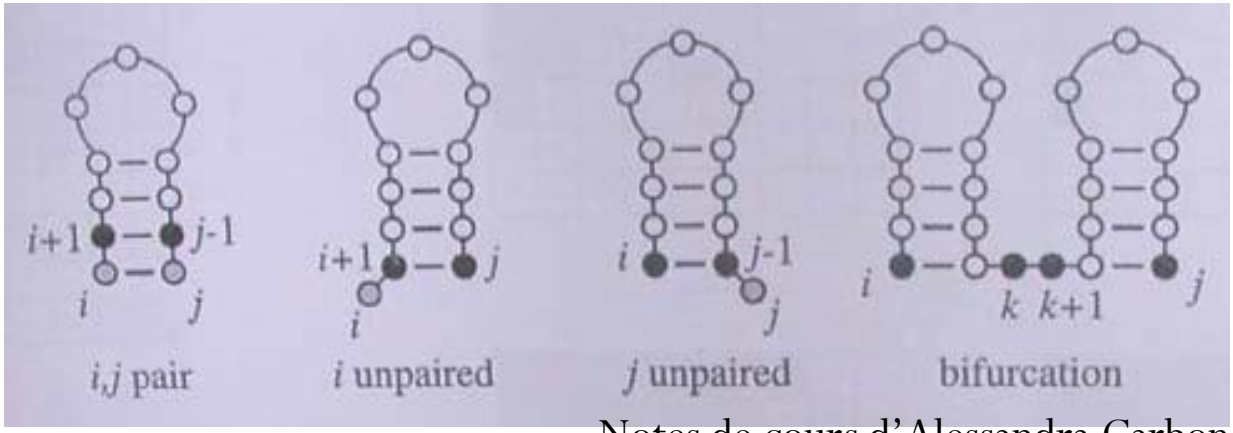


Notes de cours d'Alessandra Carbone

IV Prédiction

Algorithme de Nussinov (1978)

- (i,j) forment un appariement qu'on ajoute à la meilleure structure pour la séquence S[i+1, j-1]
- i est non-apparié et on l'ajoute à la meilleure structure pour la séquence S[i+1, j]
- j est non-apparié et on l'ajoute à la meilleure structure pour la séquence S[i, j-1]
- Combiner deux structures optimales, l'une pour la séquence S[i,k], et l'autre pour la séquence S[k+1,j], pour $i+1 \leq k \leq j-1$

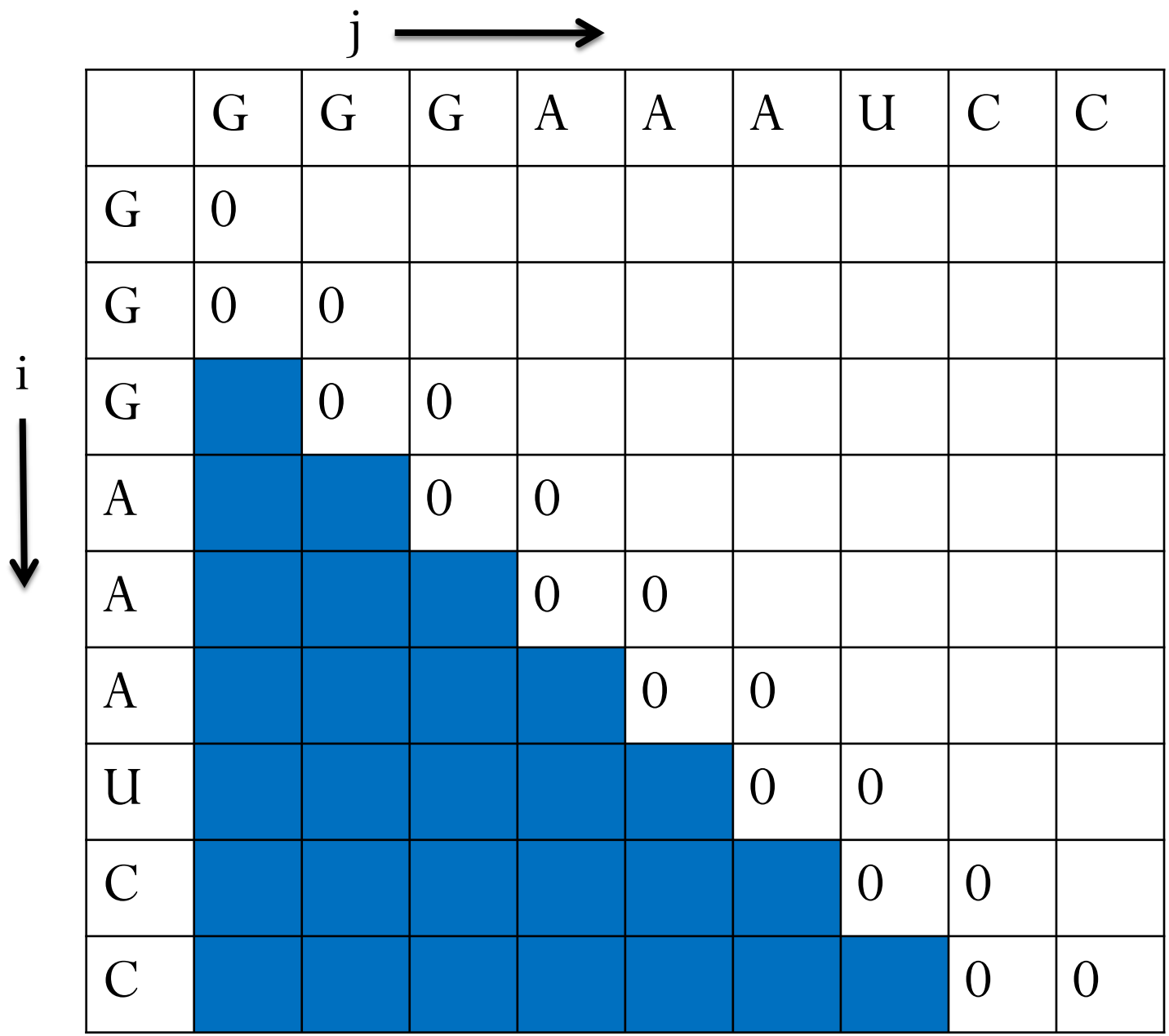


Notes de cours d'Alessandra Carbone

IV Prédiction

Algorithme de Nussinov (1978)

- D(i,j): nombre max. de paires de bases qui peuvent être formées pour la sous-séquence S[i,j].
- Remplir une Table de programmation dynamique D de n lignes et n colonnes:
- Cas de base:
  - $D(i,i) = 0$, pour $1 \leq i \leq n$
  - $D(i,i-1) = 0$ pour $2 \leq i \leq n$


j
i
G G G A A A U C C
G 0
G 0 0
G 0 0
A 0 0
A 0 0
A 0 0
U 0 0
C 0 0
C 0 0

IV Prédiction

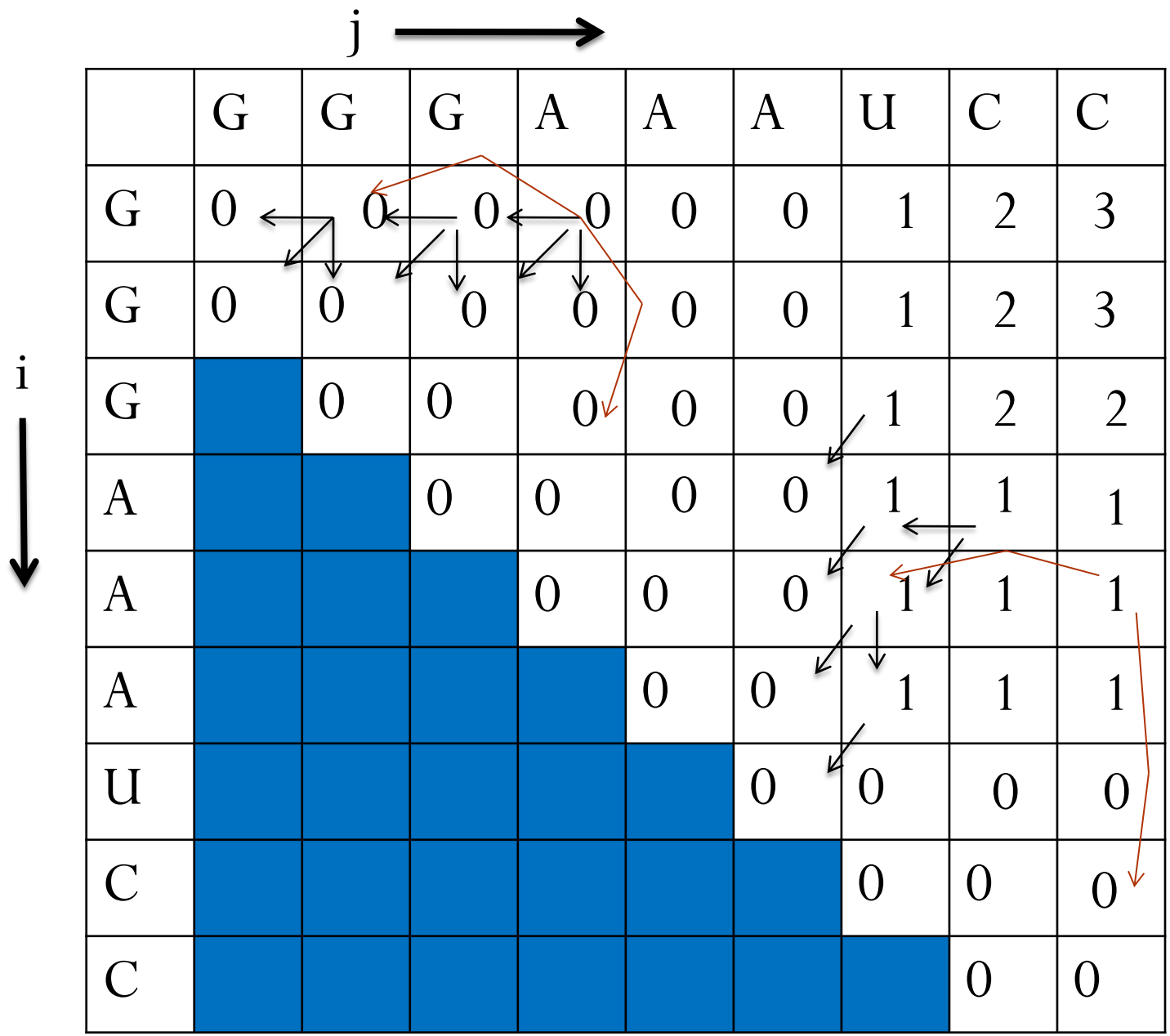
Algorithme de Nussinov (1978)

- Remplir D diagonale par diagonale, pour 0<i<j, par la relation de récurrence

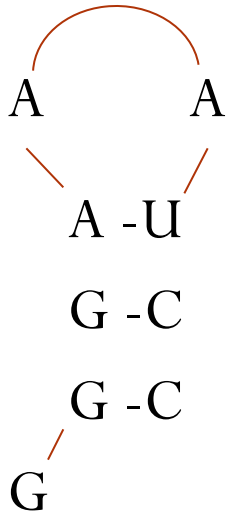
$$D(i,j) = \max \{ D(i+1,j),\ D(i,j-1),\ D(i+1,j-1)+p(S[i],S[j]),\ \max_{i+1 \le k \le j-2} [D(i,k)+D(k+1,j)] \}$$

où p(S[i],S[j]) = 1 si S[i]-S[j] forme une paire de base et 0 sinon.

		1	2	3	4	5	6	7	j=8	9
		G	G	G	A	A	A	U	C	C
1	G	0								
i=2	G	0	0							
3	G		0	0						
4	A			0	0					
5	A				0	0				
6	A					0	0			
7	U						0	0		
8	C							0	0	
9	C								0	0


j
i
G G G A A A U C C
G 0 0 0 0 0 0 1 2 3
G 0 0 0 0 0 0 1 2 3
G 0 0 0 0 0 1 2 2
A 0 0 0 0 1 1 1
A 0 0 0 0 1 1 1
A 0 0 1 1 1
U 0 0 0 0
C 0 0 0
C 0 0

	G	G	G	A	A	A	U	C	C
G	0	0	0	0	0	0	1	2	3
G	0	0	0	0	0	0	1	2	3
G		0	0	0	0	0	1	2	2
A			0	0	0	0	1	1	1
A				0	0	0	1	1	1
A					0	0	1	1	1
U						0	0	0	0
C							0	0	0
C								0	0



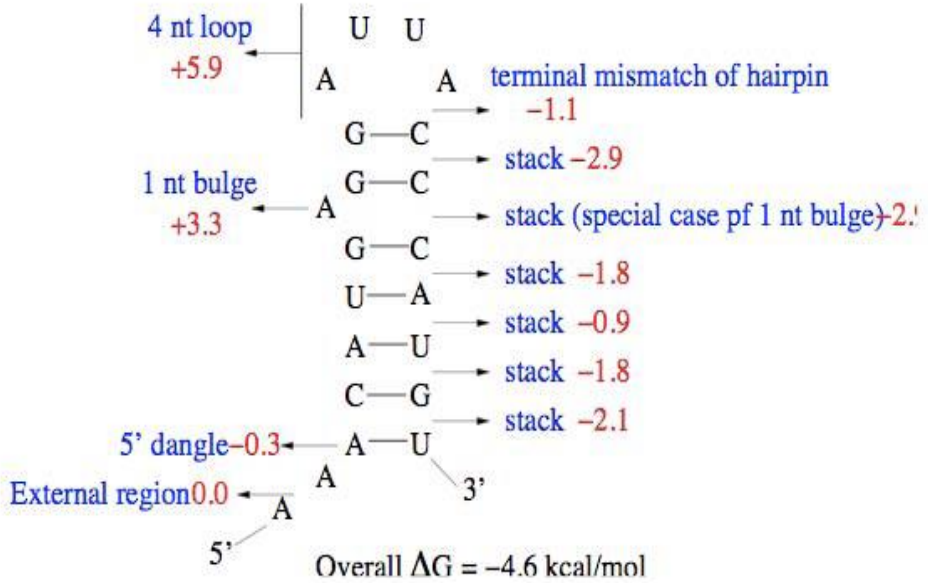
IV Prédiction

Algorithme de Nussinov (1978)

- Le « backtracking » est linéaire en temps et en espace;
- Le remplissage de la table est en $O(n^2)$ en espace et en $O(n^3)$ en temps.

IV Prédiction

Algorithme de Zuker (1981- 1989)



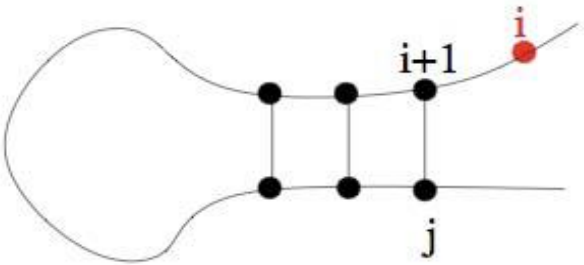
Freir rules (1986) at 37°C. Example from *Durbin, Eddy et. al.* book.

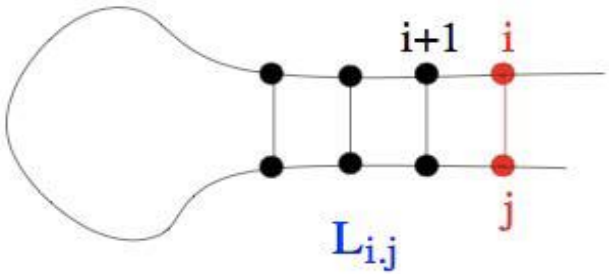
- Une différence importante avec l'algorithme de Nussinov est que l'énergie d'une paire de bases est calculée en fonction de la paire de bases précédente (stacking)

IV Prédiction

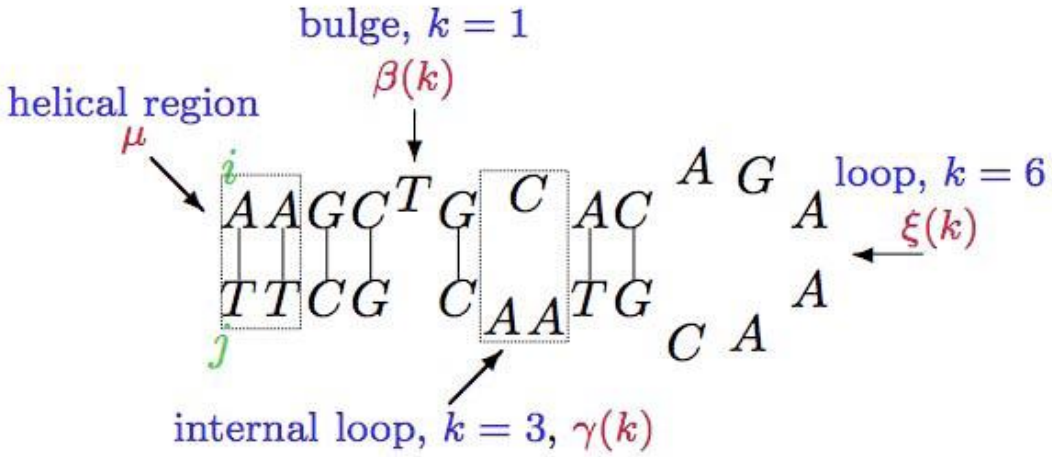
Algorithme de Zuker (1981- 1989)

$$E(S_{i,j}) = \min \begin{cases} E(S_{i+1,j}) \\ E(S_{i,j-1}) \\ \min_{i<k<j}\{E(S_{i,k}) + E(S_{k+1,j})\} \\ E(L_{i,j}) \end{cases}$$





Algorithme de Zuker



$E(L_{i,j})$:

$$\alpha(i,j) + \xi(j-i-1) \quad \text{loop}$$

$$\alpha(i,j) + \mu + E(S_{i+1,j-1}) \quad \text{helical region}$$

$$\min_{k \geq 1}\{\alpha(i,j) + \beta(k) + E(S_{i+k+1,j-1})\} \quad \text{bulge at } i$$

$$\min_{k \geq 1}\{\alpha(i,j) + \beta(k) + E(S_{i+1,j-k-1})\} \quad \text{bulge at } j$$

$$\min_{k_1,k_2 \geq 1}\{\alpha(i,j) + \gamma(k_1+k_2) + E(S_{i+1+k_1,j-1-k_2})\} \quad \text{internal loop}$$

$$\min_{i+1<k<j-2}\{E(i+1,k) + E(k+1,j-1)\} \quad \text{multiple loop}$$

		1	2	3	4	5	6	7	j=8	9
		G	G	G	A	A	A	U	C	C
1	G	0								
i=2	G	0	0						α	
3	G		0	0			β	μ		
4	A			0	0		γ	β		
5	A				0	0				
6	A					0	0			
7	U						0	0		
8	C							0	0	
9	C								0	0

IV Prédiction

Algorithme de Zuker (1981- 1989)

- Complexité:
  - Épingles à cheveux et empilements: Temps constant pour chaque (i,j) → $O(n^2)$
  - Bulge: O(n) pour chaque (i,j) → $O(n^3)$
  - Boucles internes: $O(n^2)$ pour chaque (i,j) → $O(n^4)$
- Les algorithmes ont été étendus pour retrouver des structures sous-optimales.

Références

- *Time warps, string edits, and macromolecules – The theory and practice of sequence comparison*, David Sankoff and Joseph Kruskal, CSLI Publications, 1999. Chapter 3.
- *Biological sequence analysis, Probabilistic models of proteins and nucleic acids*, R. Durbin, S. Eddy, A. Krogh and G. Mitchison, Cambridge 1998. Chapter 10.