Journées PARISTIC
Bordeaux 2005

Bioinformatique / Bioalgorithmique de l'ARN

Alain Denise
Bioinformatique
LRI Orsay
UMR CNRS 8623
Université Paris-Sud 11

Le dogme central

Réplication

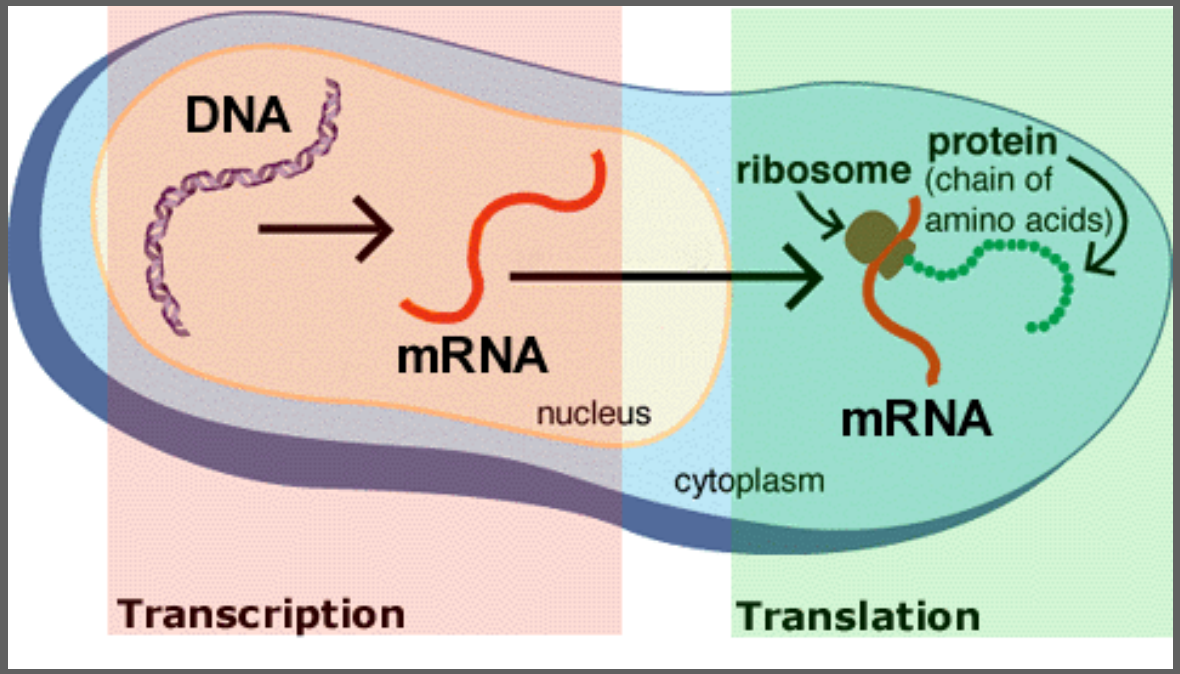
ADN

Transcription

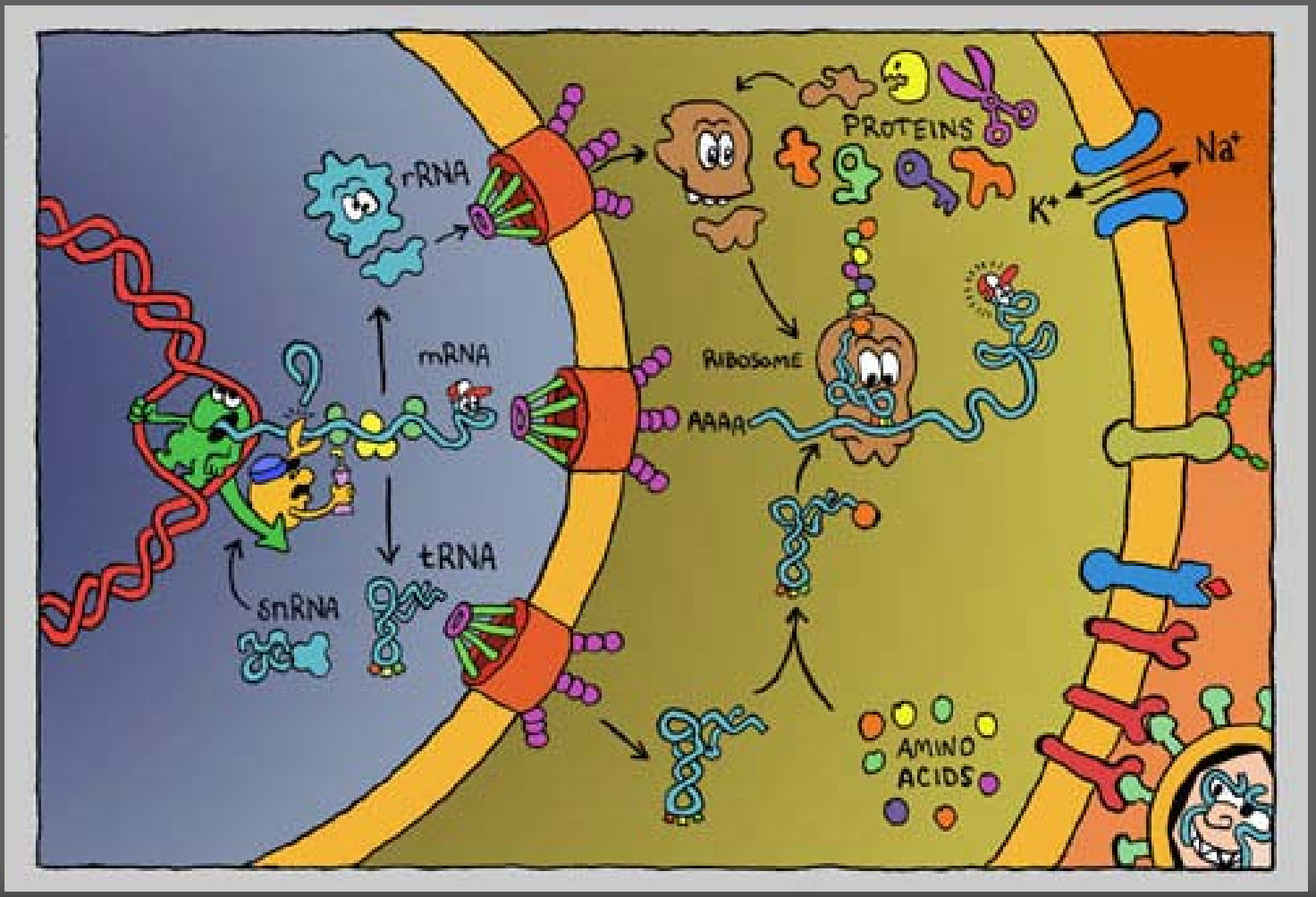
ARN

Traduction

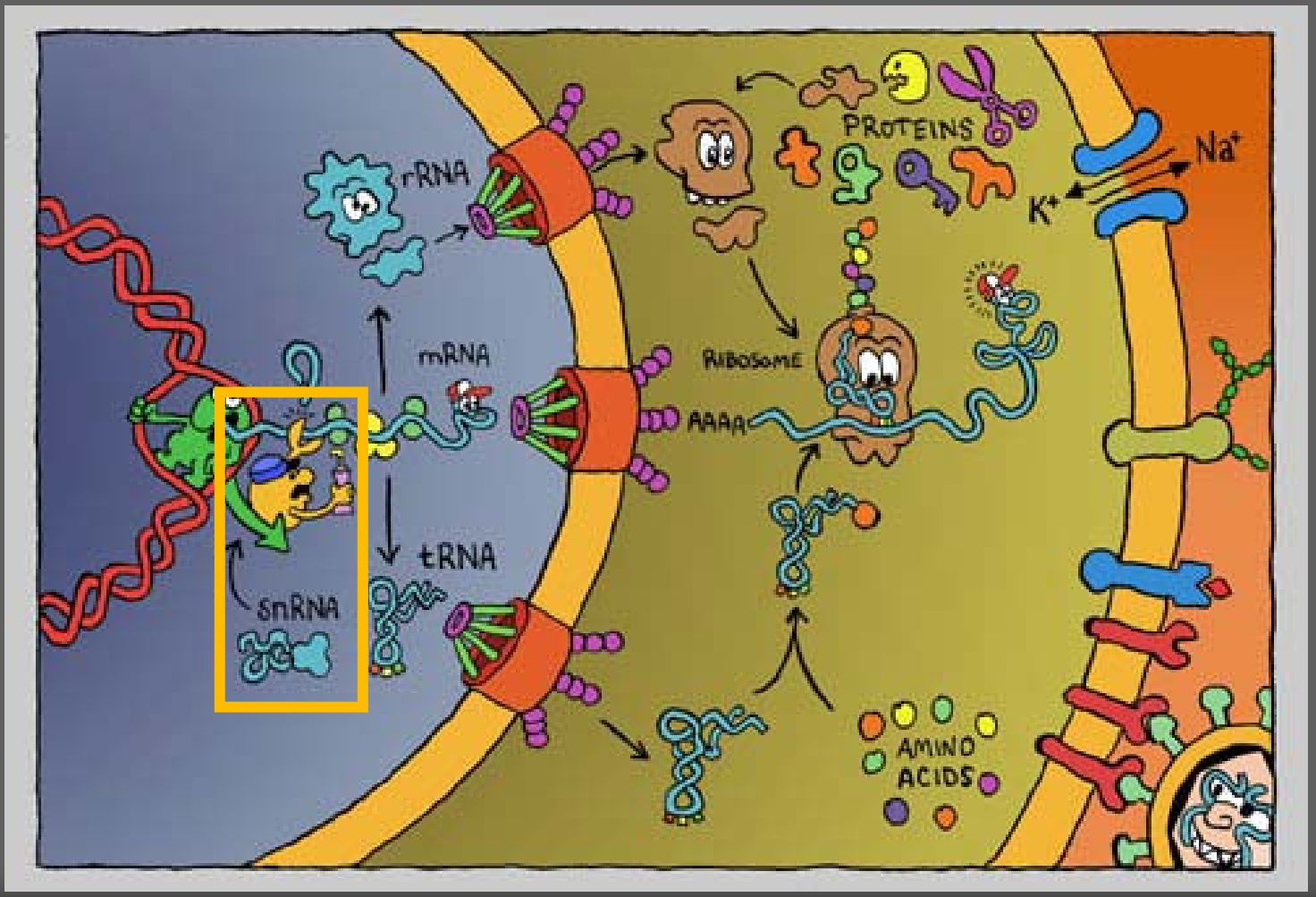
Protéines



Les multiples rôles de l'ARN



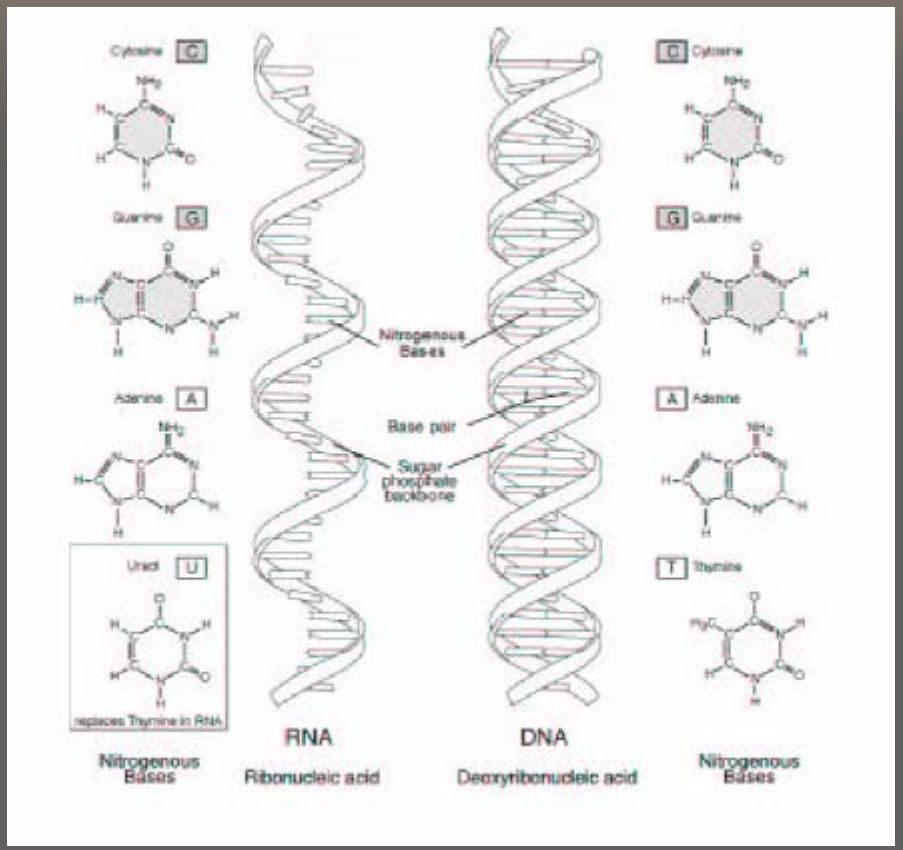
Les multiples rôles de l'ARN

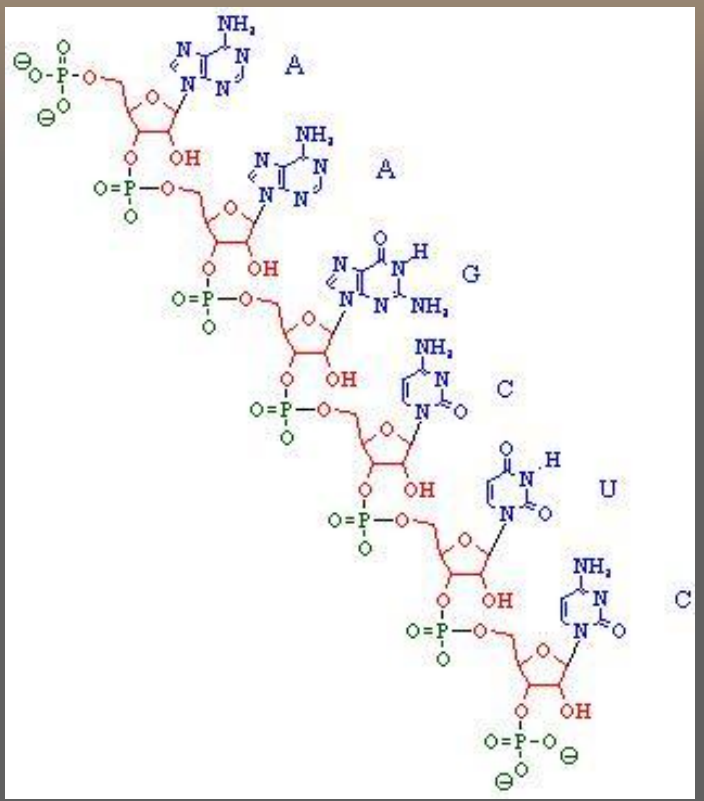


Importance de l'ARN

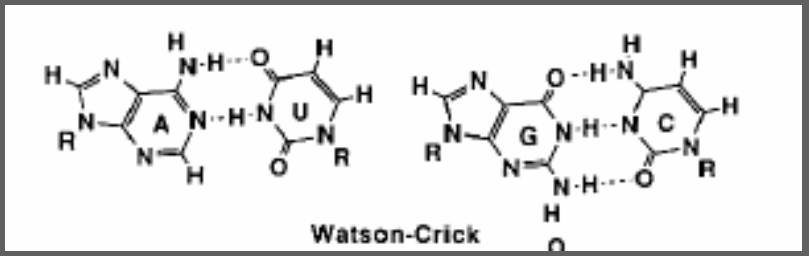
- Présente dans tous les processus cellulaires
- La seule molécule qui peut être génome aussi bien que catalyseur
- Origine de la vie : le monde à ARN
- Cible très fréquente des antibiotiques

L'ARN

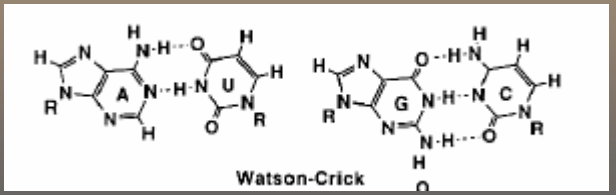

Cytosine C
Guanine G
Adenine A
Uracil U
replaces Thymine in RNA
Nitrogenous Bases
Nitrogenous Bases
Base pair
Sugar phosphate backbone
RNA
Ribonucleic acid
DNA
Deoxyribonucleic acid
C Cytosine
G Guanine
A Adenine
T Thymine
Nitrogenous Bases


A
A
G
C
U
C

…AAGCUC…


A
U
G
C
Watson-Crick

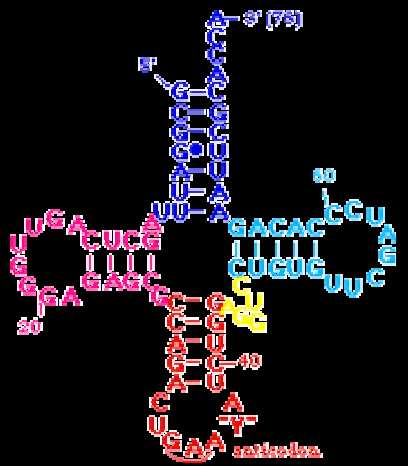
Structure de l'ARN



- Structure primaire

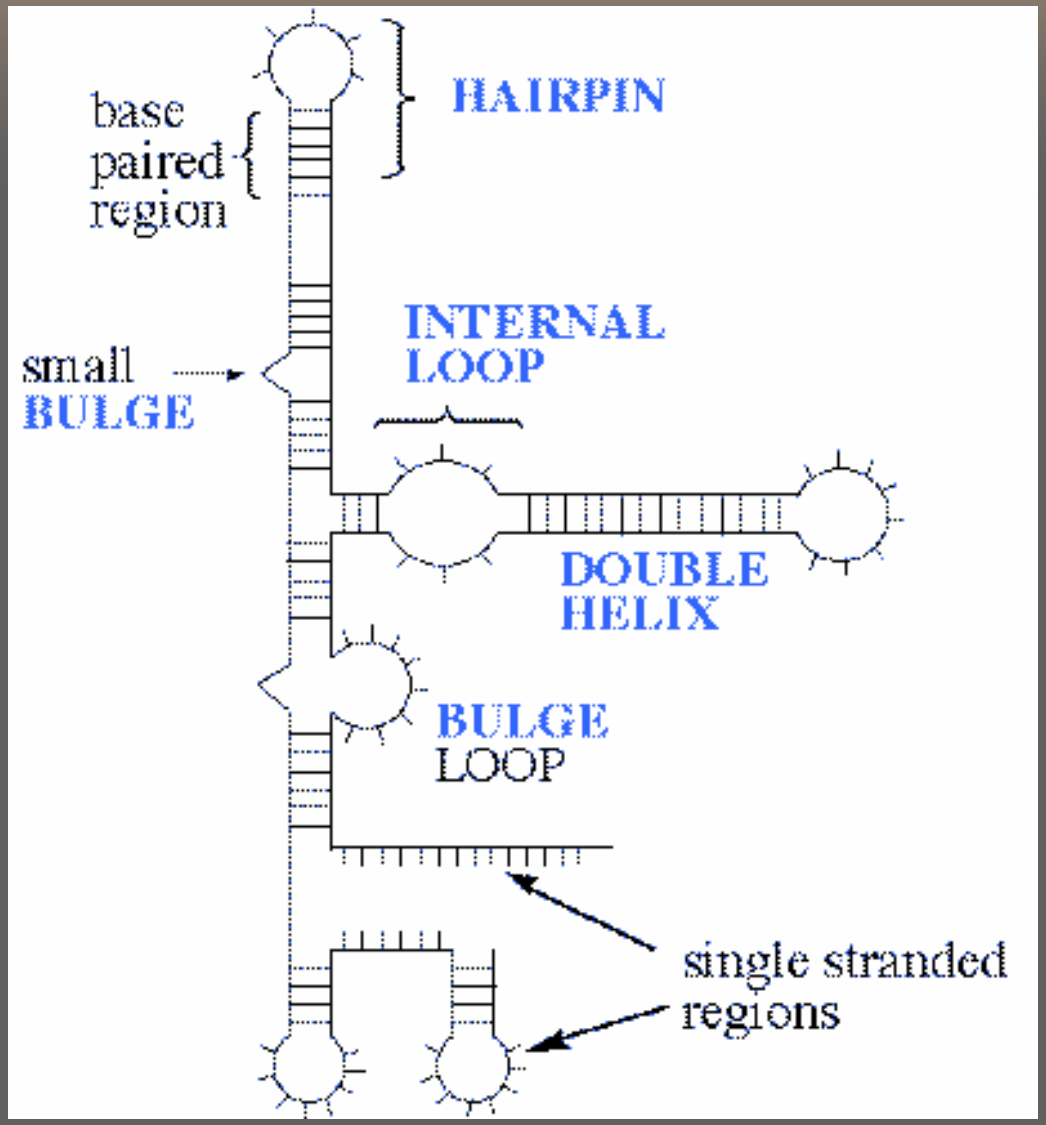
GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAUAUCUGGAGGUCCUGUGUUCGAUCCCACAGAAUUCGCACCA

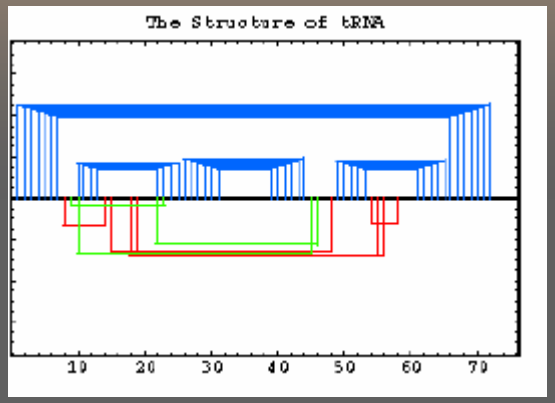
- Structure secondaire

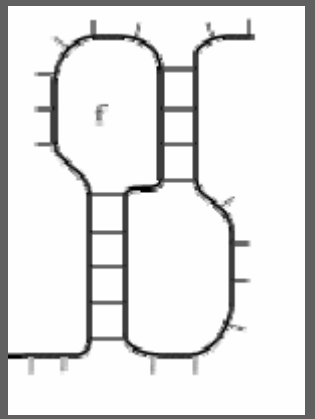


- Structure tertiaire

Structures secondaires





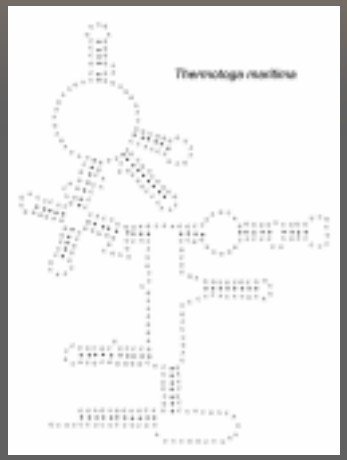


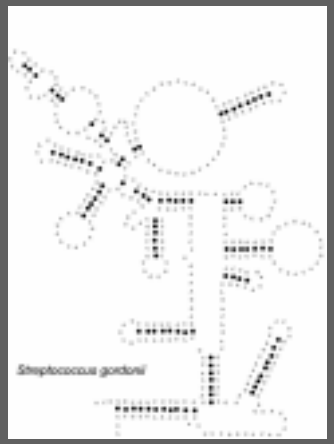
Pseudo-nœud

« Bio-Algorithmique » de l'ARN

- Prédiction de structure en fonction de la séquence
- Détection de motifs structurels dans une séquence
- Comparaison de deux ou plusieurs structures
- Détermination d'une séquence en fonction de la structure
- Recherche de sous-structures communes à deux ou plusieurs structures

Pourquoi comparer ?





- Phylogénie
- Prédiction de structures
- Recherche de motifs communs

Comparer = Calculer une **distance (ou un score)**

Edition et alignement deux à deux

On se donne un ensemble « d'opérations atomiques », chacune ayant un score (ou un coût).

Données : deux structures.

• Edition : trouver la suite d'opérations de score maximal (ou de coût minimal) permettant de transformer une structure en l'autre.

• Alignement : trouver une « sur-structure » commune aux deux structures telle que la somme des scores d'édition de chacune des structures à la sur-structure soit maximale (ou que la somme des coûts soit minimale).

Comparaison de 2 séquences

Deux séquences $v = v_1 v_2 \ldots v_n$ et $w = w_1 w_2 \ldots w_m$

Opérations d'édition :

- ins(x,i)
- suppr(x,i)
- subs(x,y,i)

CHAT - suppr(C,1) → HAT - subs(H,R,1) → RAT

(Pour les séquences : édition ~ alignement)

Comparaison de 2 séquences

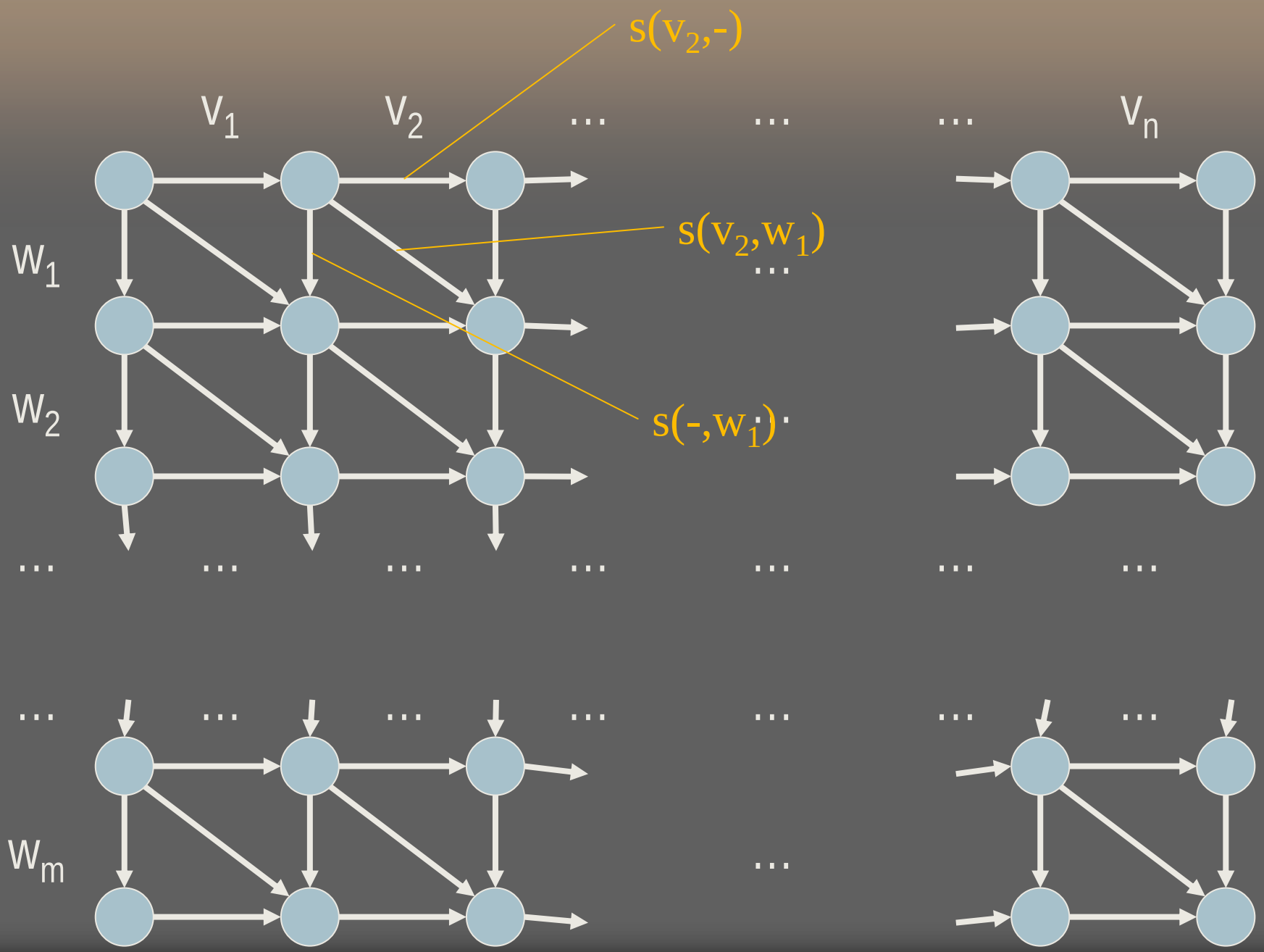
$v = v_1 v_2 \ldots v_n$ $\quad\quad$ $w = w_1 w_2 \ldots w_m$

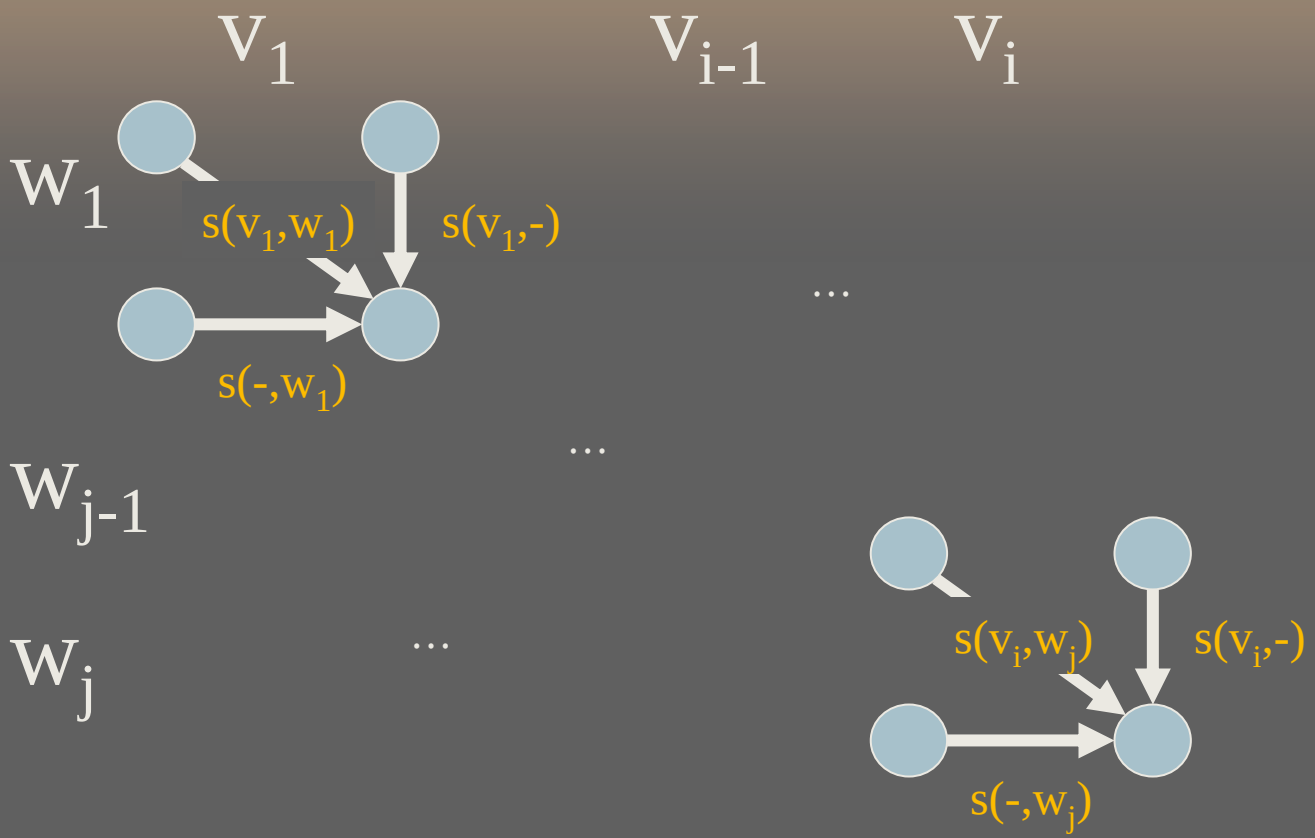
s(x,y) : coût de substitution de x en y
s(x,-) : coût de suppression de x
s(-,y) : coût d'insertion de y

D(v,w) : distance d'édition de v et w

$$D(v_1 \ldots v_i, w_1 \ldots w_j) = \text{Min} \{$$
$$D(v_1 \ldots v_{i-1}, w_1 \ldots w_{j-1}) + s(v_i, w_j)$$
$$D(v_1 \ldots v_{i-1}, w_1 \ldots w_j) + s(v_i, -)$$
$$D(v_1 \ldots v_i, w_1 \ldots w_{j-1}) + s(-, w_j)$$
$$\}$$

Needleman, Wunsch 1970, Gotoh 198


s(v2,-)
V1
V2
...
...
...
Vn
s(v2,w1)
W1
...
W2
s(-,w1)
...
...
...
...
...
...
...
...
...
...
...
...
...
Wm
...

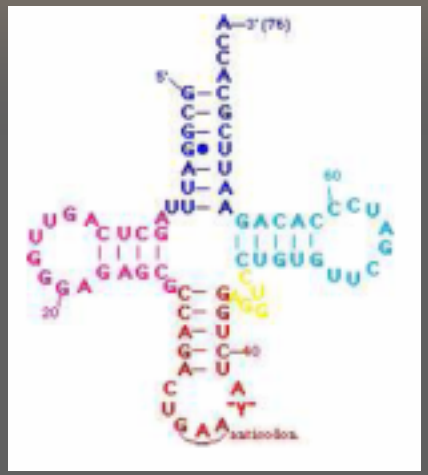


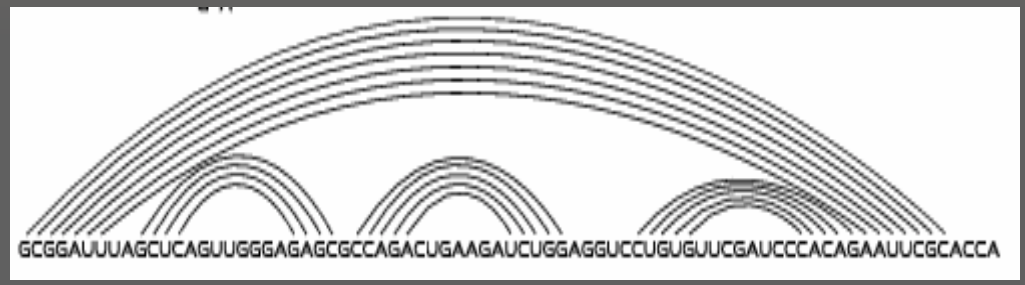
$$D(v_1 \ldots v_i, w_1 \ldots w_j) = \text{Min} \{$$

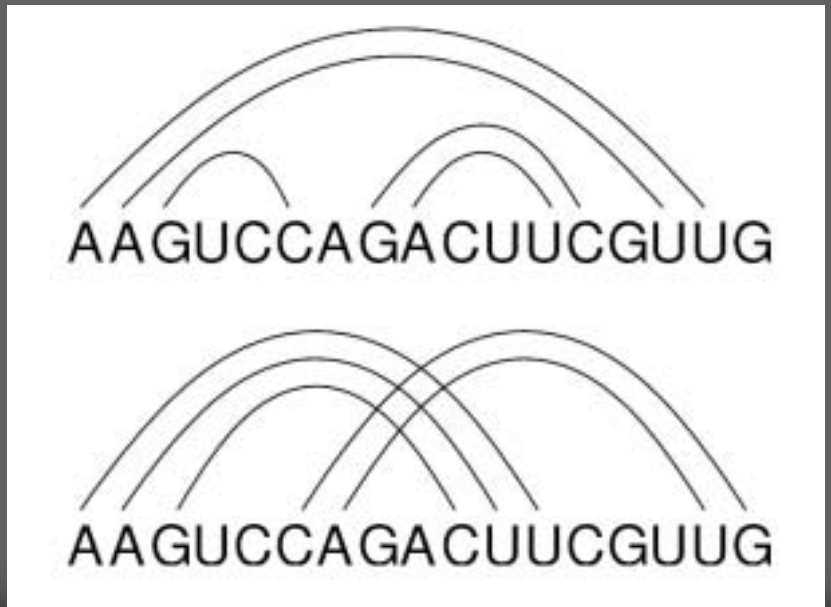
$$D(v_1 \ldots v_{i-1}, w_1 \ldots w_{j-1}) + s(v_i, w_j)$$
$$D(v_1 \ldots v_{i-1}, w_1 \ldots w_j) + s(v_i, -)$$
$$D(v_1 \ldots v_i, w_1 \ldots w_{j-1}) + s(-, w_j)$$

$$\}$$

ARN et séquences arc-annotées

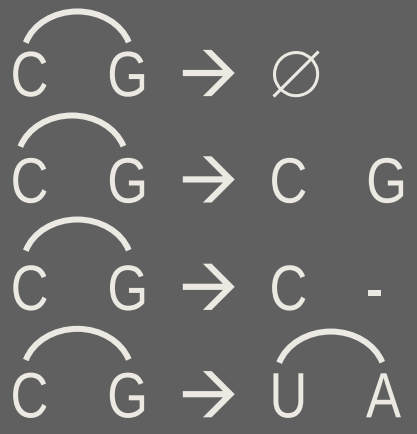




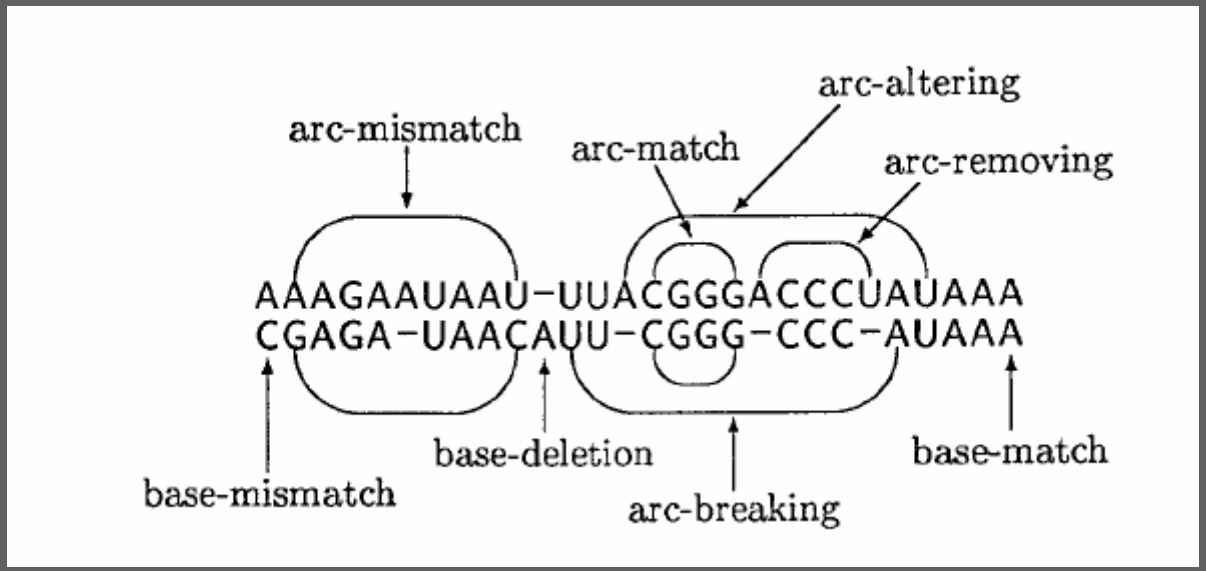


Opérations de séquences arc-annotées

- Opérations sur les bases :
  - Suppression / Insertion
  - Substitution (ou conservation)
- Opérations sur les arcs :
  - Suppression / Insertion :
  - Cassure / :
  - Altération / :
  - Substitution :



Edition de séquences arc-annotées

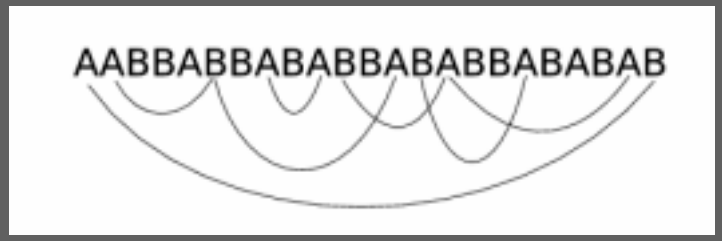


(Jiang, Lin, Ma, Zhang 2002)

Complexité de l'édition

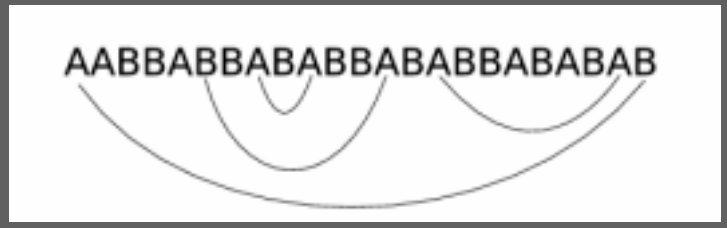
Types de séquences arc-annotées

- Générale



- Croisée

- Imbriquée



- Sans arcs

Complexité de l'édition

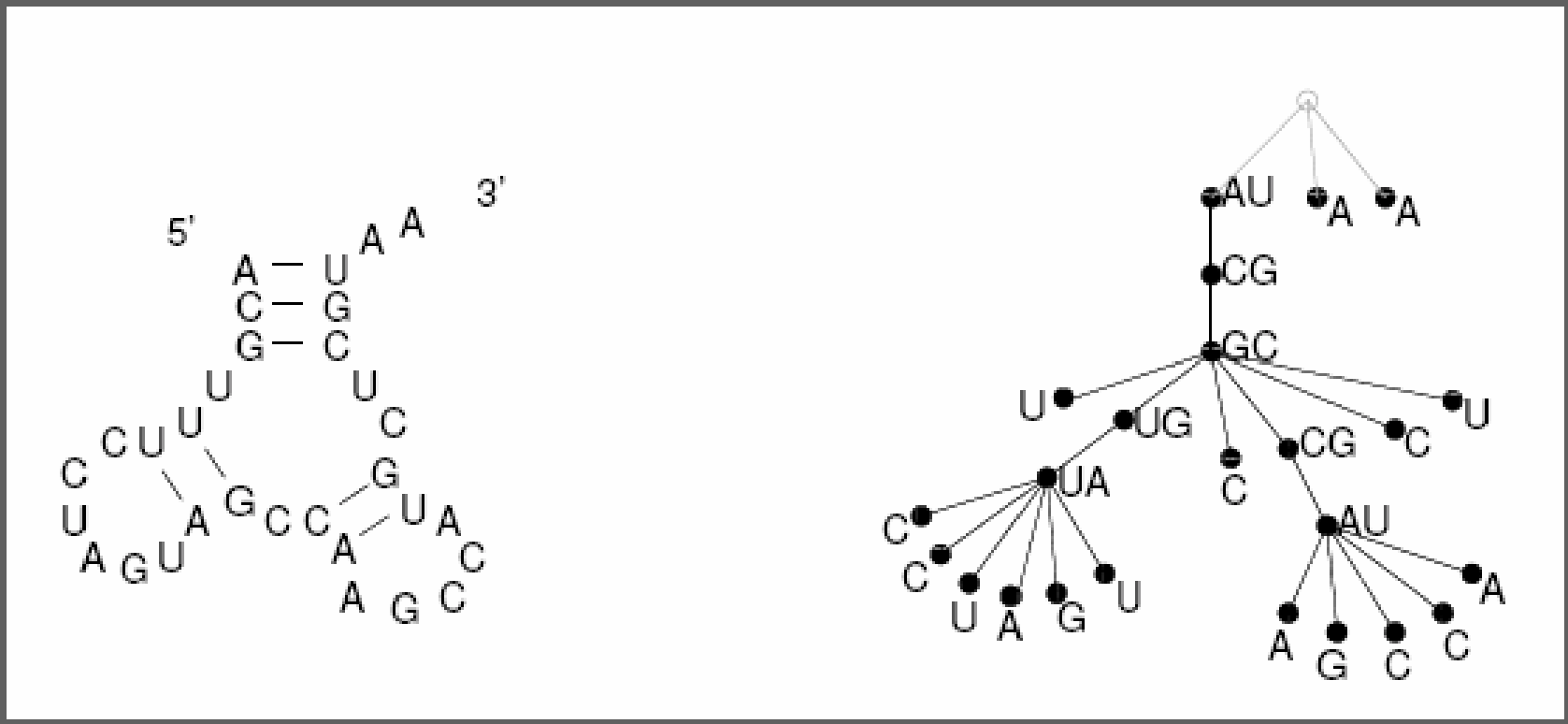
	Générale	Croisée	Imbriquée	Sans arcs
Générale	NP-complet			
Croisée		NP-complet		
Imbriquée			NP-complet	$O(nm^3)$
Sans arcs				O(nm / logn)

Si 2×Score(Altération d'arc) = Score(Cassure) + Score (Suppression), alors algorithme en $O(n^3m)$ pour Edit(croisée,imbriquée) et Edit(imbriquée, imbriquée)

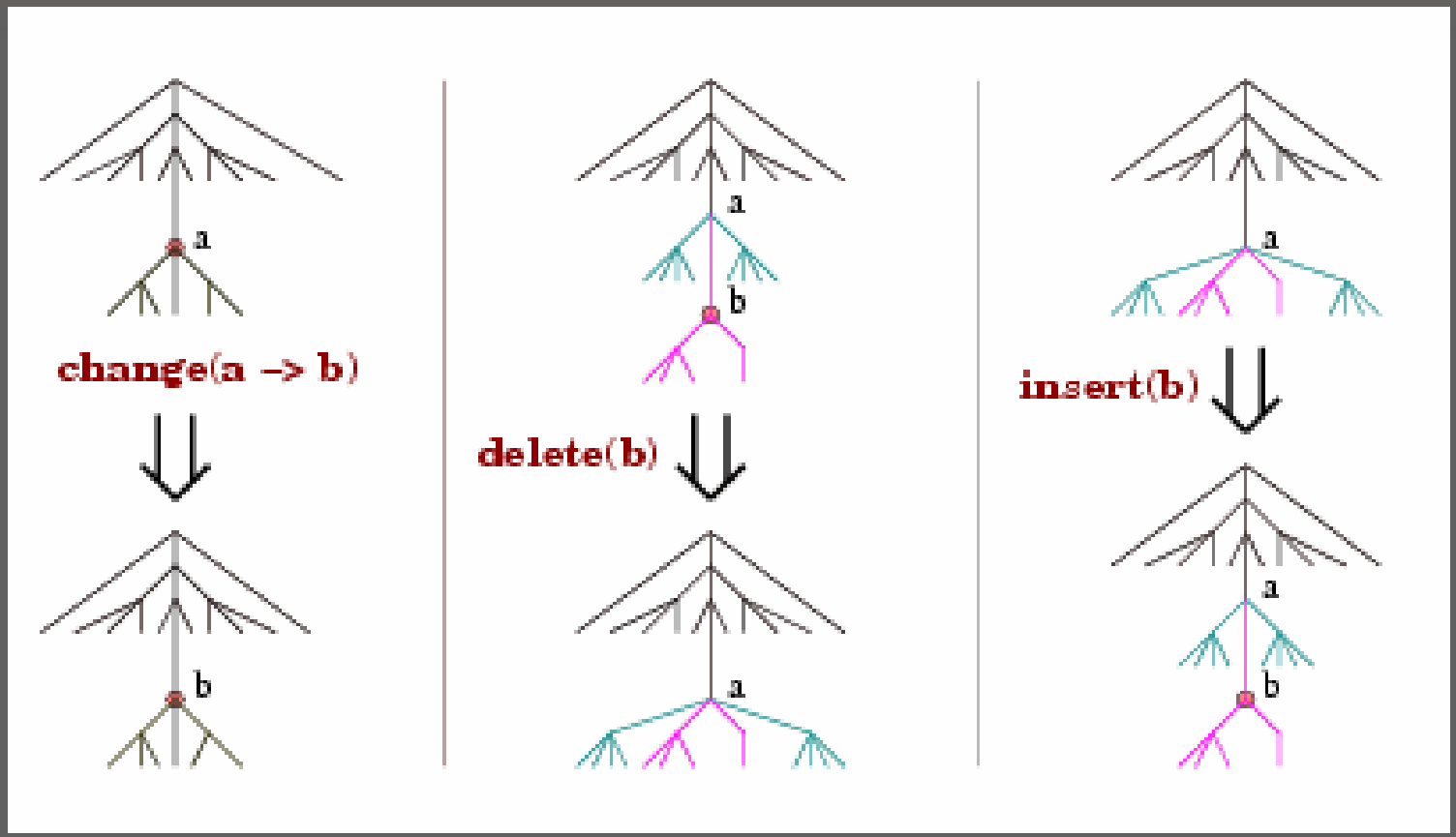
- **Jiang, Lin, Ma, Zhang 2002**
- **Blin, Fertin, Rusu, Sinoquet 2003**
- **Crochemore, Landau, Ziv-Ukelson 200**

Le cas « imbriqué-imbriqué »

→ Structures secondaires
→ Comparaison d'arbres

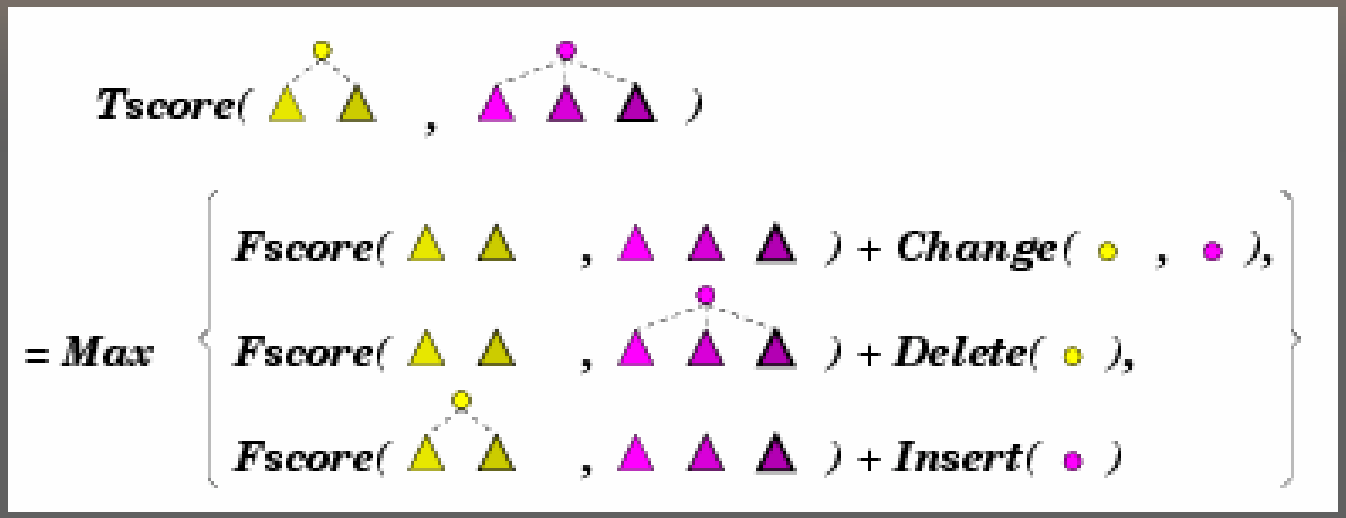


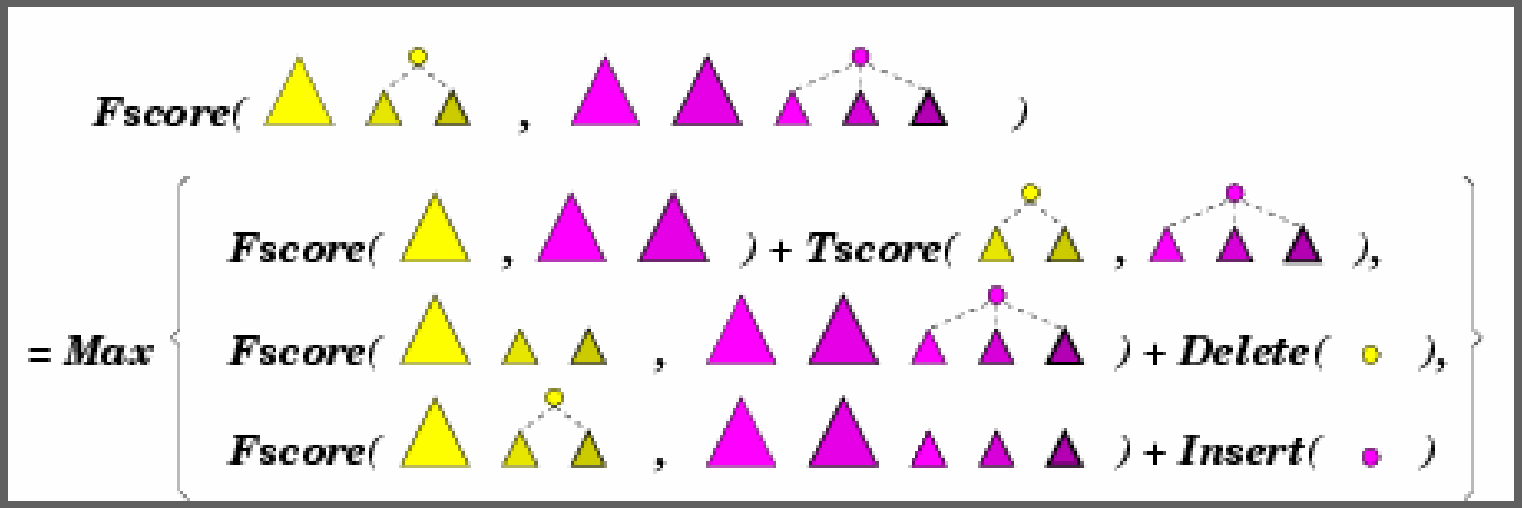
Opérations d'édition



Algorithme d'édition

Zhang, Shasha 1989





Complexité

Edition [Zhang, Shasha 1989, Klein 1998]

- **Au pire :** $O(n^4)$ **[Zhang-Shasha 1989]**
  $O(n^3 \log n)$ **[Klein 1998, Dulucq-Touzet 2003]**
- **En moyenne :** $O(n^3)$ **[Dulucq-Tichit 2003]**

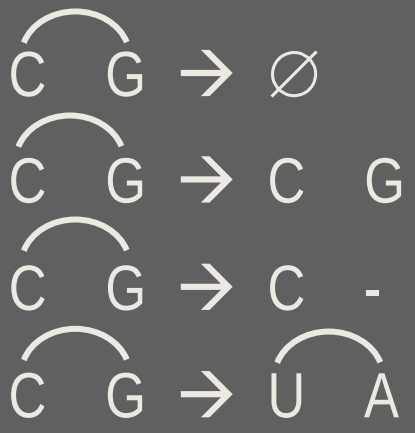
Alignement [Jiang, Wang, Zhang 1995]

- **Au pire :** $O(n^4)$

…mais EDIT(Imbriqué,Imbriqué) est NP-complet !

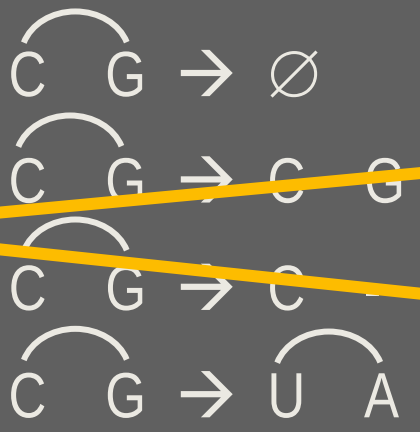
Opérations d'édition

- Opérations sur les bases :
  - Suppression / Insertion
  - Substitution
- Opérations sur les arcs :
  - Suppression / Insertion :
  - Cassure / :
  - Altération / :
  - Substitution :



Opérations d'édition : manques

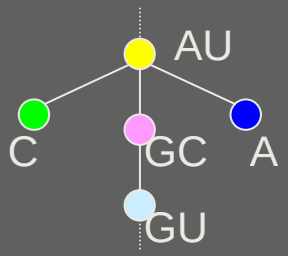
- Opérations sur les bases :
  - Suppression / Insertion
  - Substitution
- Opérations sur les arcs :
  - Suppression / Insertion :
  - ~~Cassure / :~~
  - ~~Altération / :~~
  - Substitution :



Opérations d'édition : problème

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

A-U
C A
G-C
C-U

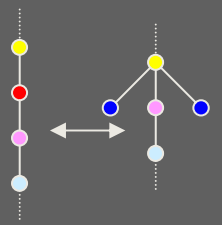


Delete( )
Insert( )
Insert( )

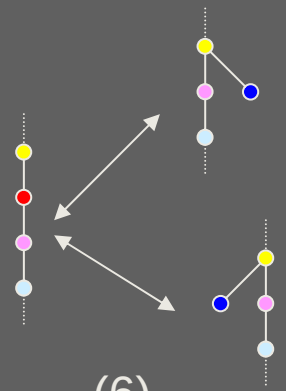
3 opérations au lieu d'une !

Opérations d'édition : ajouts

- Suppression et insertion d'une base
- Substitution de bases
- Suppression et insertion d'une paire de bases
- Substitution de paires de bases
- Appariement et désappariement (5)
- Suppression et insertion d'une base dans une paire de bases (6)



(5)

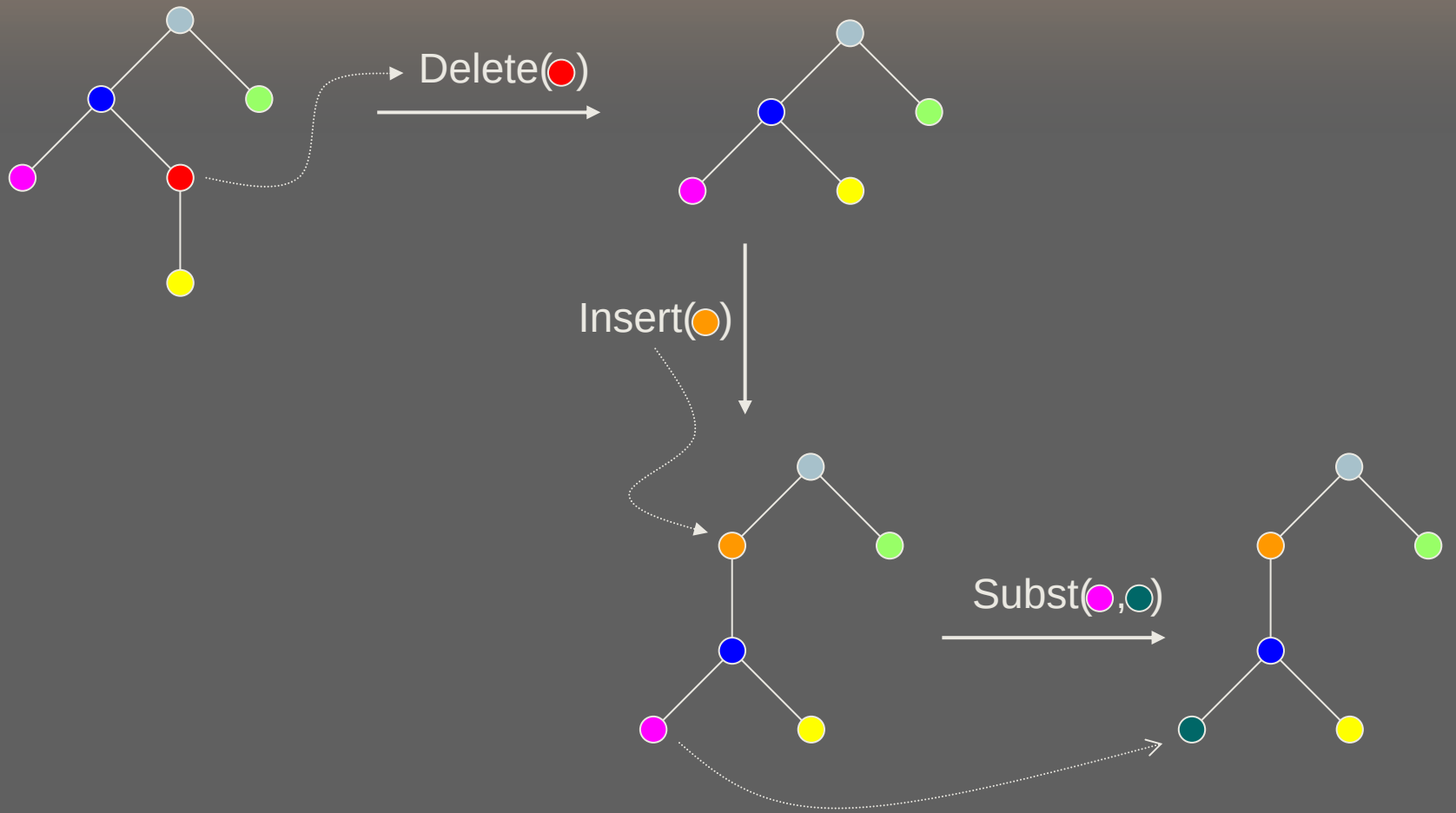


(6)

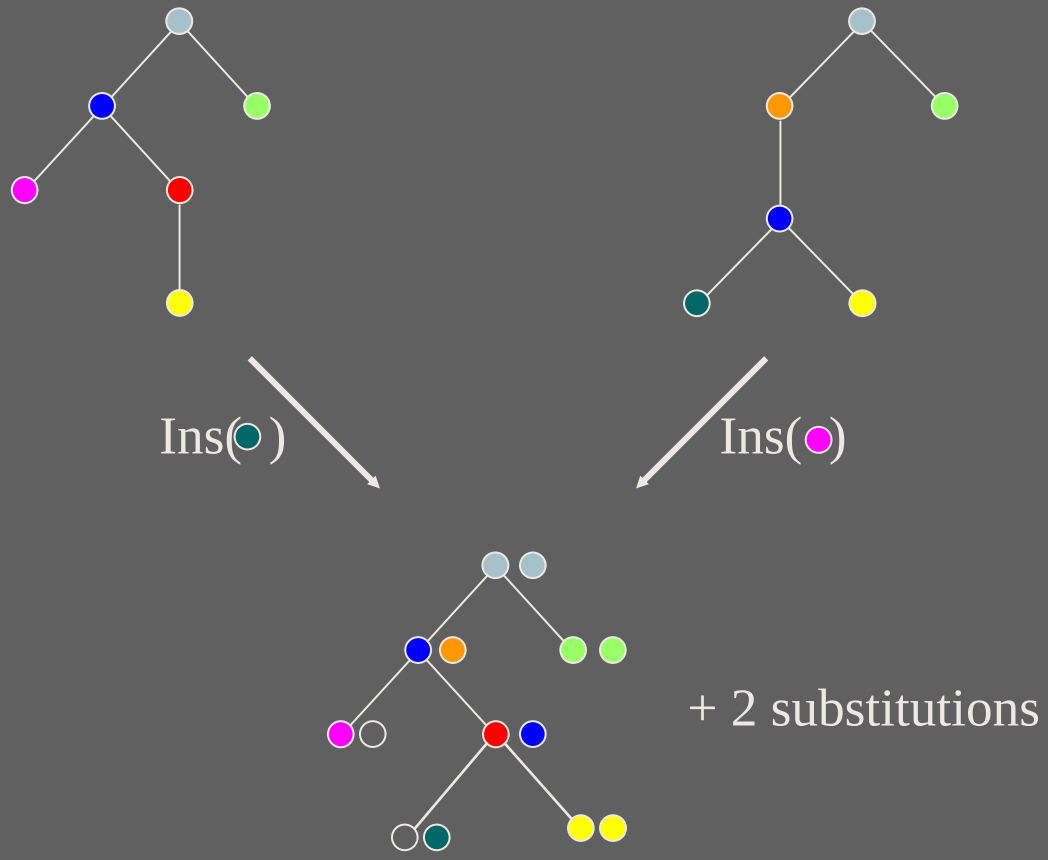
Edition et alignement d'arbres

	Opérations « arbres »	Opérations « ARN »
Edition	$O(n^3 \log n)$ [Zhang-Shasha 1989, Klein 1998]	NP-complet [Blin, Fertin, Sinoquet, Rusu 2003]
Alignement	$O(n^4)$ [Jiang, Wang, Zhang 1995]	?

Edition d'arbres ≠ Alignement



Edition d'arbres ≠ Alignement



Qu'est-ce que l'alignement dans notre cas?

Edition

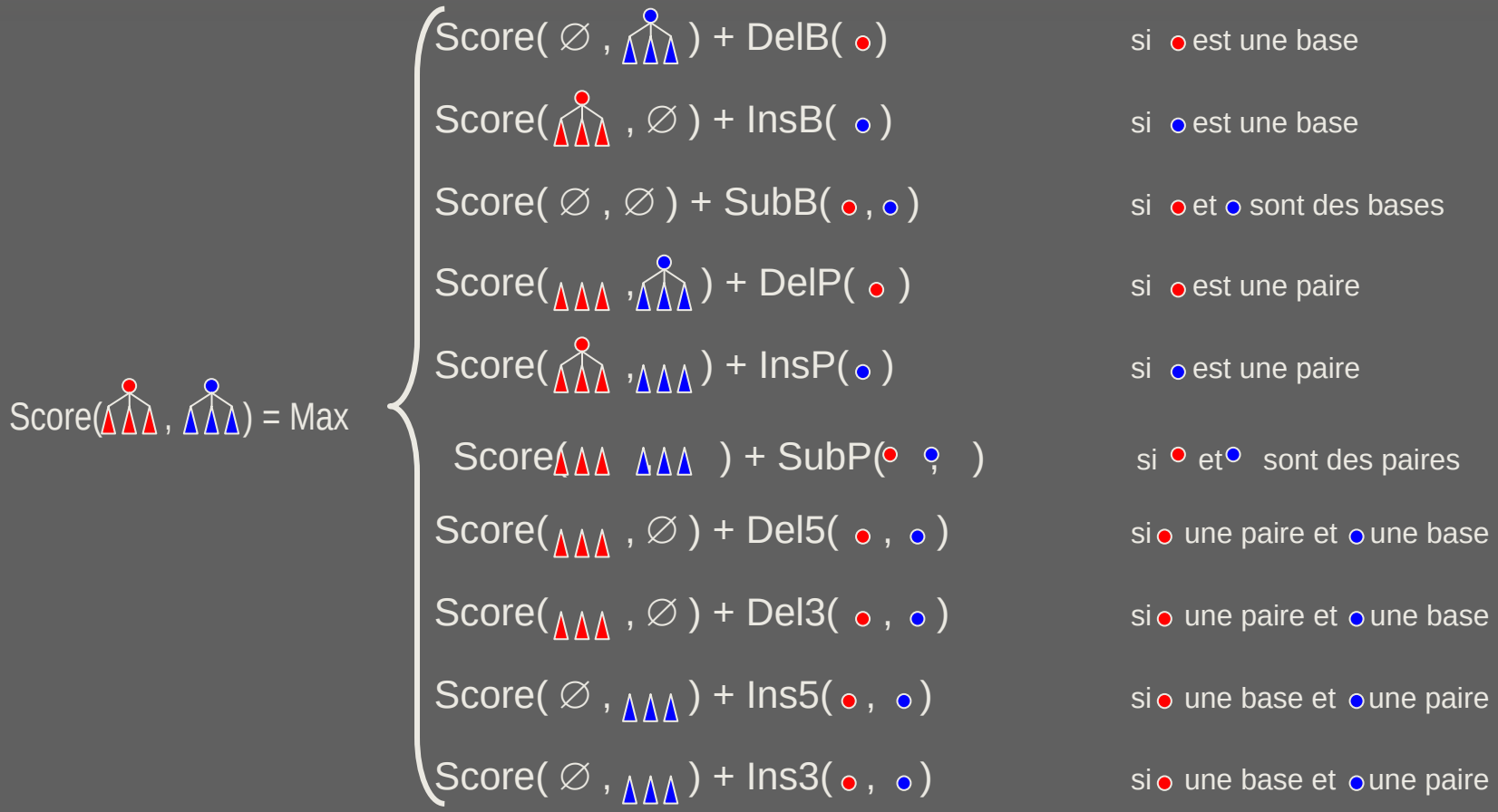
A-BCD-EFG
ABB-DF-FG

Alignement

AB---CDEFG
ABBDF---FG

Un algorithme d'alignement (1/4)

Herrbach, AD, Dulucq, Touzet 200



Un algorithme d'alignement (2/4)

Un algorithme d'alignement (3/4)

Un algorithme d'alignement (4/4)

Score (∅ , ∅) = 0

Score ( , ∅) = { Score (∅ , ∅) + DelB() si  est une base ; Score ( , ∅) + DelP() si  est une paire

Score ( , ∅) = Score ( , ∅) + Score ( , ∅)

Score (∅ , ) = { Score (∅ , ∅) + InsB() si  est une base ; Score (∅ , ) + InsP() si  est une paire

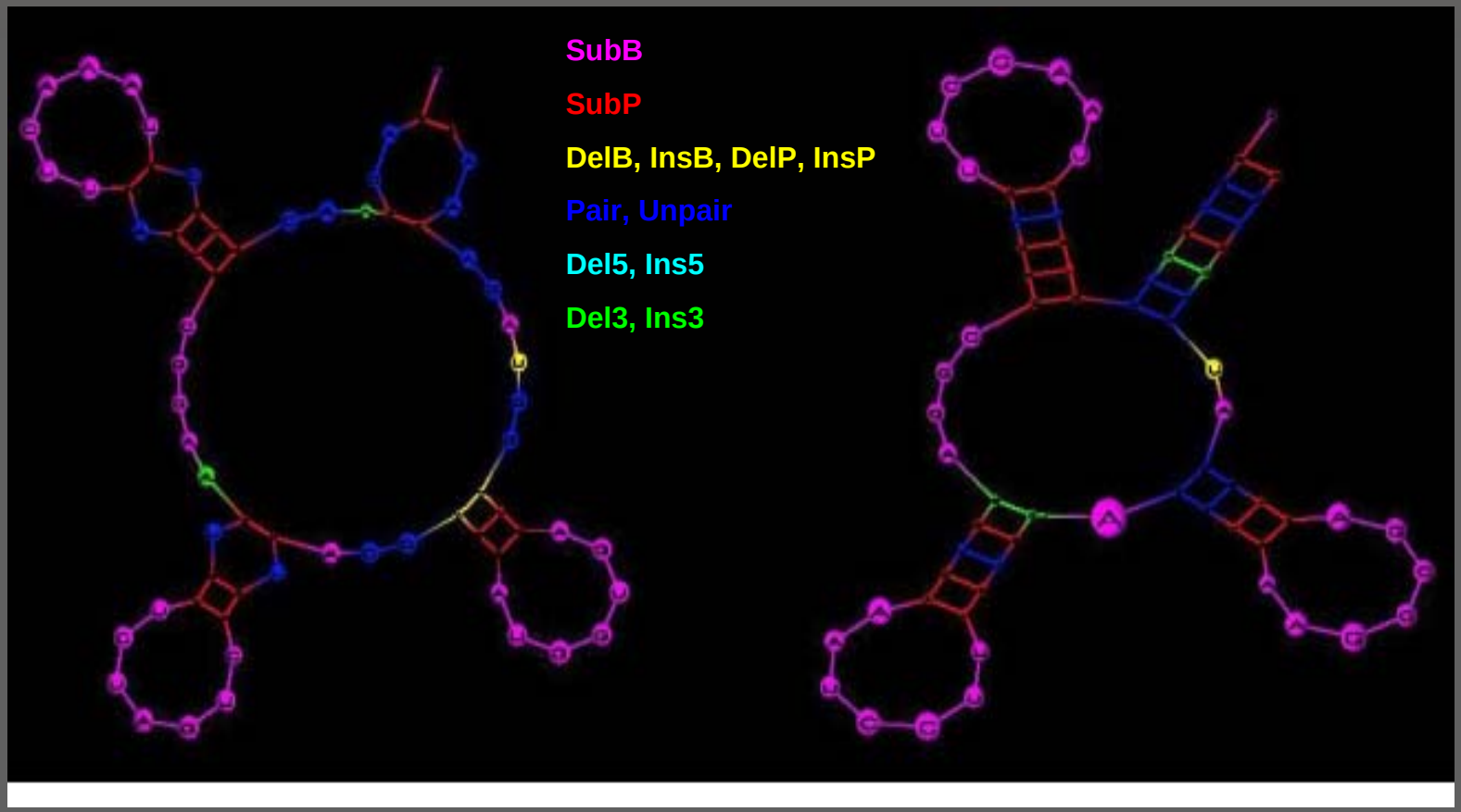
Score (∅ , ) = Score (∅ , ) + Score (∅ , )

Edition et alignement d'arbres

	Schéma « arbres »	Schéma « ARN »
Edition	$O(n^3 \log n)$ [Zhang-Shasha 1989, Klein 1998]	NP-complet [Blin, Fertin, Sinoquet, Rusu 2003]
Alignement	$O(n^4)$ [Jiang, Wang, Zhang 1995]	$O(n^4)$ [Herrbach, AD, Dulucq, Touzet 2005]

Exemple : deux ARNt

Image avec Tulip (David Auber, LaBR

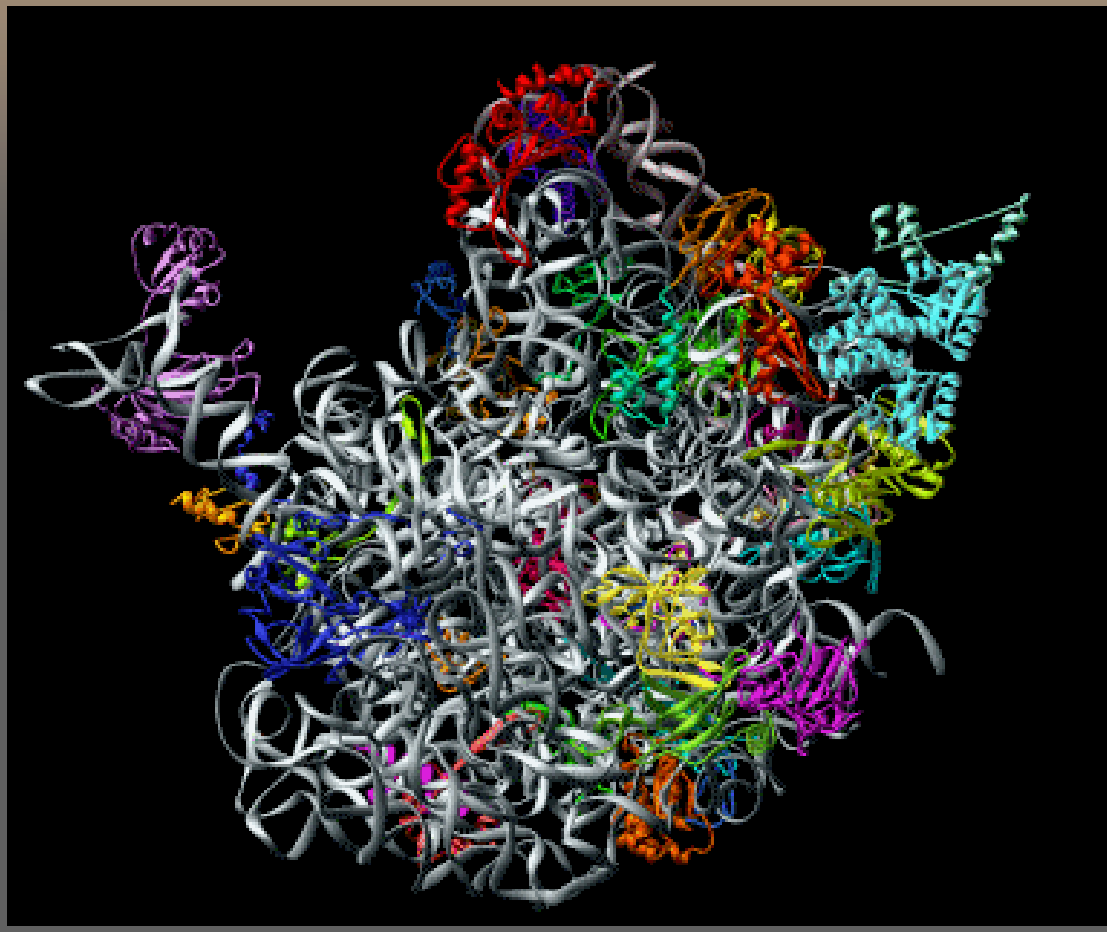


Homo sapiens

Bacillus subtilis

Crédits

- Serge Dulucq
- Claire Herrbach
- Yann Ponty
- Michel Termier
- Laurent Tichit
- Hélène Touzet
- Eric Westhof



ACI IMPBIO

NavGraphe
ACI MdD