

INFORMATIQUE: LA BOÎTE NOIRE DE LA GÉNOMIQUE

damien.eveillard@univ-nantes.fr



UNIVERSITÉ DE NANTES



PLAN DE COURS

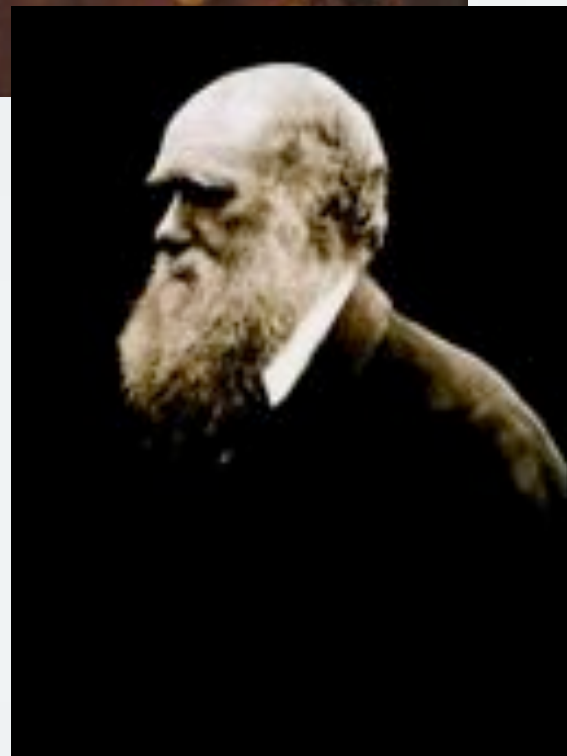
1.Motivations

2.Reconstruction de génome

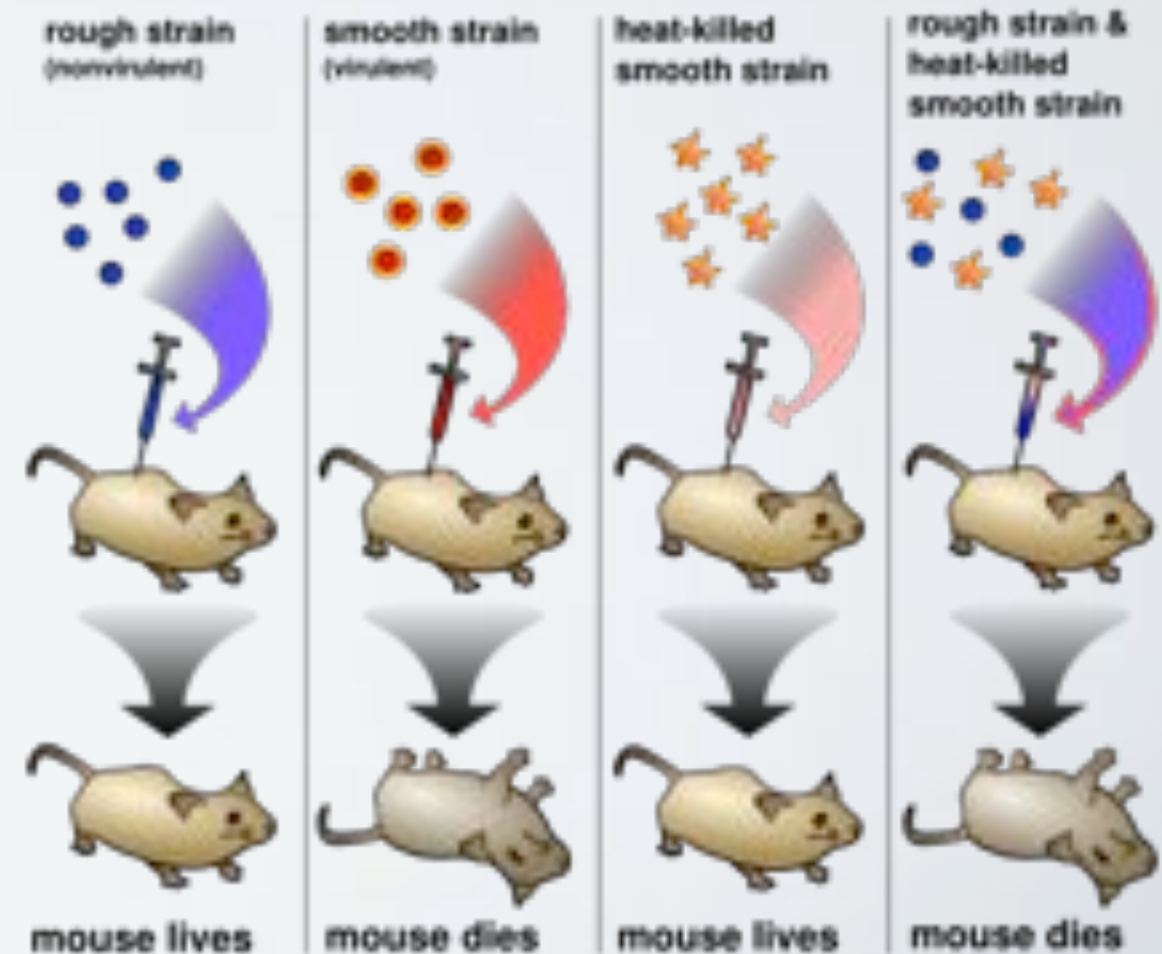
3.Annotation de génomes

4.Détermination de la fonction biologique

APPROCHE NATURALISTE



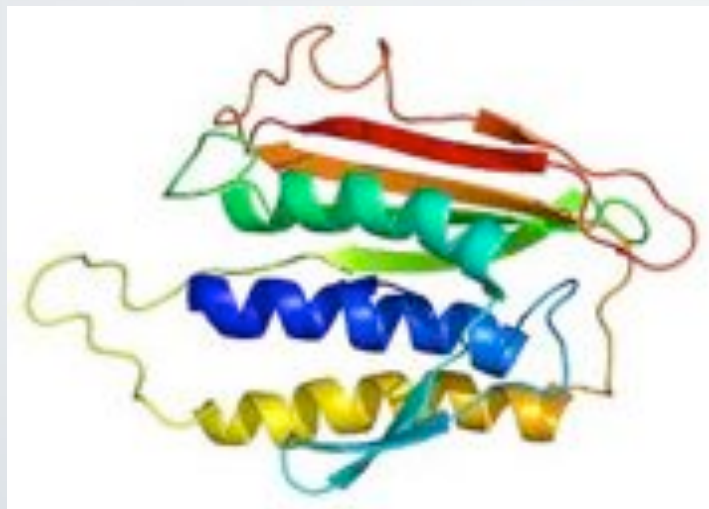
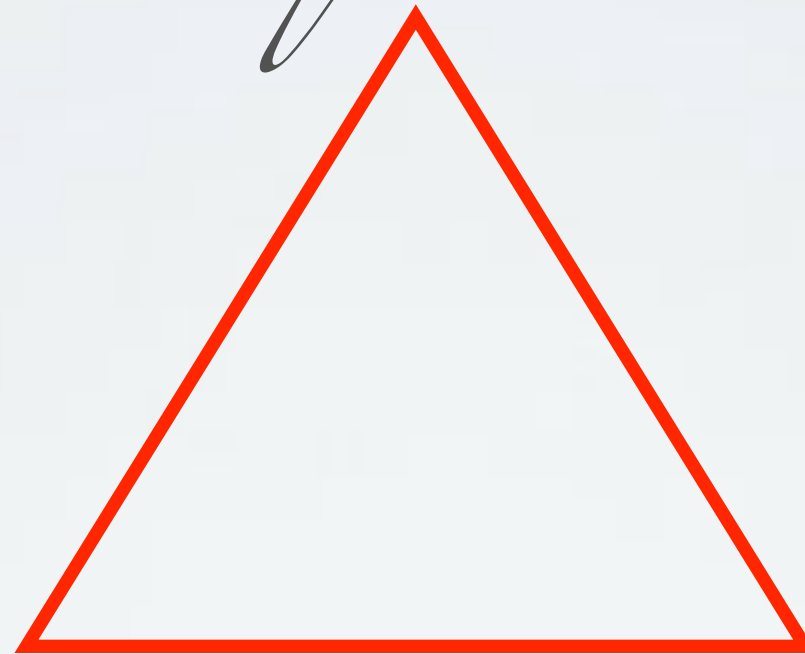
APPROCHE EXPÉRIMENTALISTE



BIOLOGIE MODERNE

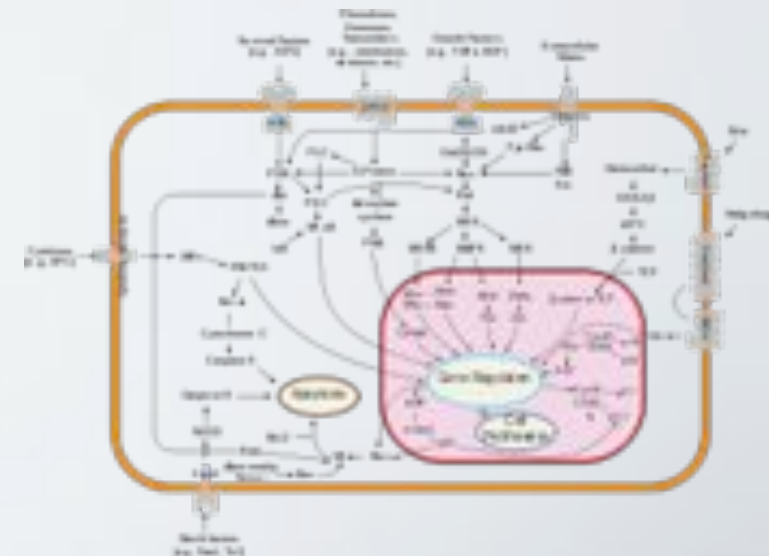
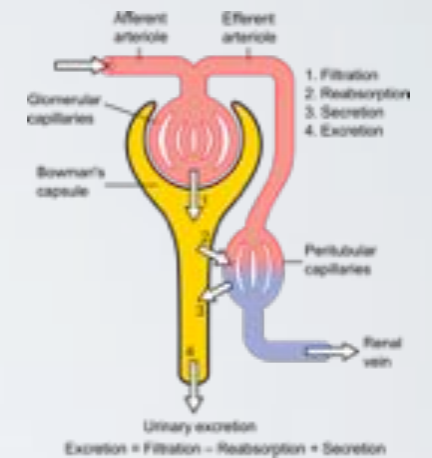


Séquence



Structure

Fonction



POURQUOI LA INFORMATIQUE ?

- Stocker les données **(base de données)**
- Ranger les données **(classification)**
- Reasonner sur les abstractions du vivant **(intelligence artificielle)**
- Prédire les comportements des systèmes naturels **(simulation)**

POURQUOI LA BOINFORMATIQUE ?

- Médecine : médecine personnalisée...
- Pharmaceutique
- Biotechnologies
- Ecologie
- Génétique

PROBLÈME DU SÉQUENÇAGE

comment l'Informatique a sauvé le
séquençage du génome humain



- Project
- Files
- Sub-Projects
- Sequences
- Annotations
- Alignments
- Gene and protein
- Map

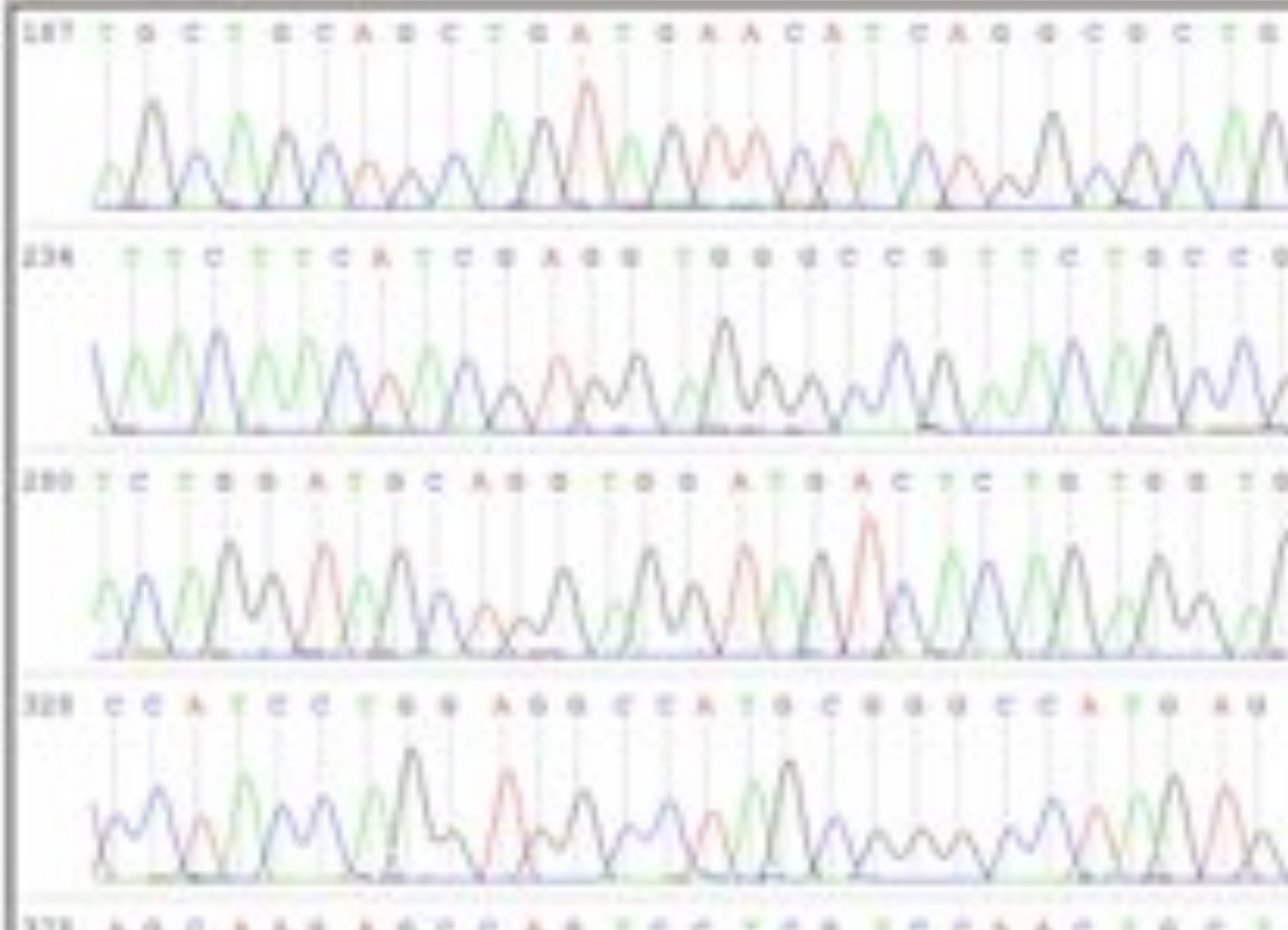
U.S. Govt. Lab

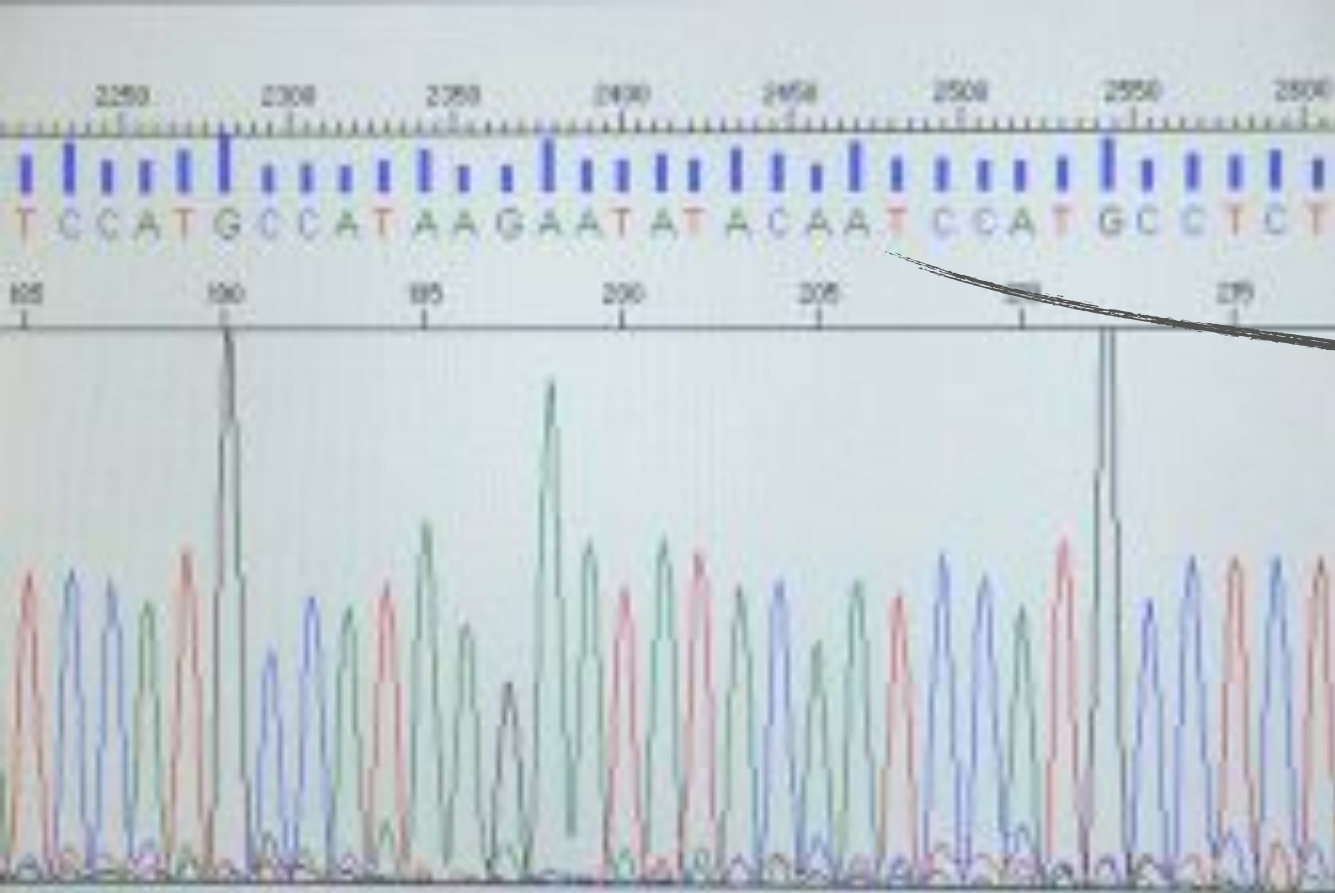
☐ Sequence ☐ Annotation

☐ Overlay ☐ Trace view

☐ Trace

Size: 1000 bp
 Sample: 1
 Lane: 1
 Date: 2000-01-11, 10:00:00
 Time: 2000-01-11, 10:00:00
 User:





G A A T T C A G T T A

- Technique **que** pour des petites séquences (4000 à 7000 pb)
- On «coupe» le génome en plein de petites séquences

AGCATGCTGCAGTCATGCT

AGCATG

TAGGCTA

CTGCAGTCATGCTTAGGCTA

AGCATGCTGCAGTCATGCT

AGCATG

CTGCAGTCATGCT**TAGGCTA**

OK

AGCATGCTGCAGTCATGCT

~~AGCATG~~

NON

~~CTGCAGTCATGCTTAGGCTA~~

OK

AGCATGCTGCAGTCATGCT

CTGCAGTCATGCTTAGGCTA

OK

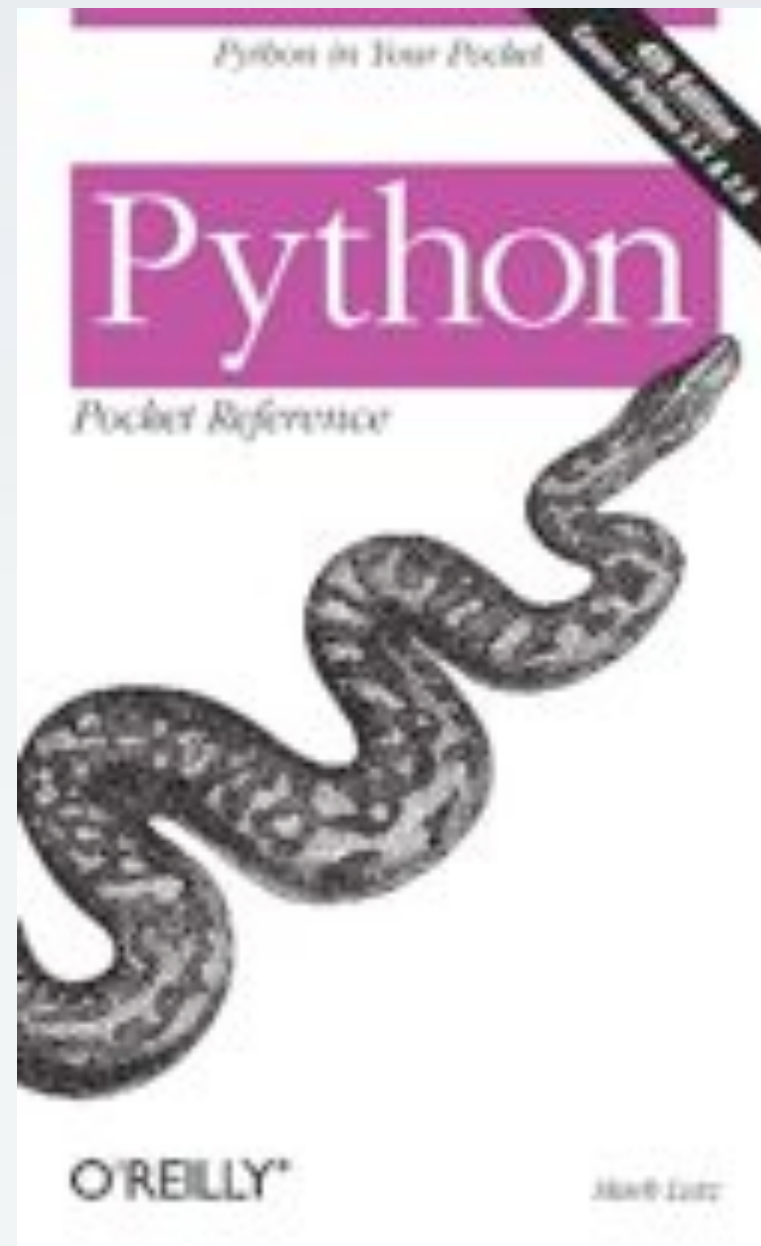
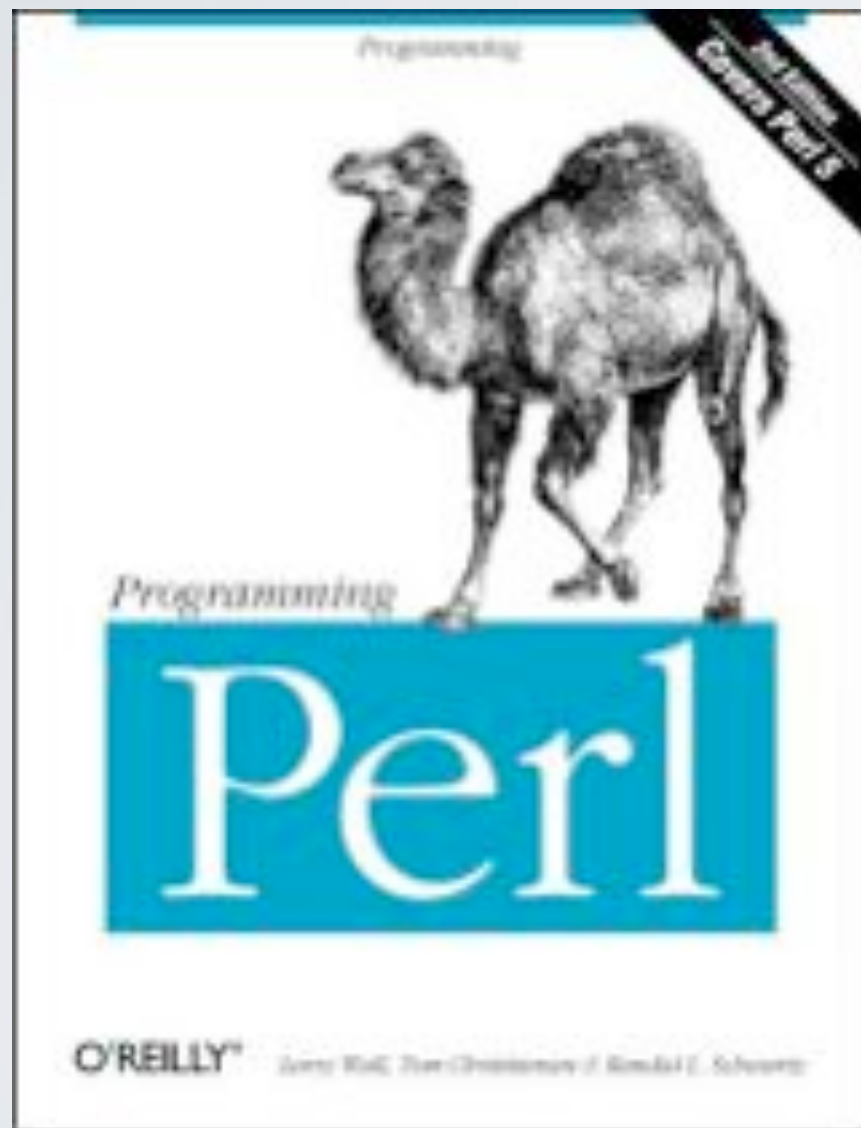
AGCATG**CTGCAGTCATGCT**TAGGCTA

g n me originel

PROBLÈME

- Pour le génome humain : $3 \cdot 10^9$ caractères (pb)
- 3 gigabytes de stockage
- + annotation : 1 à 10 terabytes d'information
- trop de séquences : besoin de stocker et automatiser

DE LA BIOLOGIE VERS L'INFORMATIQUE ?



ALIGNEMENT DE SÉQUENCES

Programmation dynamique

ALGORITHME : NEEDLEMAN/ WUNSCH

	séquence	taille
séquence A	G A A T T C A G T T A	11 (m)
séquence B	G G A T C G A	7 (n)

UNE VERSION SIMPLIFIÉE (PEU D'IMPORTANCE À LA SIMILARITÉ)

On considère un système de notation simplifié

- $S(A_i, B_j) = 1$ si les bases (en séquence A position i et séquence B position j) sont les mêmes
- $S(A_i, B_j) = 0$ si les bases (en séquence A position i et séquence B position j) sont différentes
- $w = 0$ si une base manque sur la séquence A ou la séquence B

ALGORITHME EN 3 ÉTAPES

a.Initialisation

b.Remplissage de la matrice (scoring)

c.Retour de trace (alignement)

A) INITIALISATION

- On crée une matrice de taille $(m+1, n+1)$

		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0											
G	0											
A	0											
T	0											
C	0											
G	0											
A	0											

B) REMPLISSAGE DE LA MATRICE

- On part en haut à gauche et on recherche le score maximum à chaque position de la matrice : $M_{i,j}$
- Pour trouver $M_{i,j}$ on doit connaître les valeurs à gauche, au dessus et à la diagonale de (i,j) : $M_{i-1,j}$ $M_{i,j-1}$ $M_{i-1,j-1}$

```
Mi,j = MAXIMUM[  
    Mi-1,j-1 + S(Ai,Bj) #identité ou pas sur la  
    diagonale  
    Mi,j-1 + w, #délétion séquence A  
    Mi-1,j + w #délétion séquence B]
```

B) REMPLISSAGE DE LA MATRICE

$$M_{i,j} = \text{MAXIMUM} [$$

$\mathbf{M}_{i-1, j-1} + \mathbf{S}_{i,j}$, #identité ou pas sur la diagonale

$M_{i,j-1} + w,$ #délétion séquence 1

$M_{i-1,j} + w$ #déletion séquence 2]

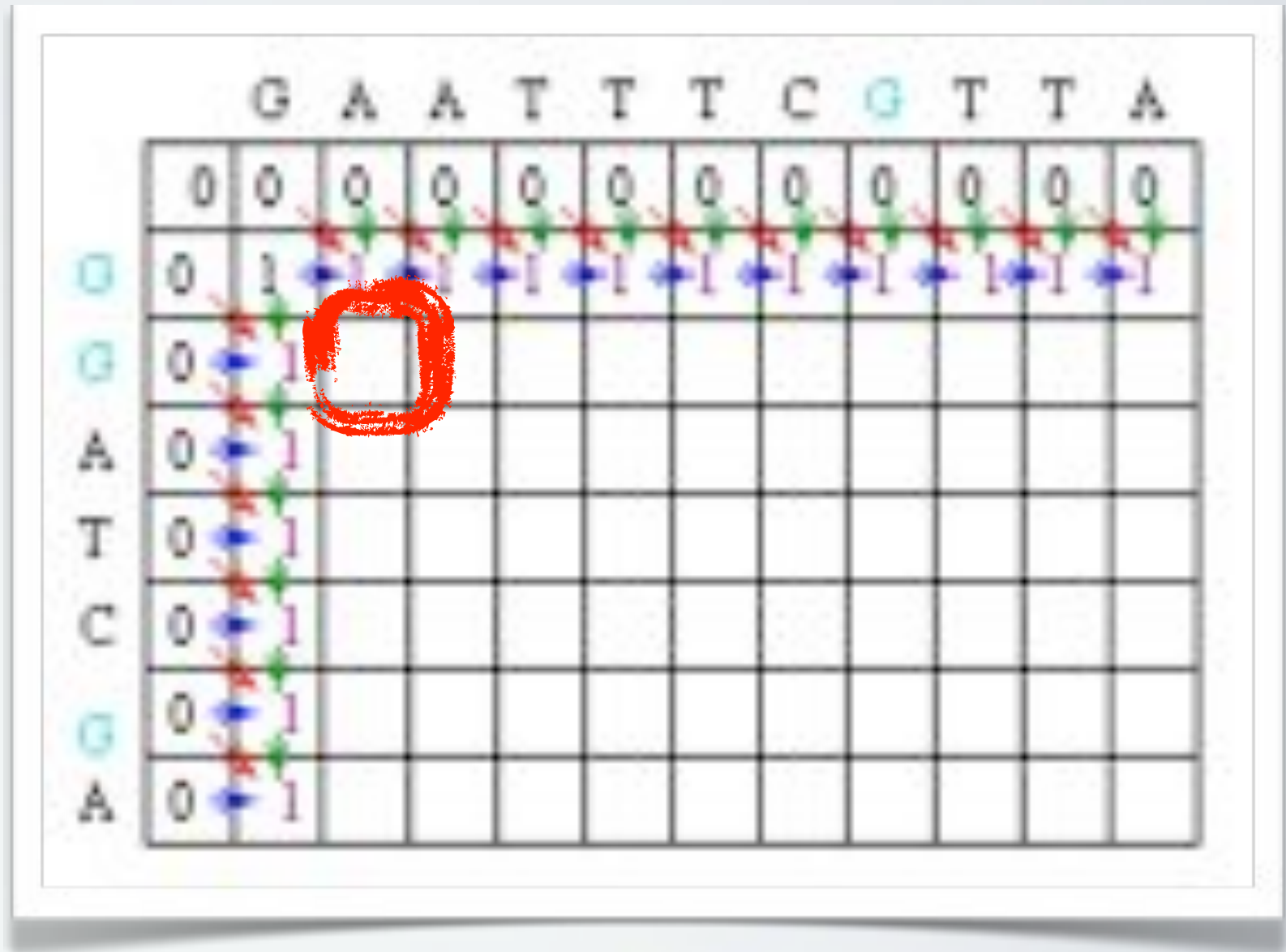
$M_{1,1} = \text{MAX}[M_{0,0} + 1, M_1 = \text{MAX}[1, 0, 0] = 1.$

$$\begin{aligned} M_{1,1} &= \text{MAX}[M_{0,0} + 1, M_{1,0} + 0, M_{0,1} + 0] \\ &= \text{MAX} [1, 0, 0] = 1. \end{aligned}$$

B) REMPLISSAGE DE LA MATRICE

La valeur de la déléation
est 0 :

- On peut remplir la colonne l avec des l
- On peut remplir la ligne l avec des l



B) REMPLISSAGE DE LA MATRICE

		G	A	A	T	T	C	A	G	T	T	A
		0	0	0	0	0	0	0	0	0	0	0
G		0	1	1	1	1	1	1	1	1	1	1
G		0	1	1	1	1	1	1	2	2	2	2
A		0	1	2	2	2	2	2	2	2	2	3
T		0	1	2	2	3	3	3	3	3	3	3
C		0	1	2	2	3	3	3	4	4	4	4
G		0	1	2	2	3	3	3	4	4	5	5
A		0	1	2	3	3	3	3	4	5	5	6

C) RETOUR DE TRACE

- Une fois la matrice remplie, on obtient le meilleur score d'alignement (6)
- On veut **visualiser** le meilleur alignement
- On commence en bas à droite de la matrice : $M(m,n)$

C) RETOUR DE TRACE

- On recherche le prédécesseur

		G	A	A	T	T	C	A	G	T	T	A
		0	0	0	0	0	0	0	0	0	0	0
G		0	1	1	1	1	1	1	1	1	1	1
G		0	1	1	1	1	1	1	2	2	2	2
A		0	1	1	2	2	2	2	2	2	2	3
T		0	1	2	2	3	3	3	3	3	3	3
C		0	1	2	2	3	3	4	4	4	4	4
G		0	1	2	2	3	3	4	4	5	5	5
A		0	1	2	3	3	3	4	5	5	5	6

C) RETOUR DE TRACE

- On recherche le prédécesseur

T A
|
A
_

		G	A	A	T	T	C	A	G	T	T	A
		0	0	0	0	0	0	0	0			
G		0	1	1	1	1	1	1	1			
G		0	1	1	1	1	1	1	2			
A		0	1	1	2	2	2	2	2			
T		0	1	2	2	3	3	3	3			
C		0	1	2	2	3	3	4	4			
G		0	1	2	2	3	3	4	4	5	5	5
A												6

C) RETOUR DE TRACE

- On recherche le prédécesseur

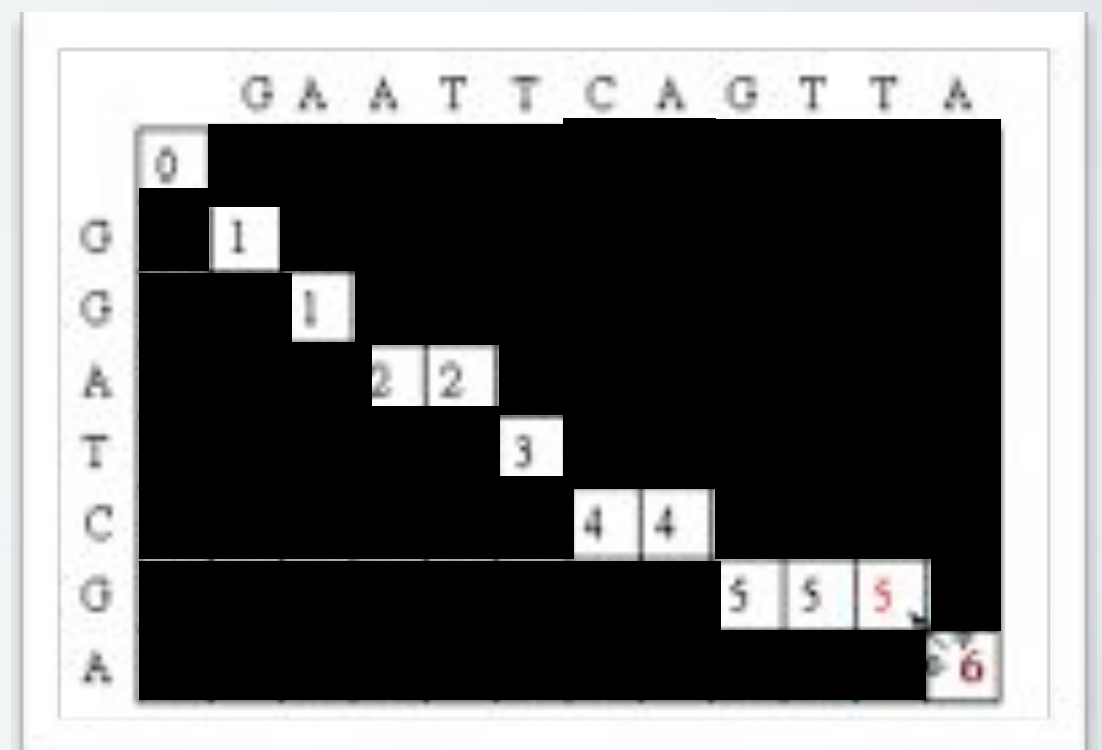
T T A
|
_ _ A

		G	A	A	T	T	C	A	G	T	T	A
		0	0	0	0	0	0	0				
G		0	1	1	1	1	1	1				
G		0	1	1	1	1	1	1				
A		0	1	1	2	2	2	2				
T		0	1	2	2	3	3	3				
C		0	1	2	2	3	3	4				
G		0	1	2	2	3	3	4	4	5	5	5
A												6

C) RETOUR DE TRACE

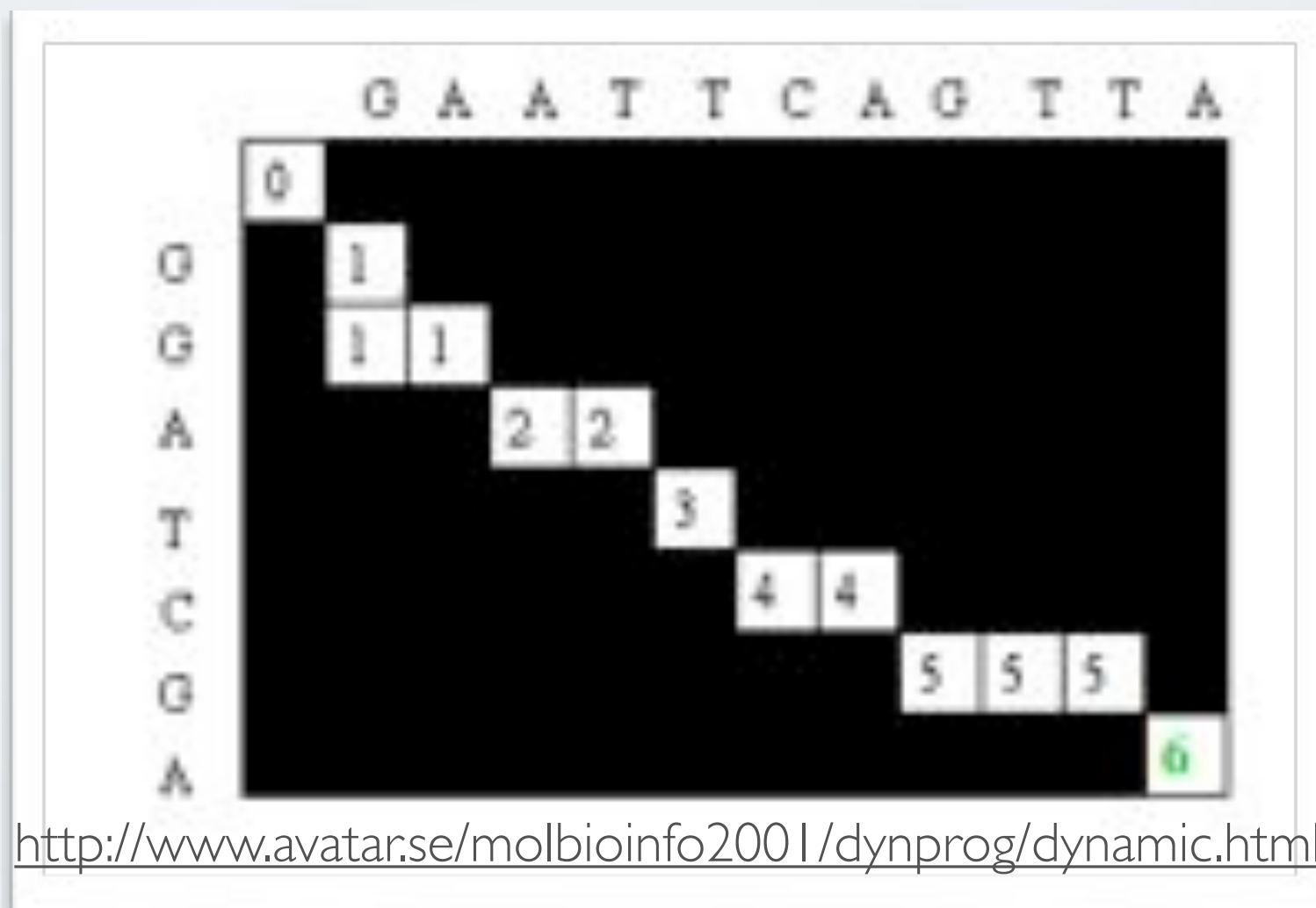
- On recherche le prédécesseur

G	A	A	T	T	C	A	G	T	T	A
G	G	A	_	T	C	_	G	_	_	A



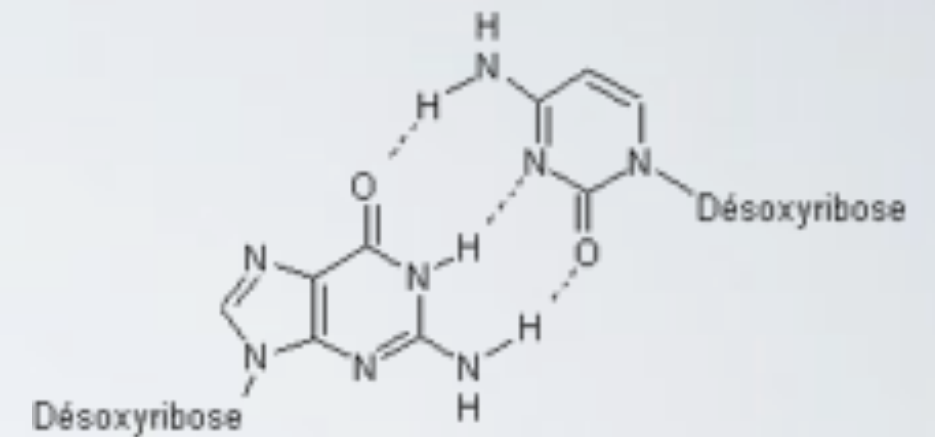
C) RETOUR DE TRACE

G	_	A	A	T	T	C	A	G	T	T	A
G	G	_	A	_	T	C	_	G	_	_	A

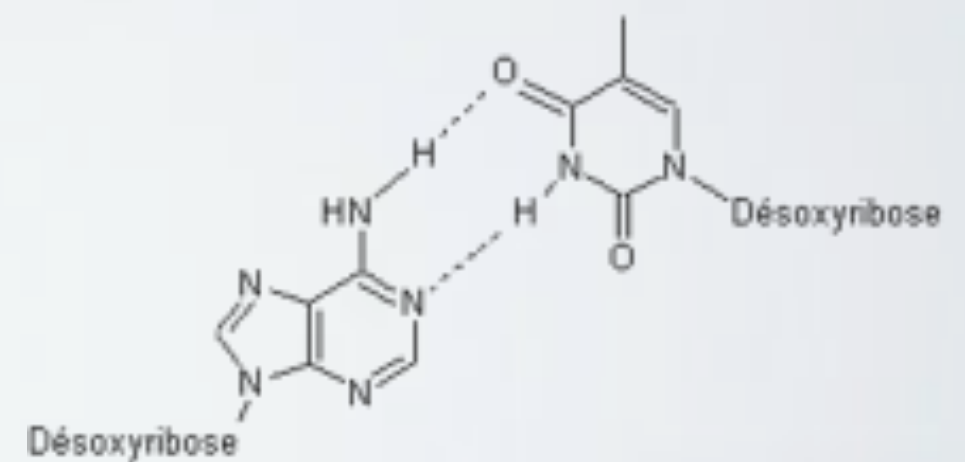


EN PRATIQUE

	A	G	C	T
A	10	-1	-3	-4
G	-1	7	-5	-3
C	-3	-5	9	0
T	-4	-3	0	8



Liaisons hydrogène guanine-cytosine



Liaisons hydrogène adénine-thymine