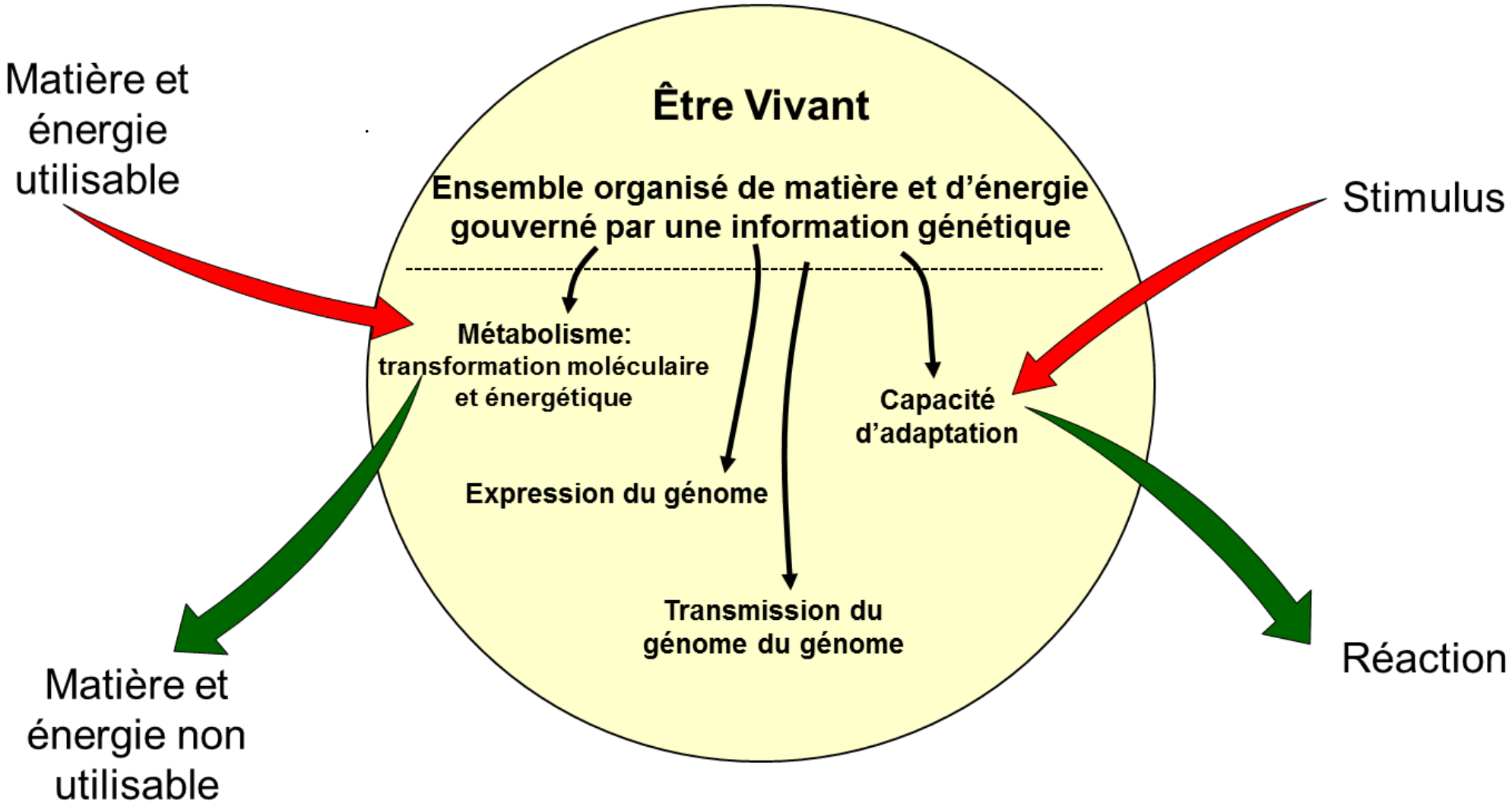
LU3IN013:
Initiation à la recherche

Thématique Bioinformatique:

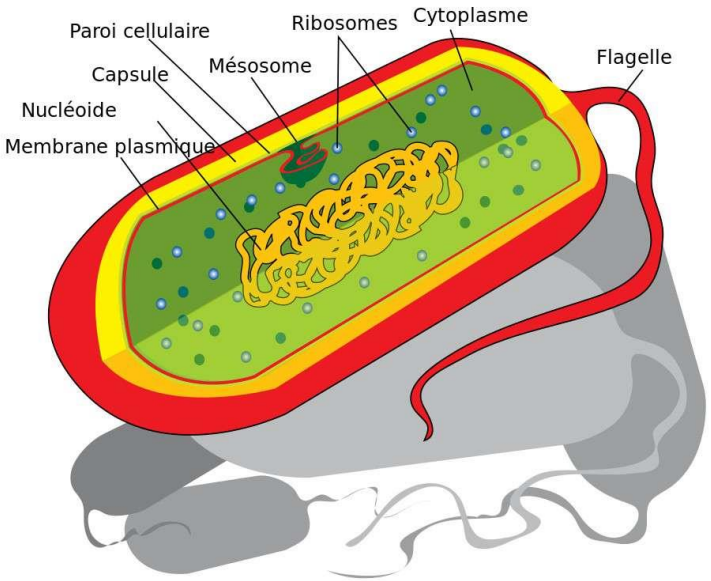
Recherche d'information dans les génomes des organismes vivants

Le génome, une information à la base du fonctionnement des êtres vivants

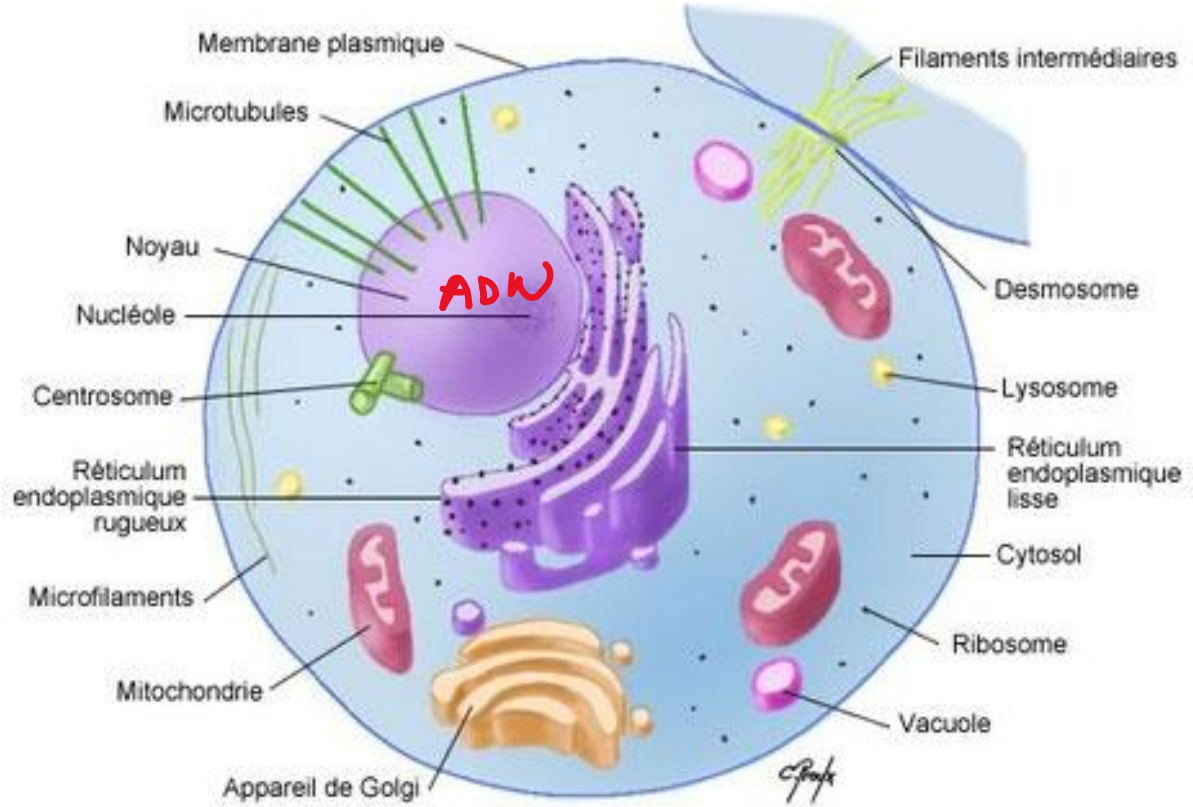


Les cellules, unite de bases des êtres vivants

Cellule procaryote



Cellule eucaryote



Cours 1

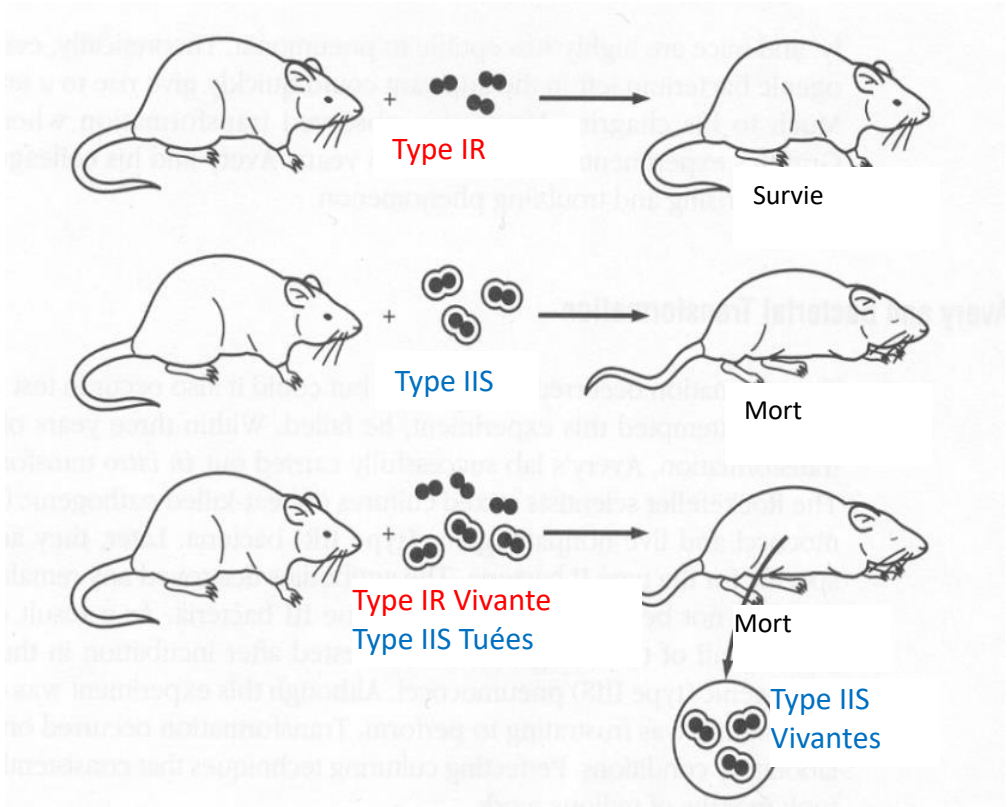
Qu’est-ce qu’un génome?

Support de l’information génétique?

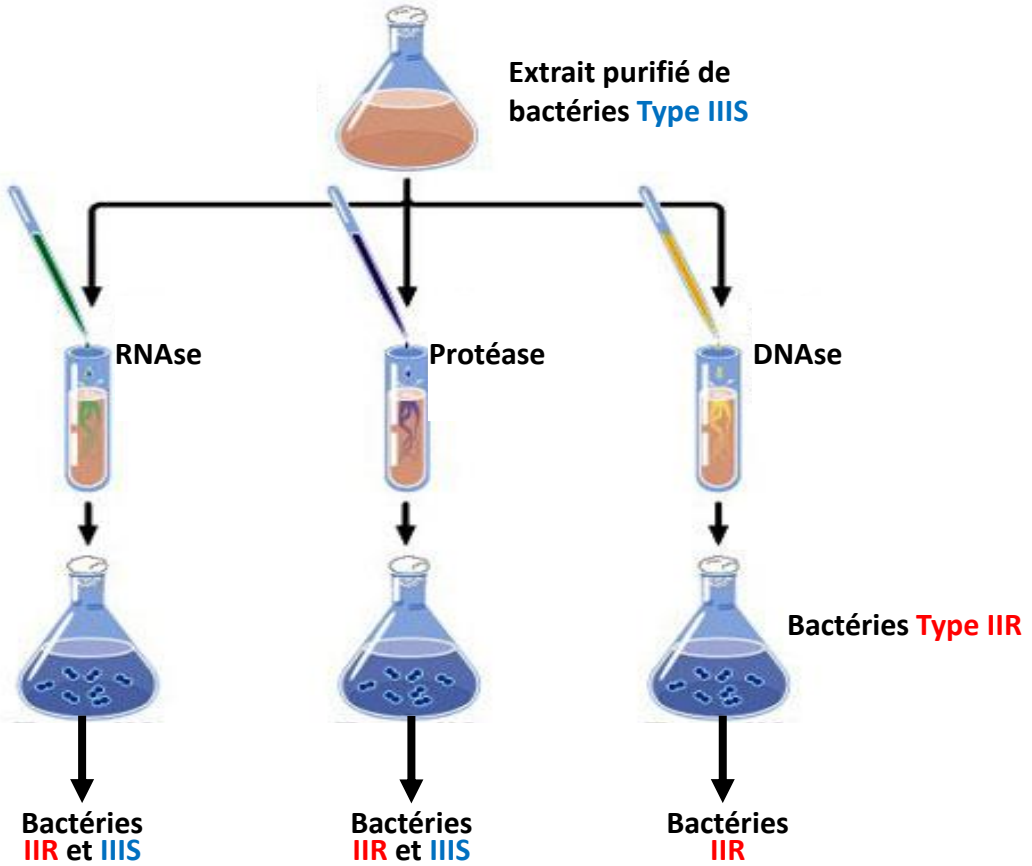
Structure de l’information contenu dans les génome?

La découverte de l'ADN comme support de l'information génétique

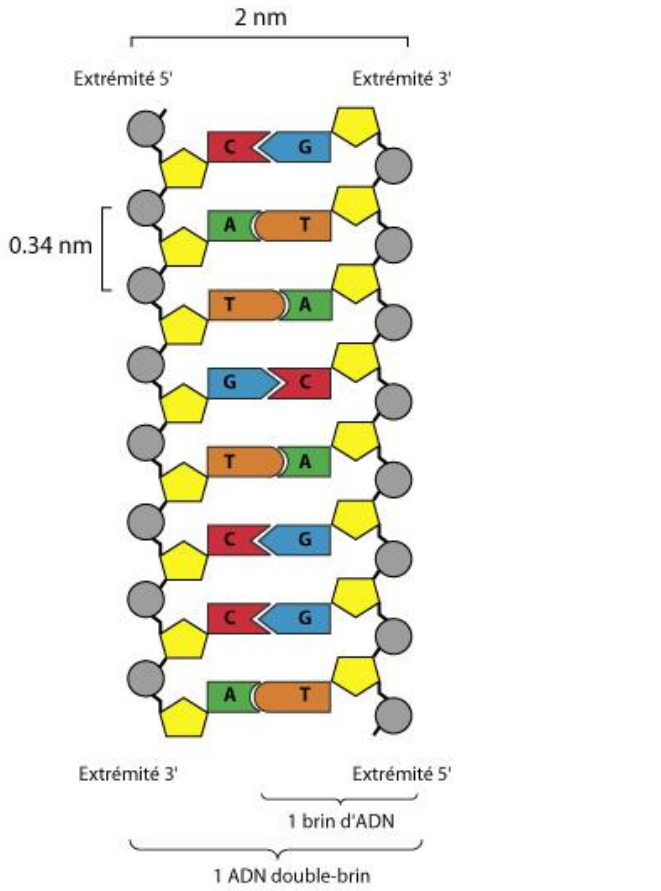
Expérience de transformation de Griffith (1928)

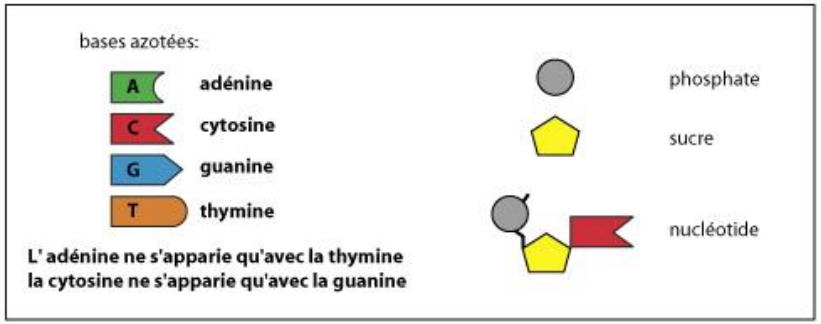


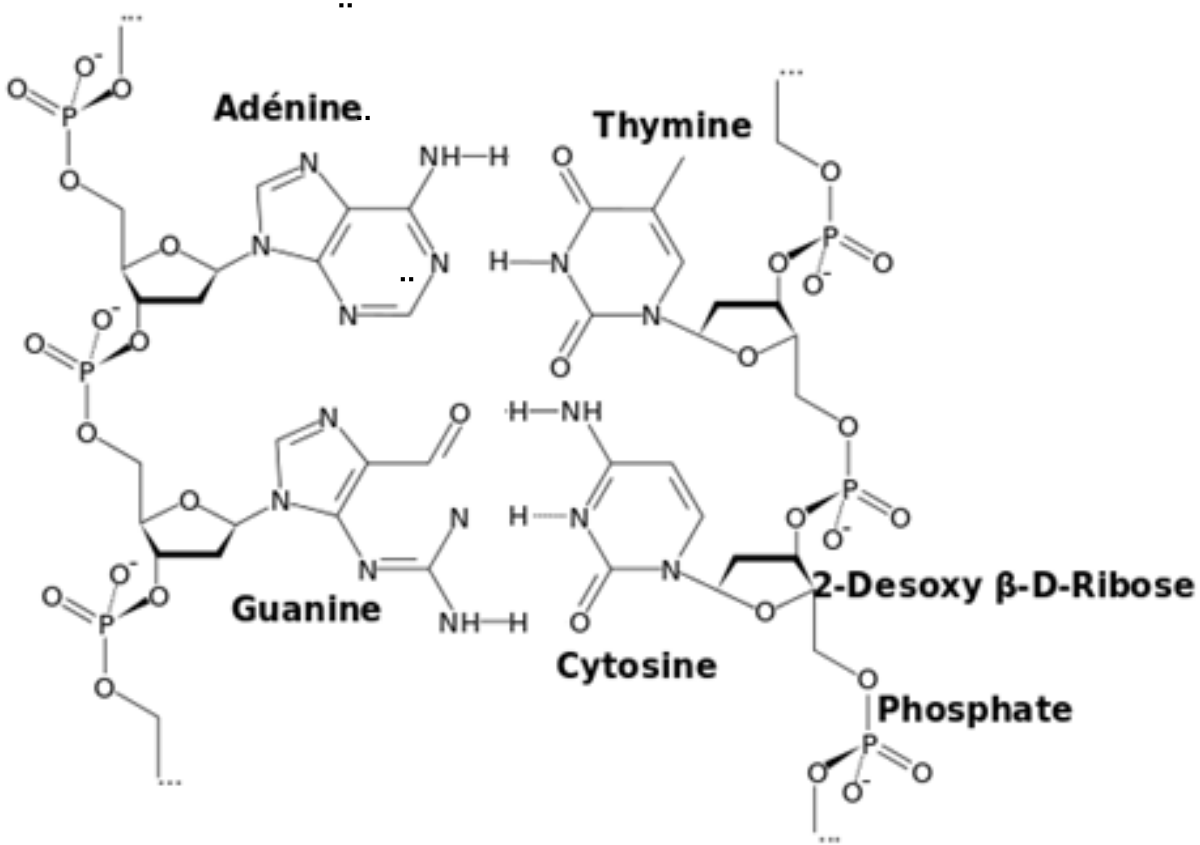
Purification du facteur transformant par Avery (1943)



La molécule d'ADN (structure élucidée en 1953 par Watson et Crick)

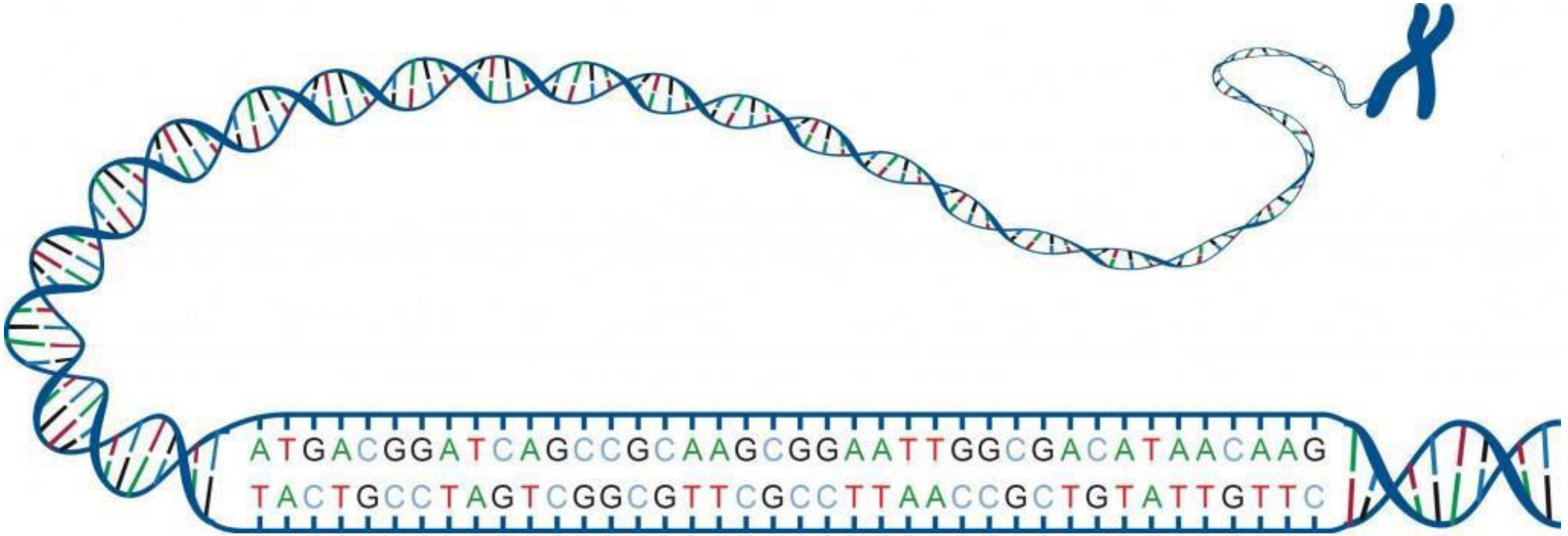




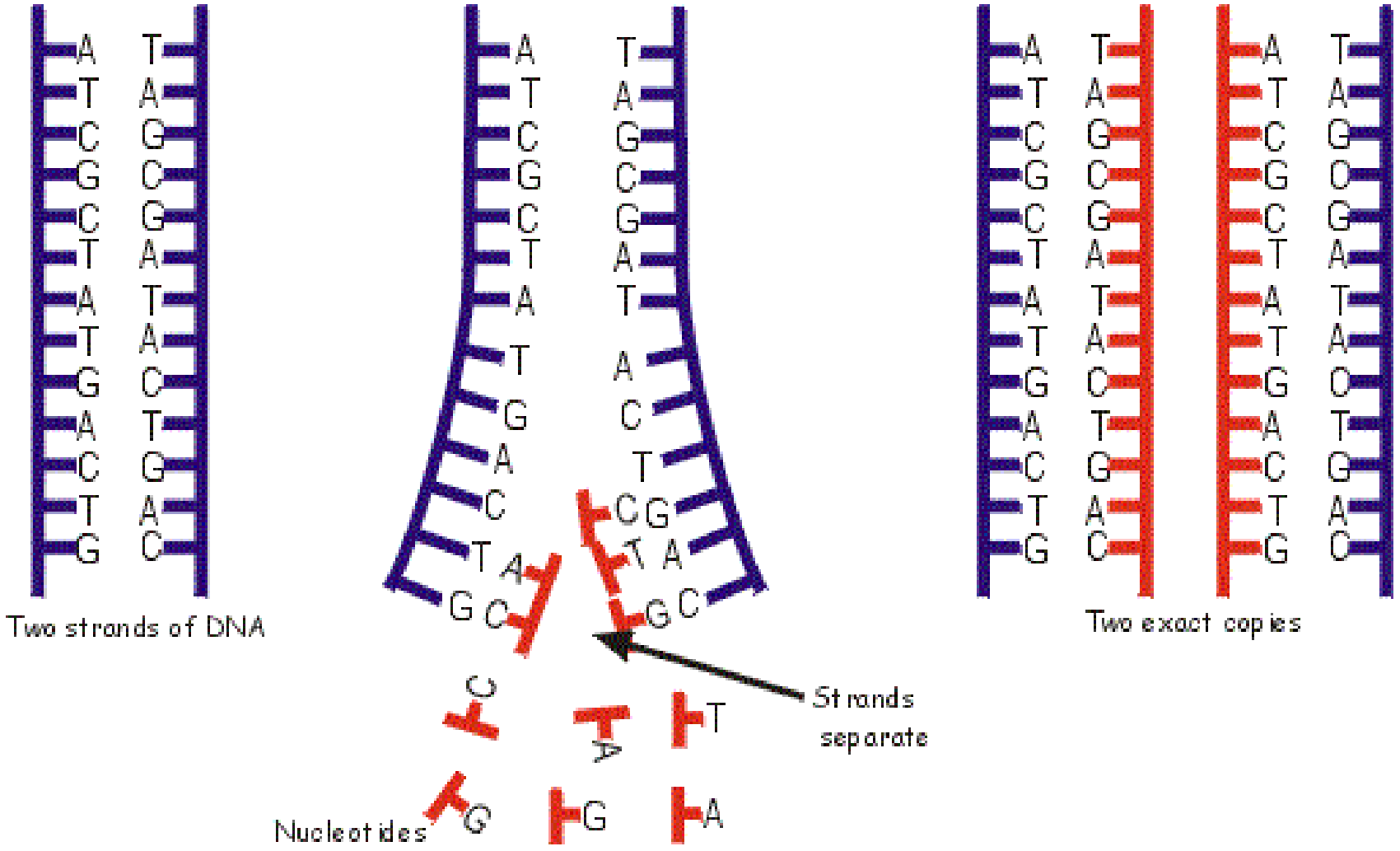


Support et Organisation de l'information des génomes

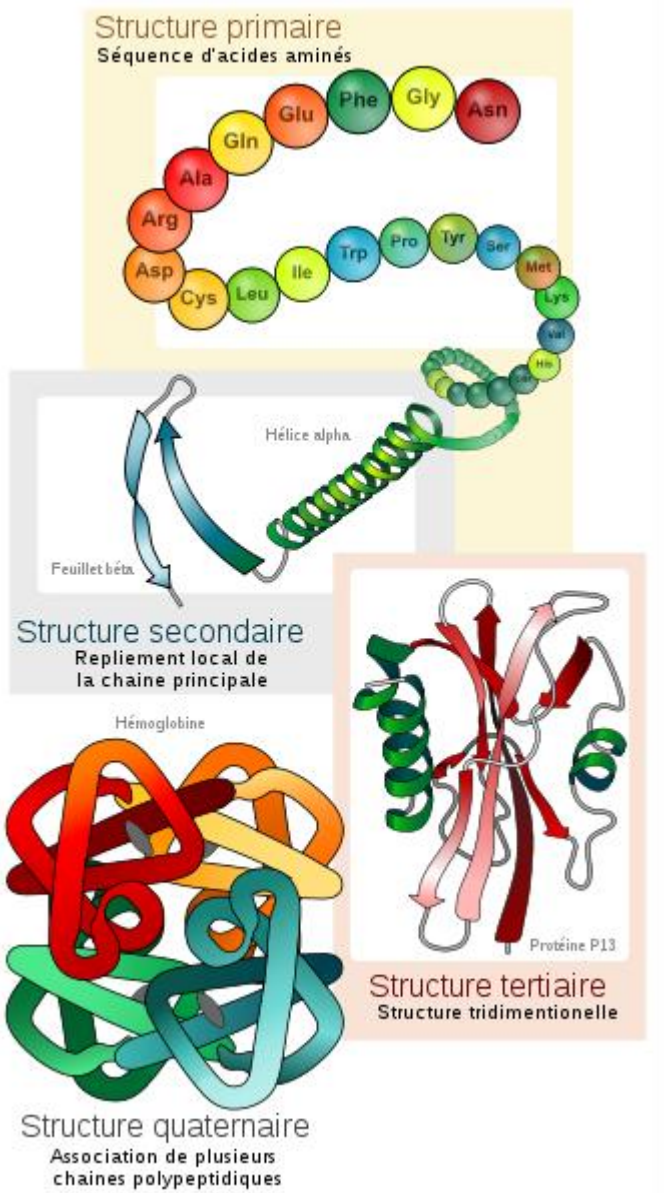
1 chromosome= 1 molécule d'ADN = 2 brins d'ADN avec des séquences complémentaires

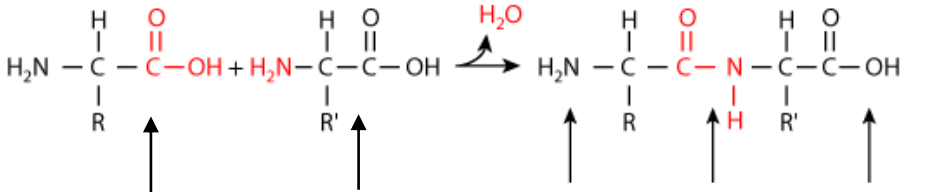


La réplication semi-conservative de la molécule d'ADN



Les protéines

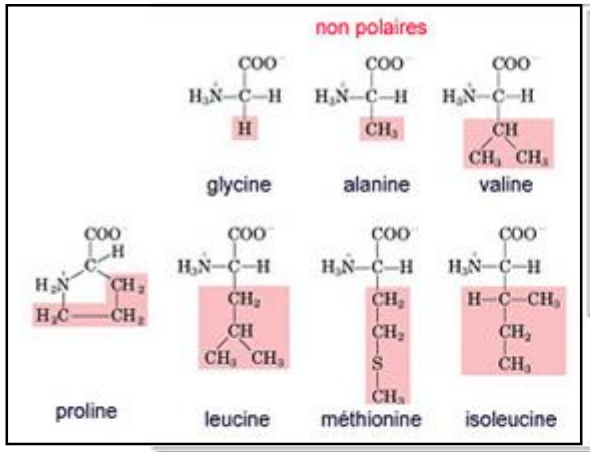


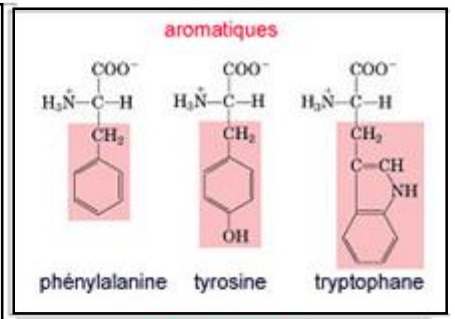


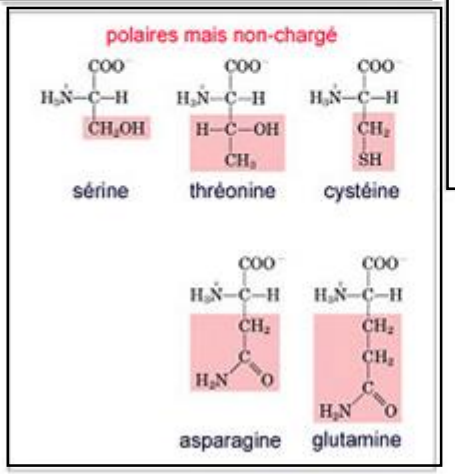
Les 20 acides aminés

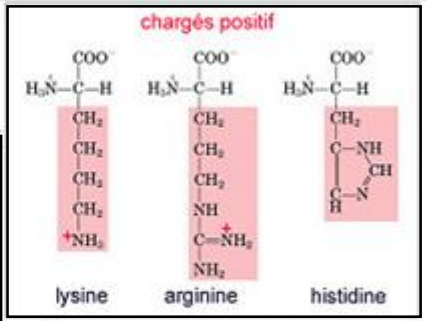
Acide glutamique	Glu	E
Acide aspartique	Asp	D
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Cystéine	Cys	C
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I

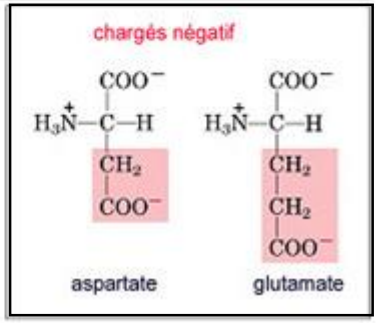
Leucine	Leu	L
Lysine	Lys	K
Méthionine	Met	M
Phénylalanine	Phe	F
Proline	Pro	P
Sérine	Ser	S
Thréonine	Thr	T
Tryptophane	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V



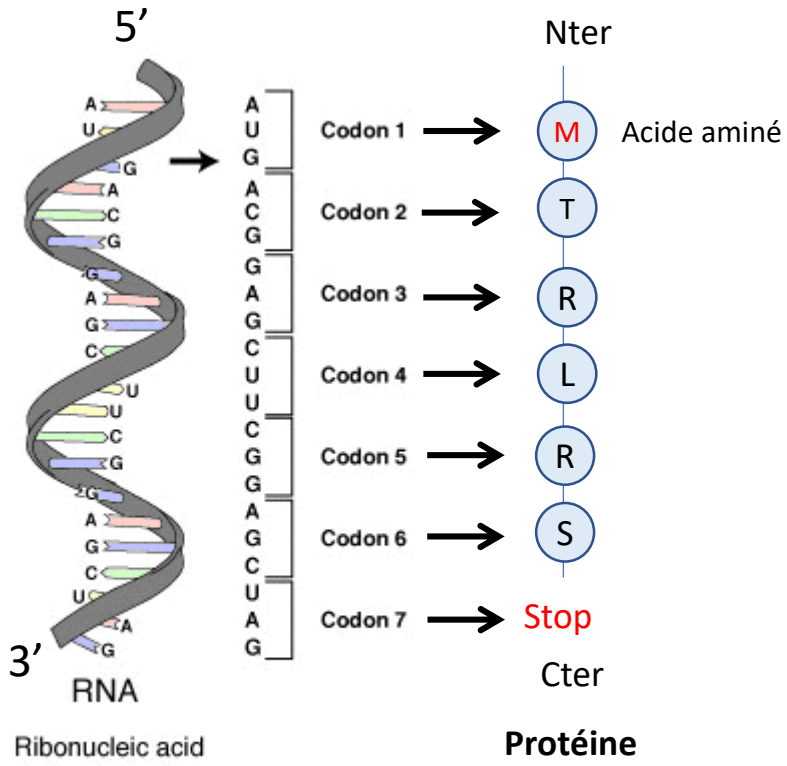






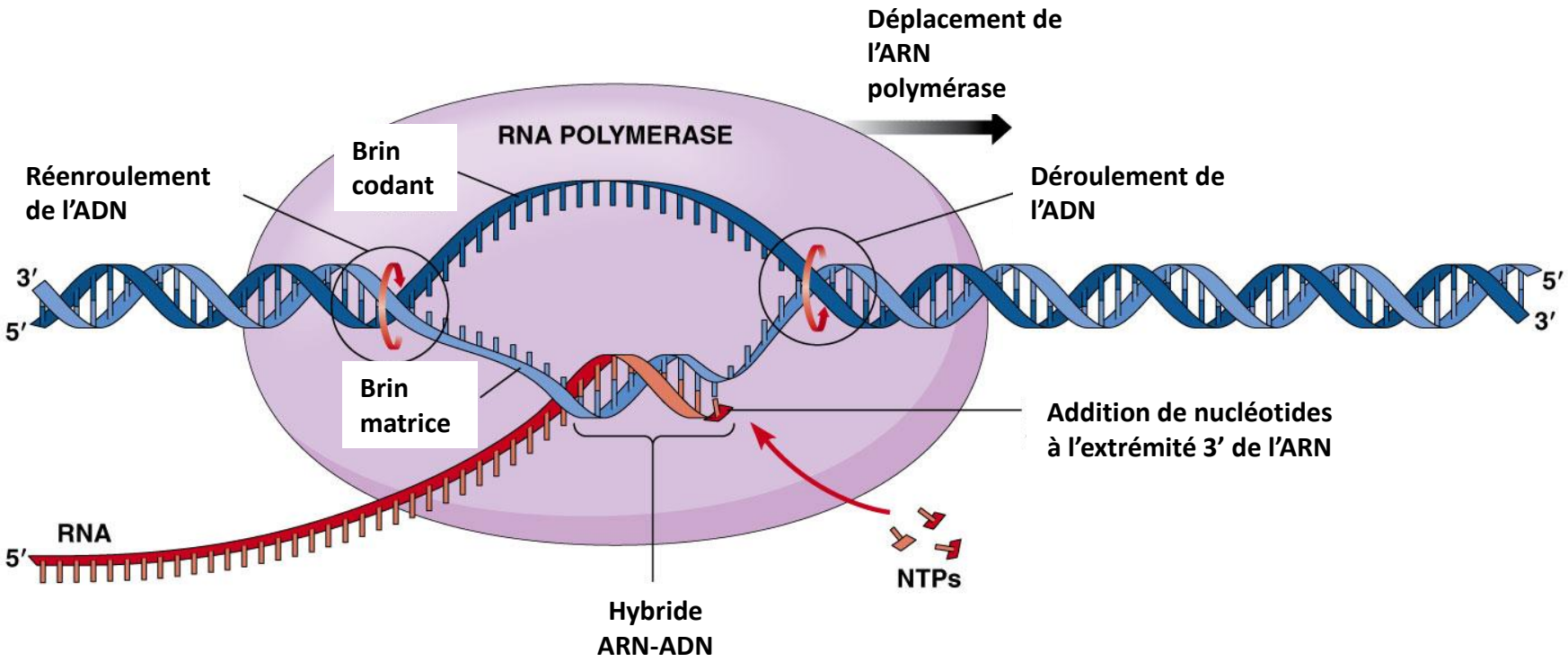


Le code génétique



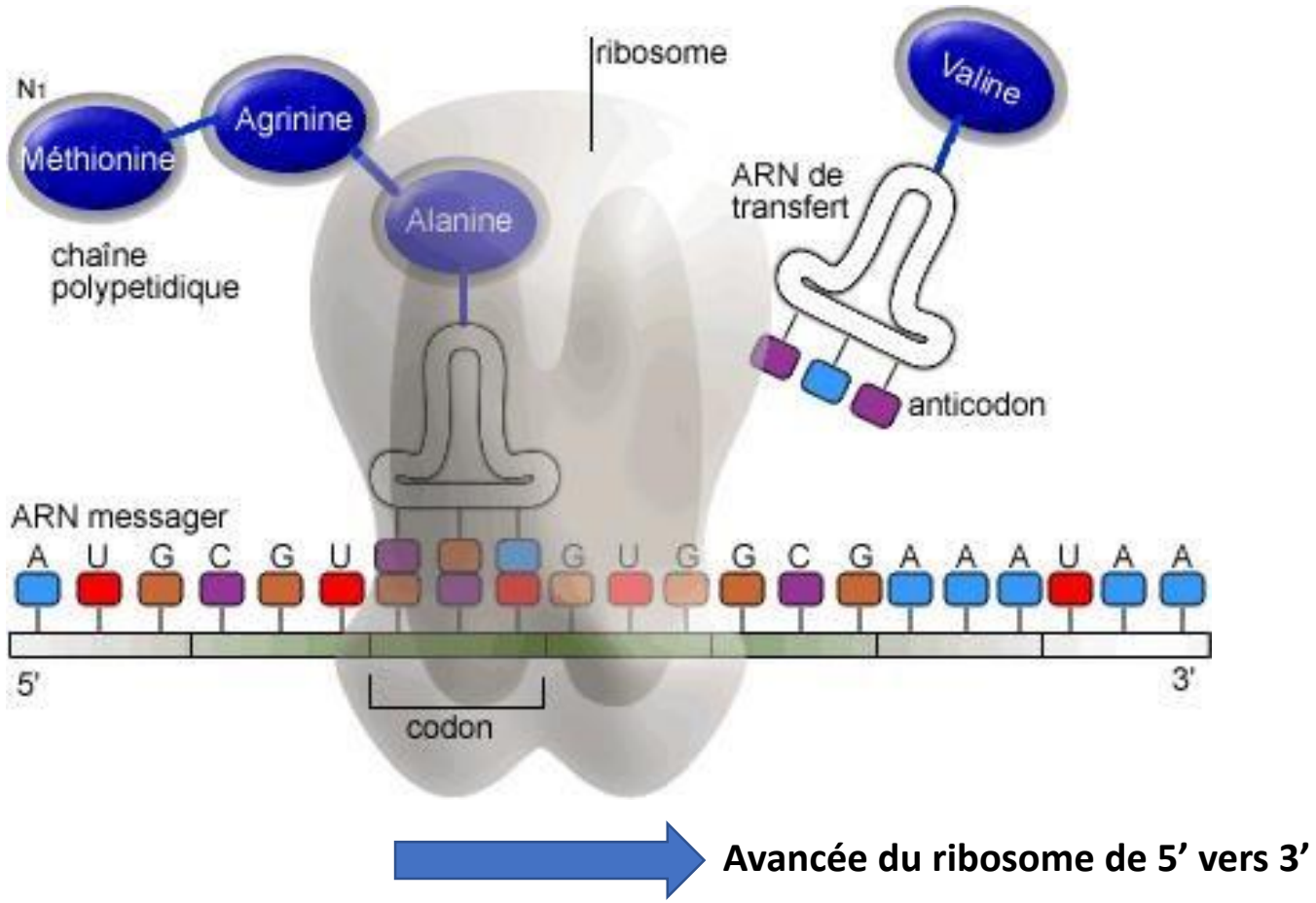
1st letter	Second Letter U		Second Letter C		Second Letter A		Second Letter G		3rd letter
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

La transcription



La transcription en vidéo

La traduction



Avancée du ribosome de 5’ vers 3’

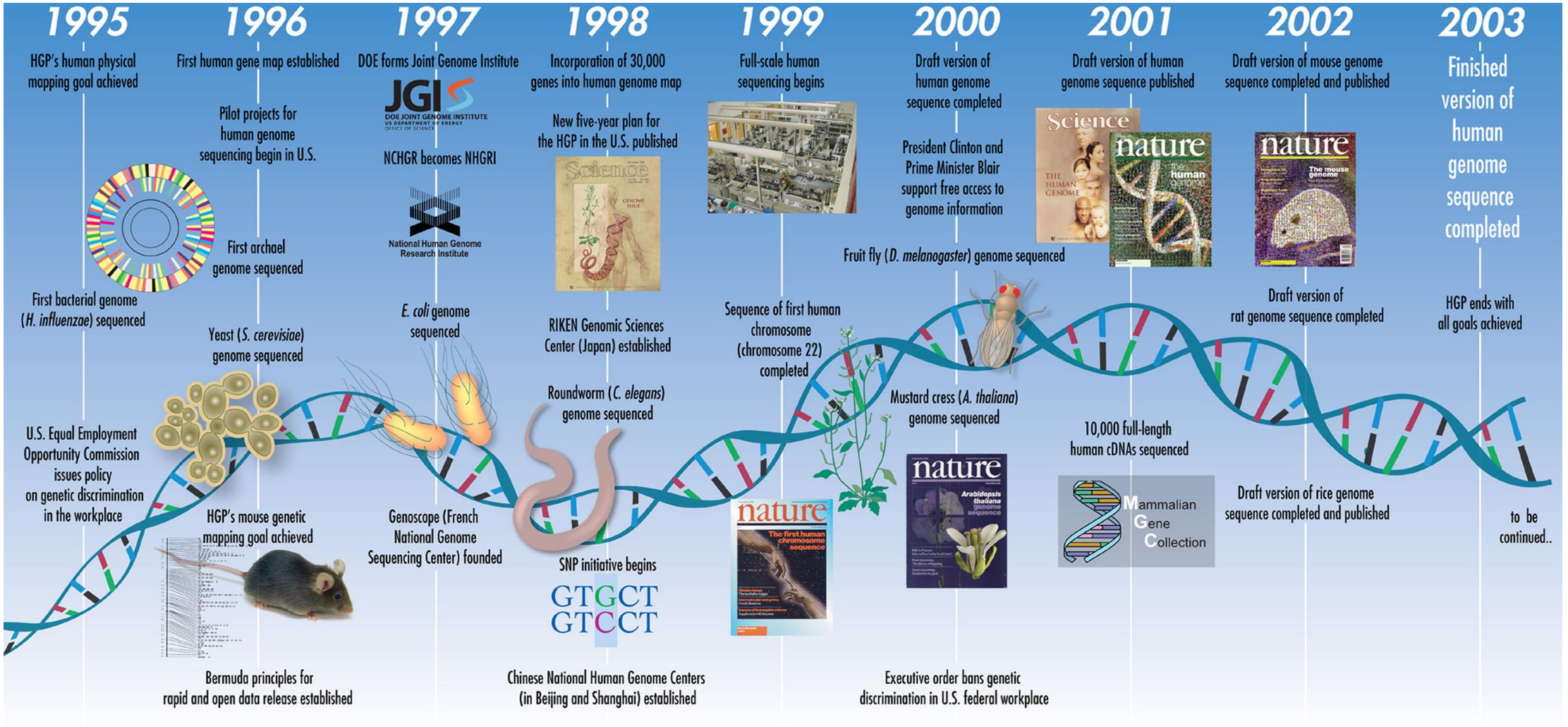
La traduction en vidéo

Cours 2:
Recherche d'information dans les génomes

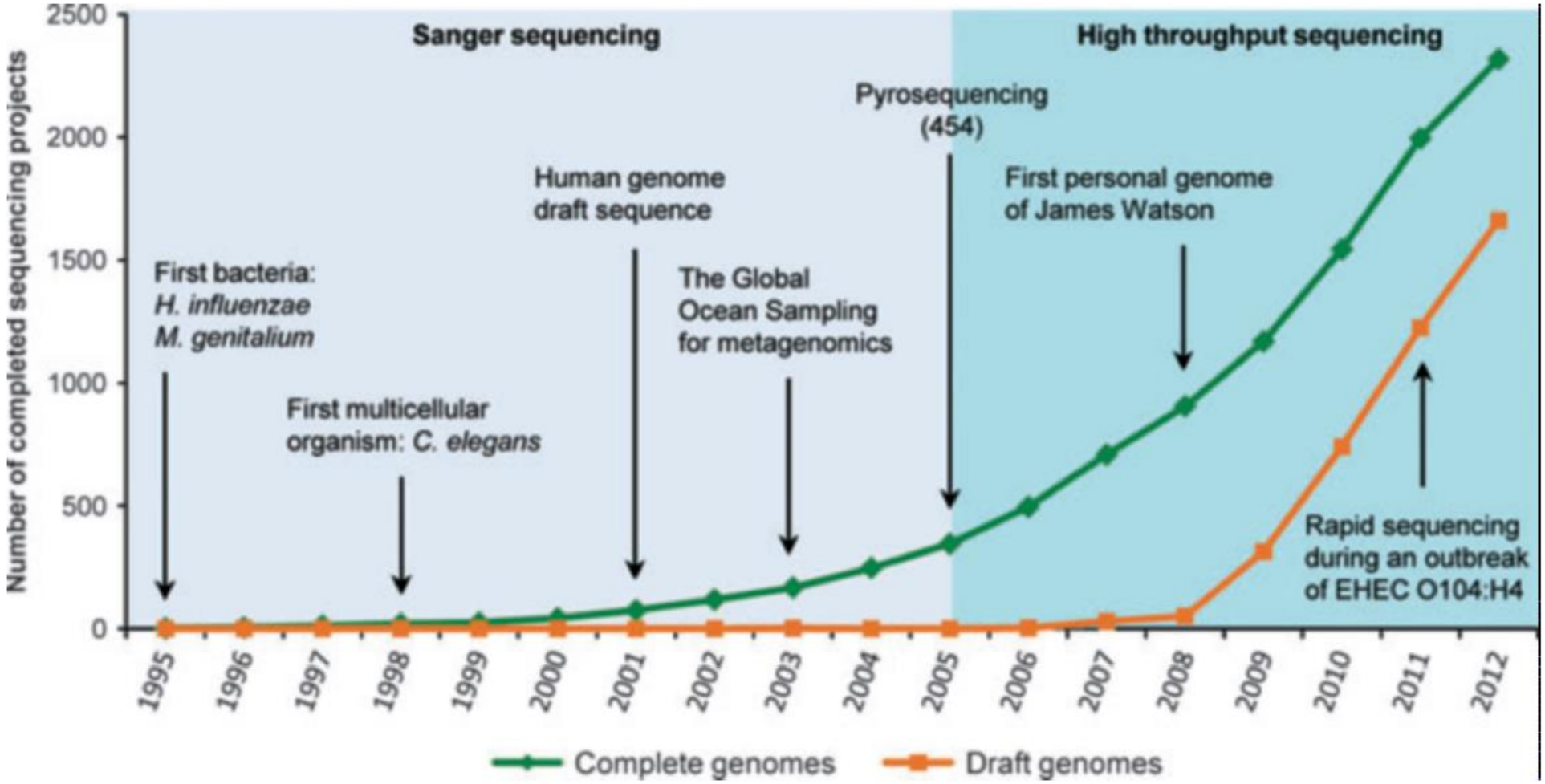
De la séquence du génome à son annotation

1) A la recherche des séquences codants
2) Les motifs d'activation des séquences codantes

Entrée de la biologie dans l'ère du Big Data:
Les programmes de séquençage des génomes



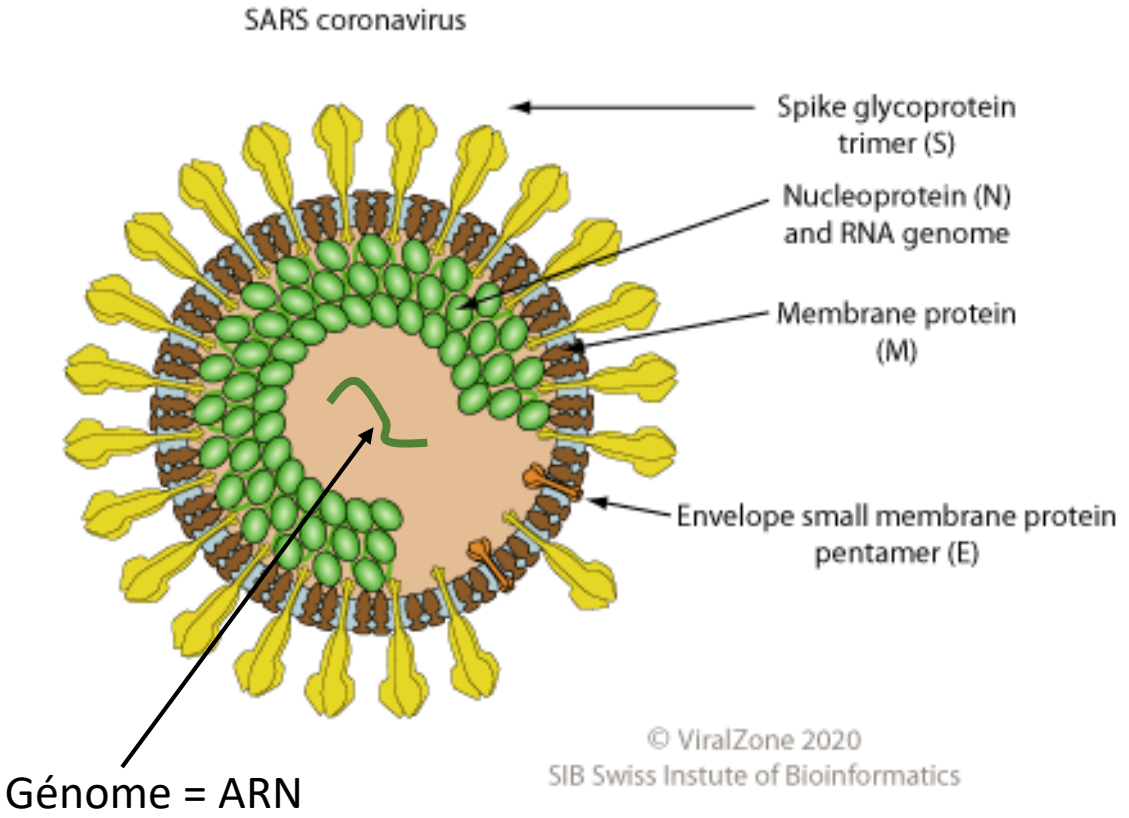
Entrée de la biologie dans l'ère du Big Data:
Une augmentation exponentielle des données de génomique



1) Recherche des séquences codantes

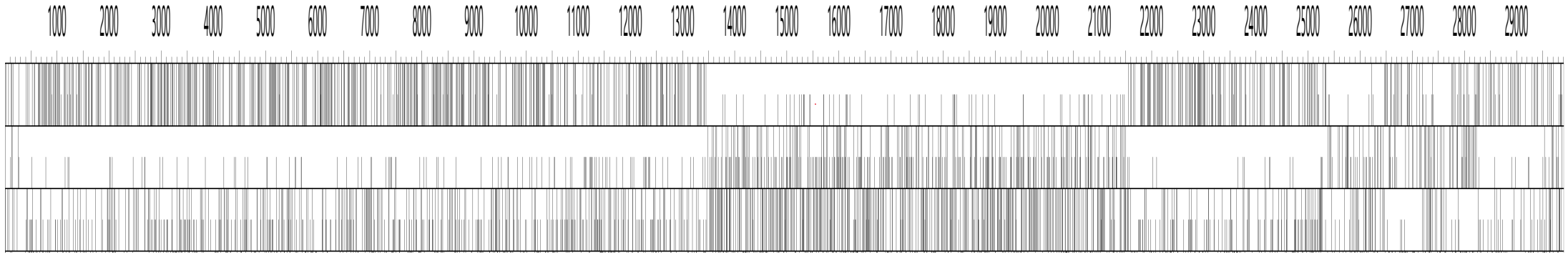
Exemple 1: le génome du sarv-co2
https://viralzone.expasy.org/30

Un génome particulier:
- Compact: 29 903 nucléotide
- Un unique brin d'ARN
- Séquence au format Fasta (ID: NC_045512.2):
https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2?report=fasta



Génome = ARN

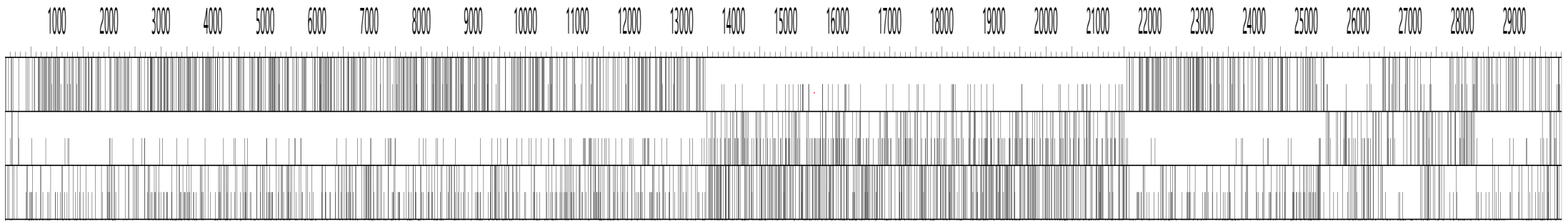
ORF MAP du genome du sarvcov-2



ORF (Open reading Frame): Région située entre deux codon STOP
CDS (Coding Séquence) Région située entre un START et un STOP incuse dans une ORF et codant potentiellement une protéine.

.

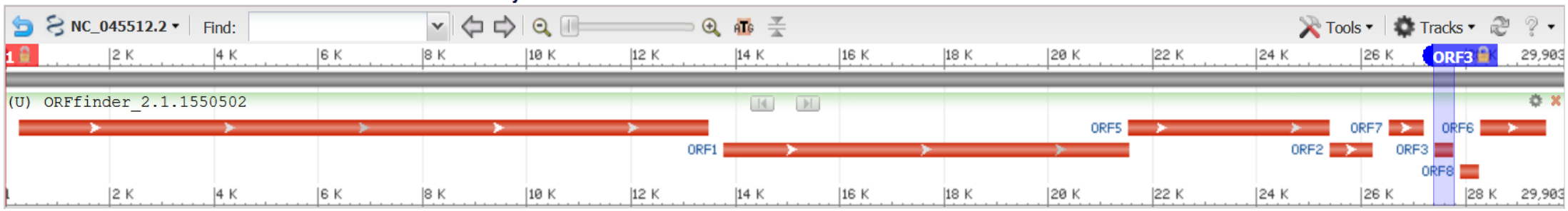
ORF MAP du genome du sarvcov-2



Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome

ORFs found: 8 **Genetic code: 1** **Start codon: 'ATG' only** **Nested ORFs removed**

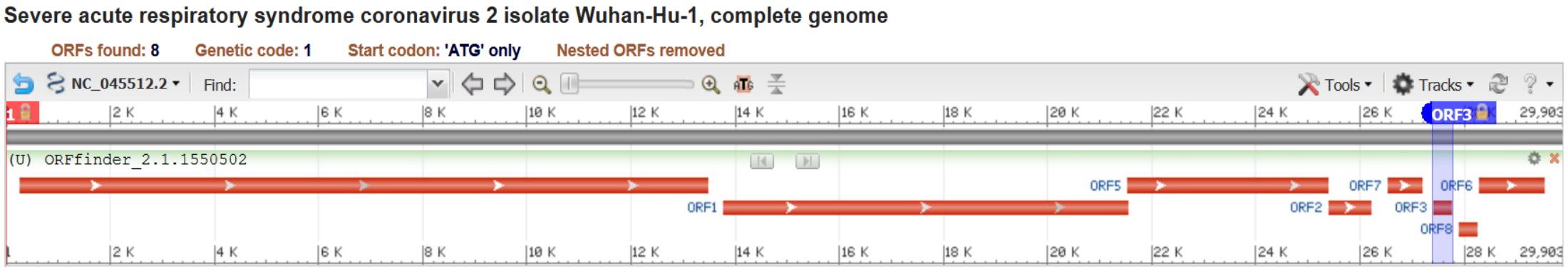
ORF Finder



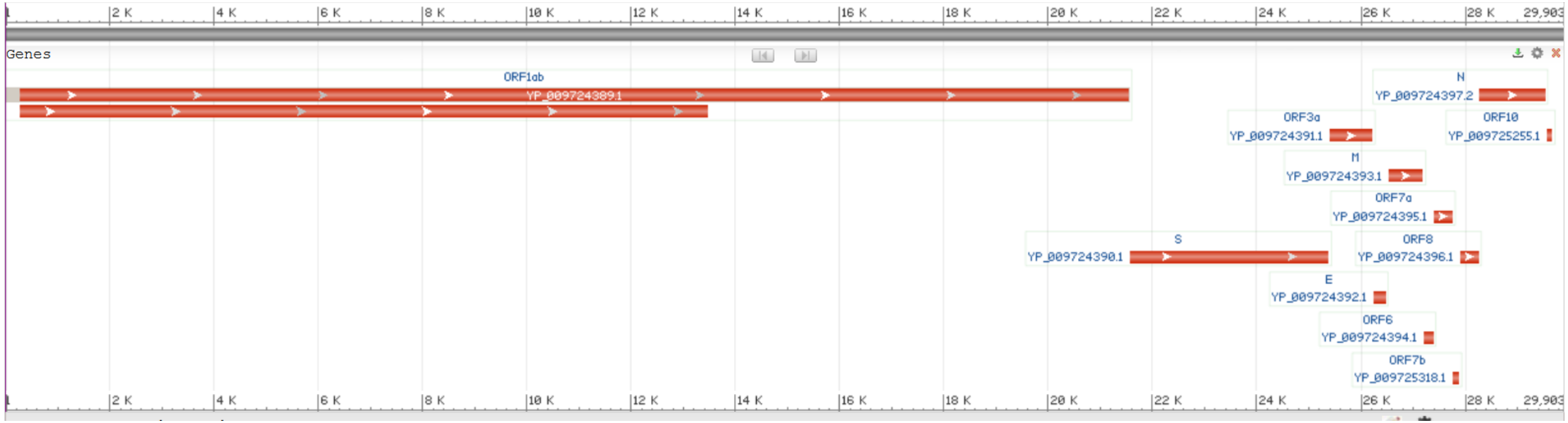
Label	Strand	Frame	Start	Stop	Length (nt | aa)
ORF1	+	1	13768	21555	7788 | 2595
ORF2	+	1	25393	26220	828 | 275
ORF3	**+**	**1**	**27394**	**27759**	**366 | 121**
ORF4	+	2	266	13483	13218 | 4405
ORF5	+	2	21536	25384	3849 | 1282
ORF6	+	2	28274	29533	1260 | 419
ORF7	+	3	26523	27191	669 | 222
ORF8	+	3	27894	28259	366 | 121

ORF MAP du genome du sarvcov-2

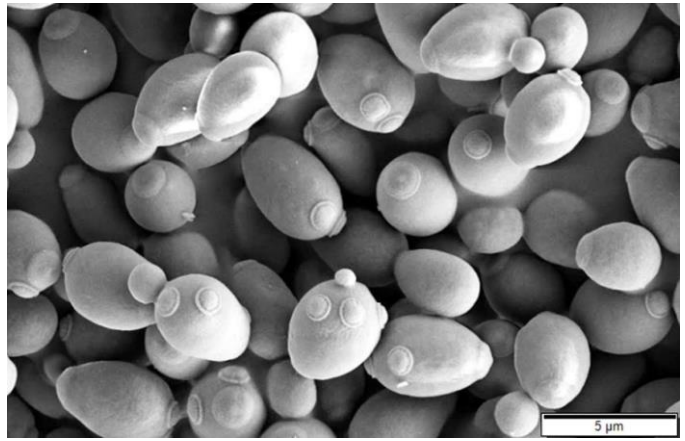
ORF Finder



Annotation biologique dans sequence viewer



1) Recherche des séquences codantes

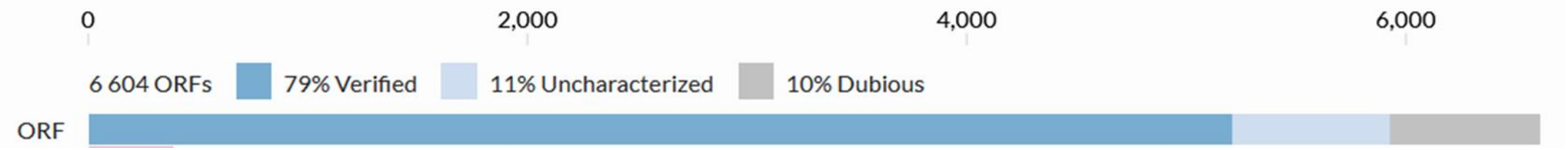


Exemple 2: Recherche dans le génome de la levure *S. cerevisiae*

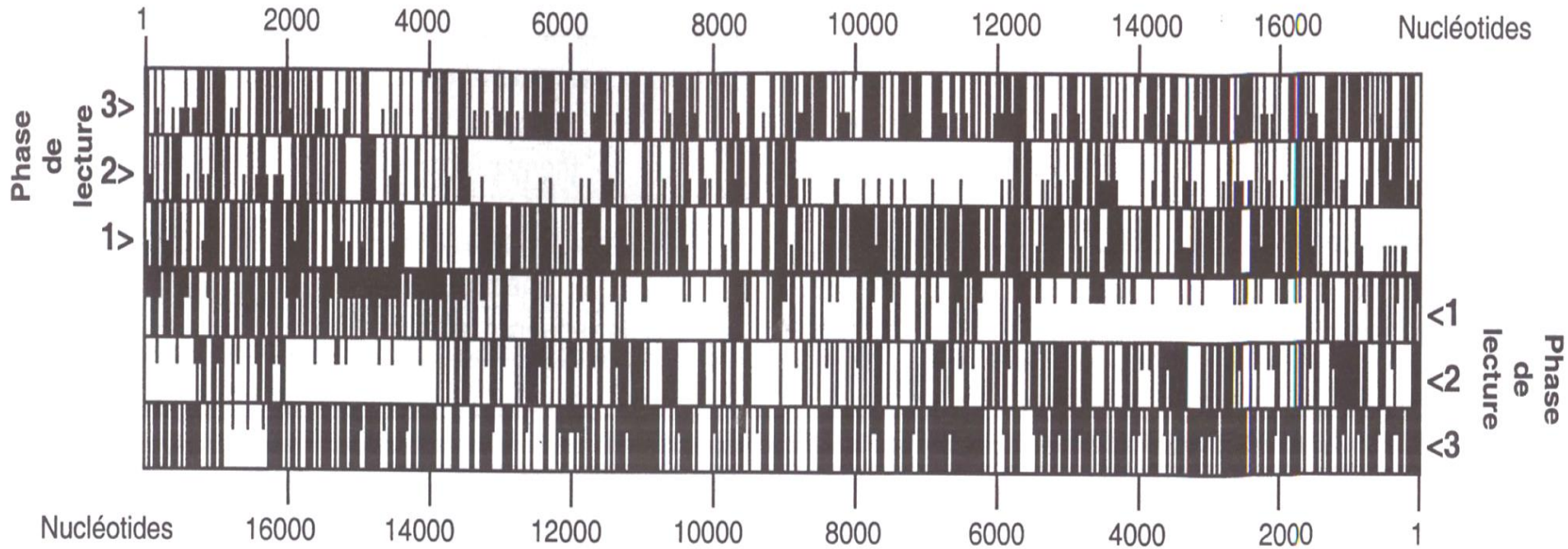
Une base de donnée très détaillée: SGD (Saccharomyces Genome Database)
https://www.yeastgenome.org/genomesnapshot#genome-inventory

- 17 molécules d'ADN: 16 chromosome nucléaire + 1 chromosome mitochondrial
- Taille total du génome nucléaire: 12 071 326 paire de bases (pb)

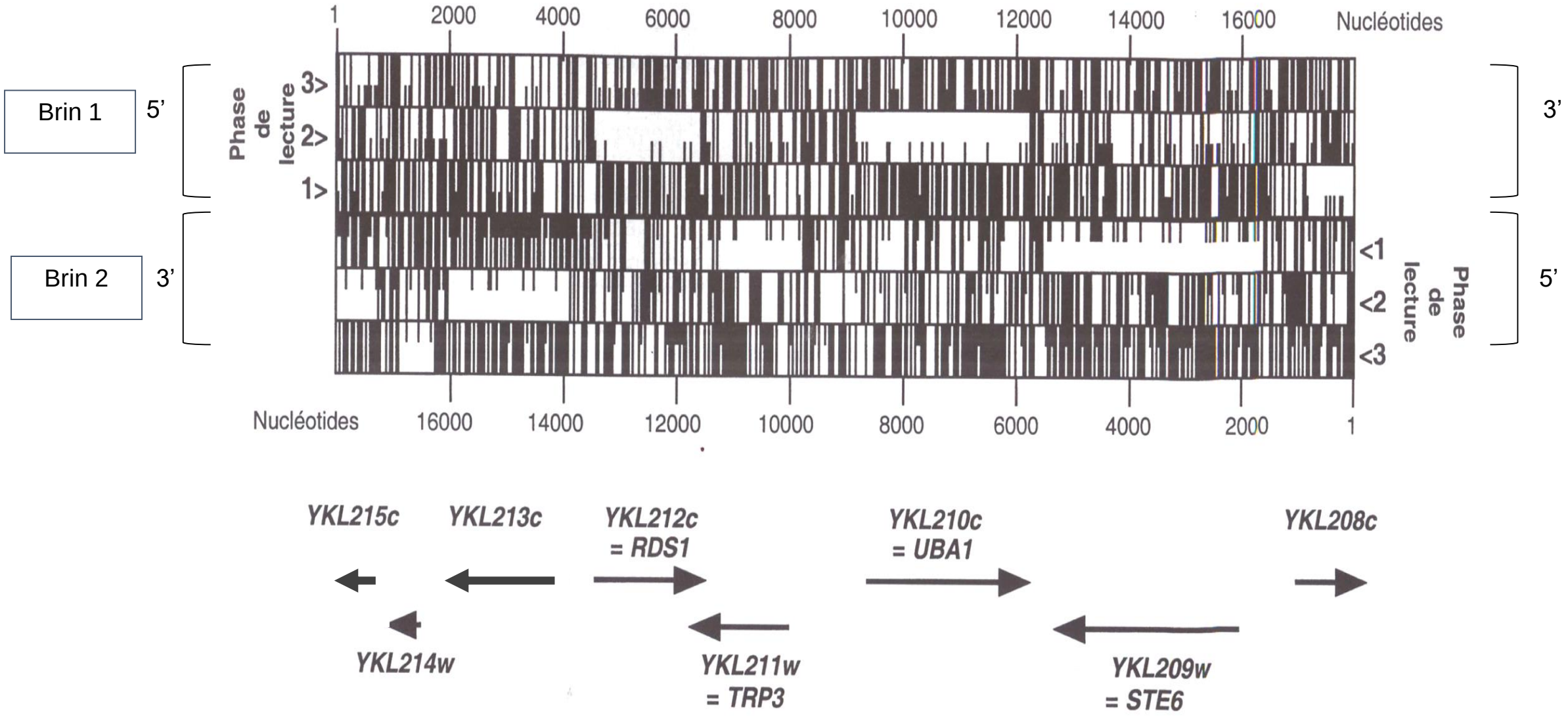
Un génome très bien annoté



ORF MAP d'un fragment de chromosome de *S. cerevisiae*

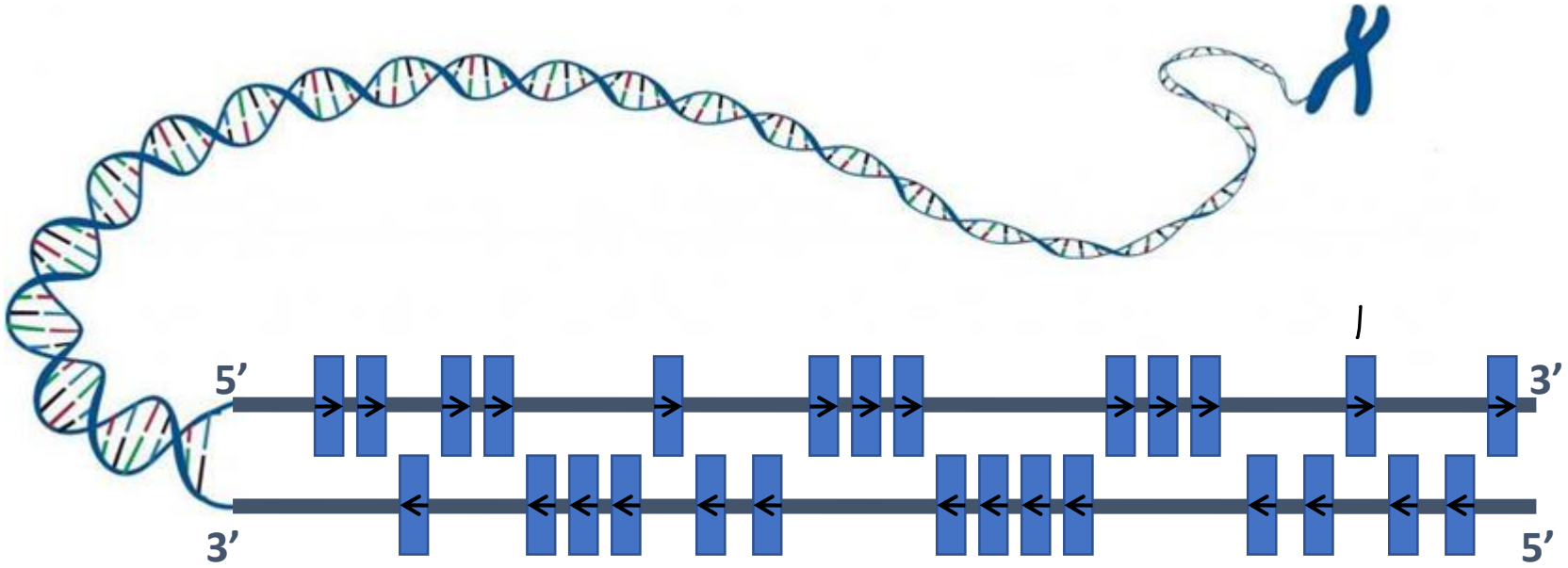


ORF MAP d'un fragment de chromosome de *S. cerevisiae*



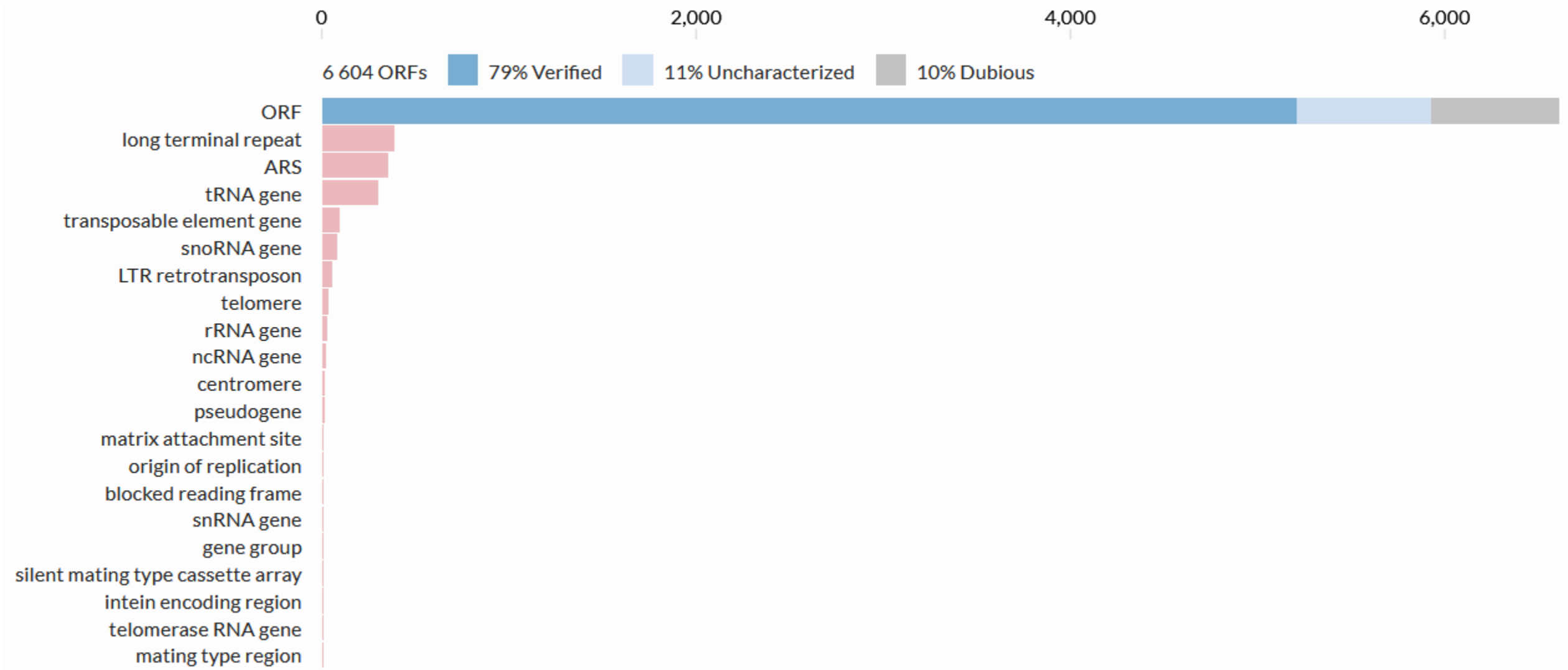
Bilan

1 chromosome= 1 molécule d'ADN = 2 brins d'ADN avec des séquences complémentaires
=> des centaines/milliers de séquences codant des protéines (CDS)



<u>Détail d'une CDS</u>

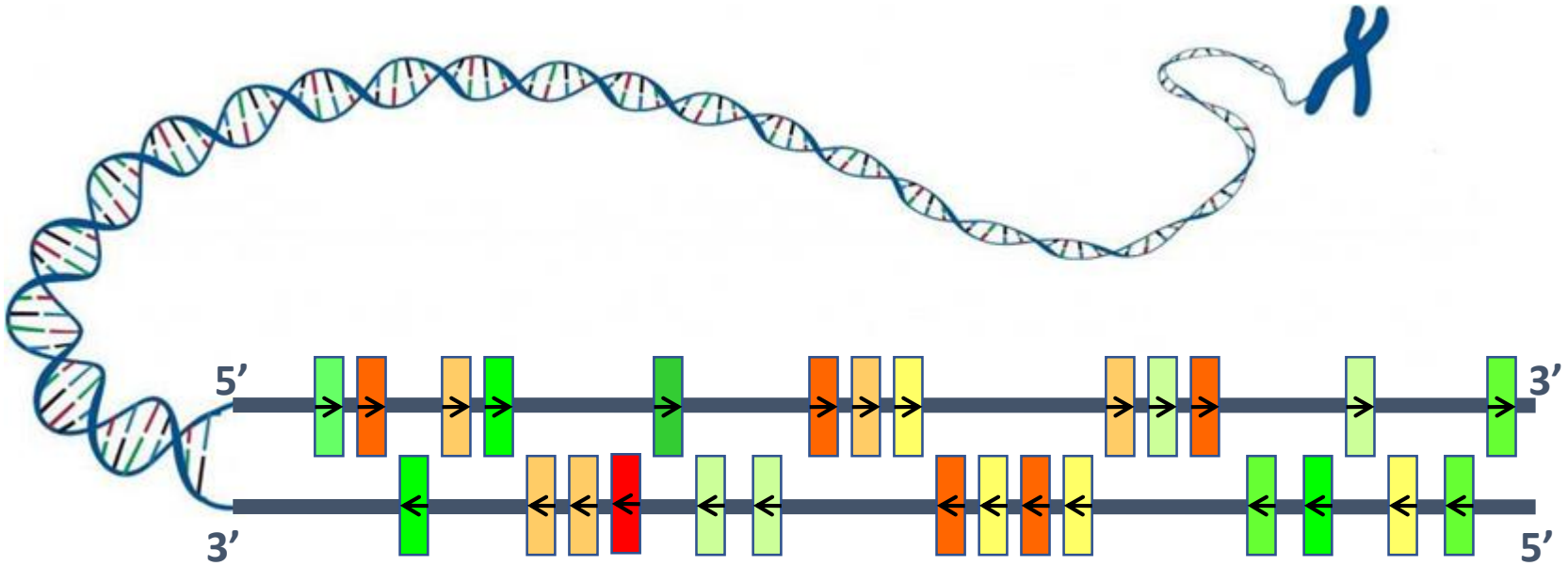
Les gènes codants ne sont pas les seules information contenues dans les génomes



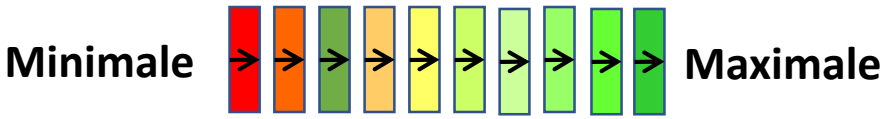
Différents types de séquences annotées dans le génome de *S. cerevisiae*

2) Les motifs d'activation des séquences codantes

1 chromosome= 1 molécule d'ADN = 2 brins d'ADN avec des séquences complémentaires
⇒des centaines/milliers de séquences codant des protéines (CDS)



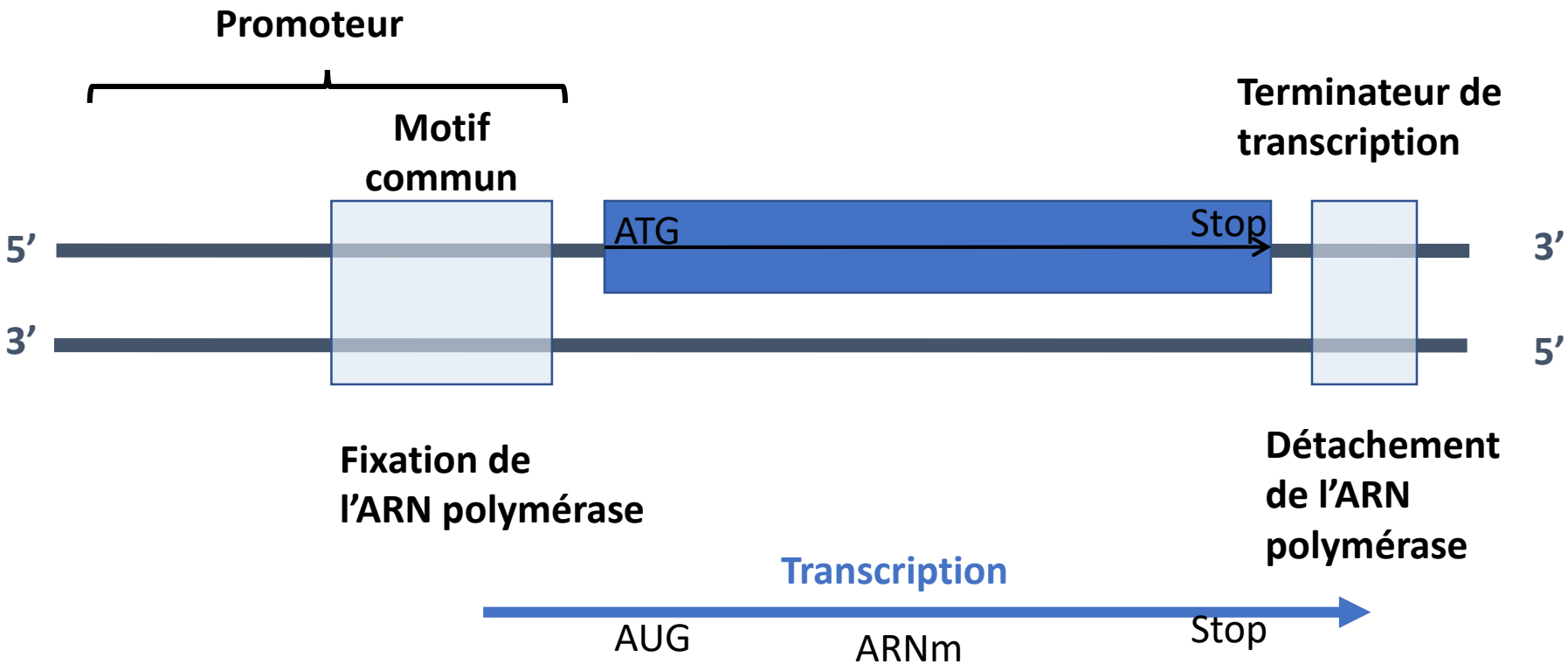
⇒Une activation variable des CDS en fonction des conditions cellulaire et environnementale



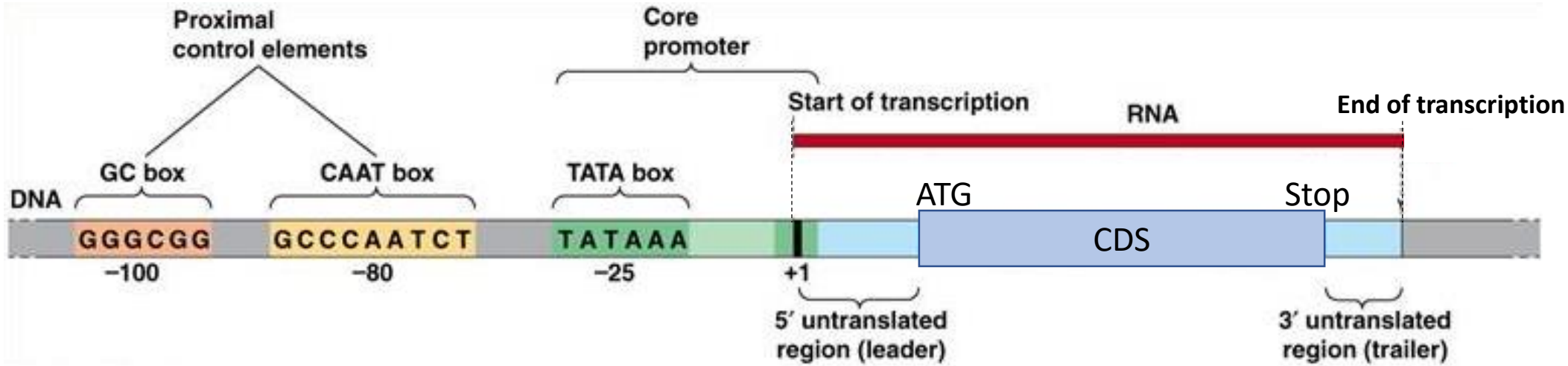
Echelle de niveau d'activation

Le promoteur des CDS:
le site d'activation de la transcription des CDS

Détail d'une CDS avec son promoteur

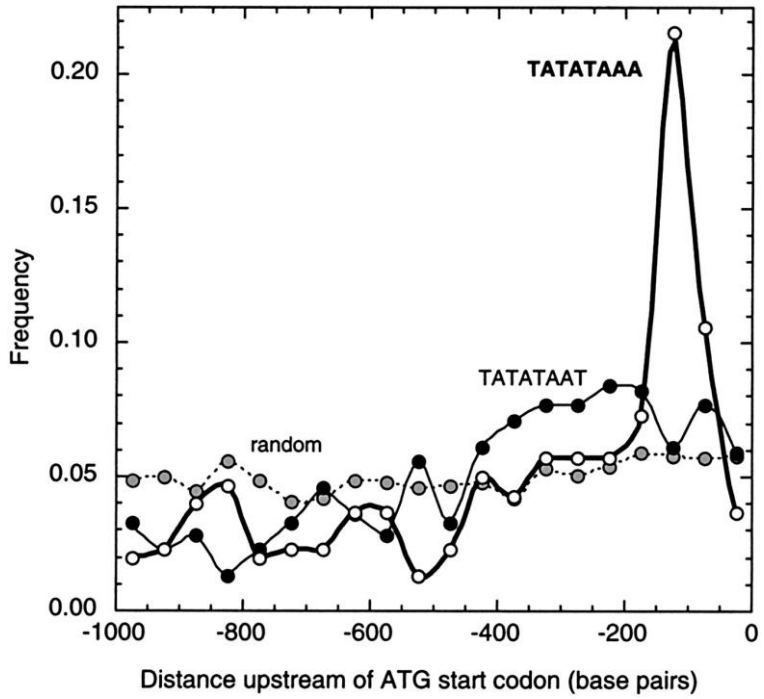


Détail de la structure globale de la régions promotrice d'un gène eucaryote



Exemple de recherche de motifs:
Recherche de la TATA box dans le génome de la levure *S. cerevisiae*

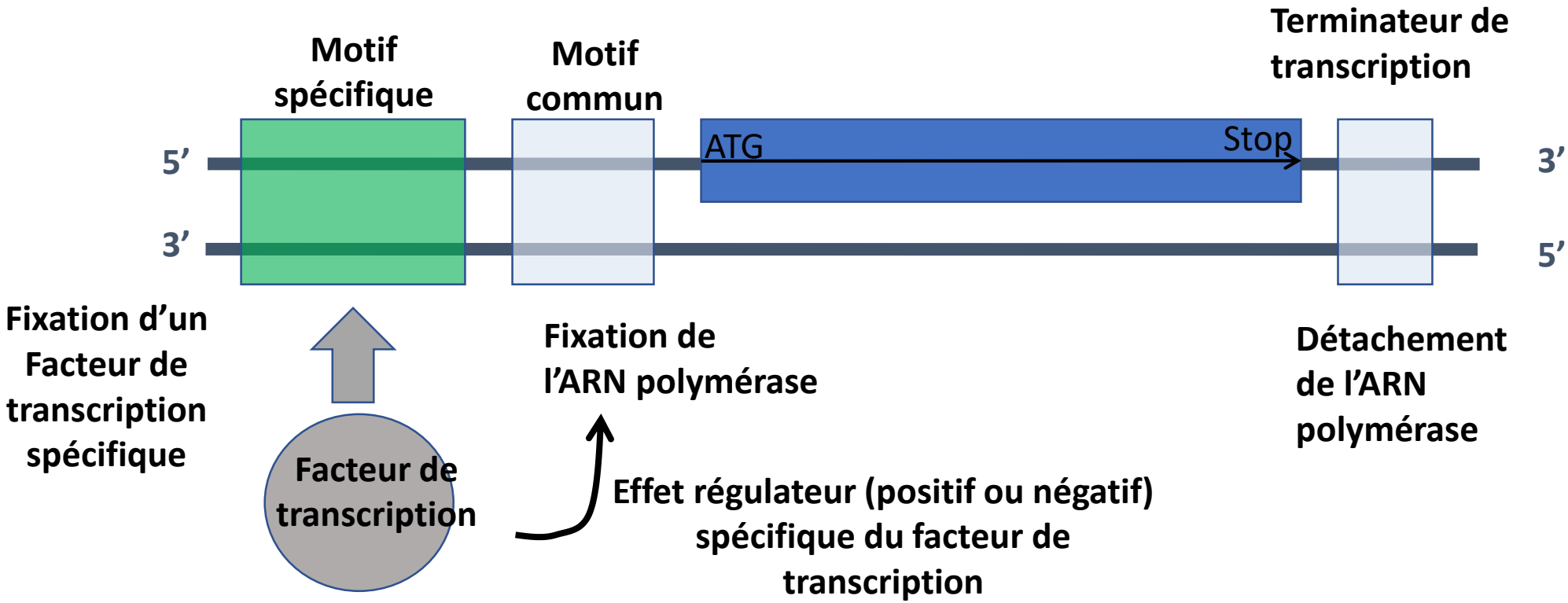
Consensus du motif TATA box : TATA(A/T)A(A/T)(A/G) → IUPAC nucleotide code: TATAWAWR
Position de la TATA box:
20 à 40 bases avant le site initiation de la transcription, ie 50 à 200 base avant l'ATG initiateur



→ Utilisation d'un outil en ligne de recherche de motif:
http://rsat.sb-roscoff.fr/genome-scale-dna-pattern.cgi

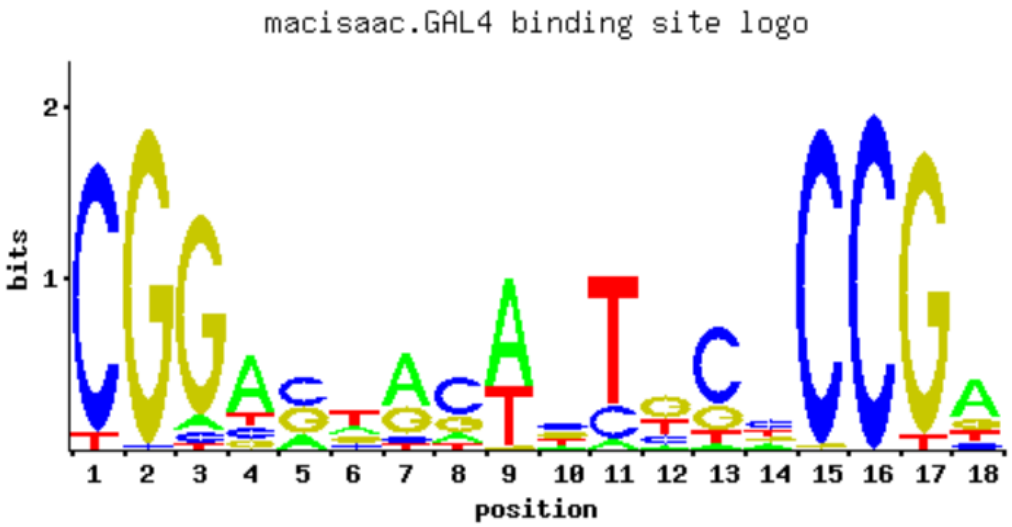
1477 occurrence sur 6604 CDS: ~ 20% des CDS

Le promoteur des CDS:
une région régulatrice de l'activation génique



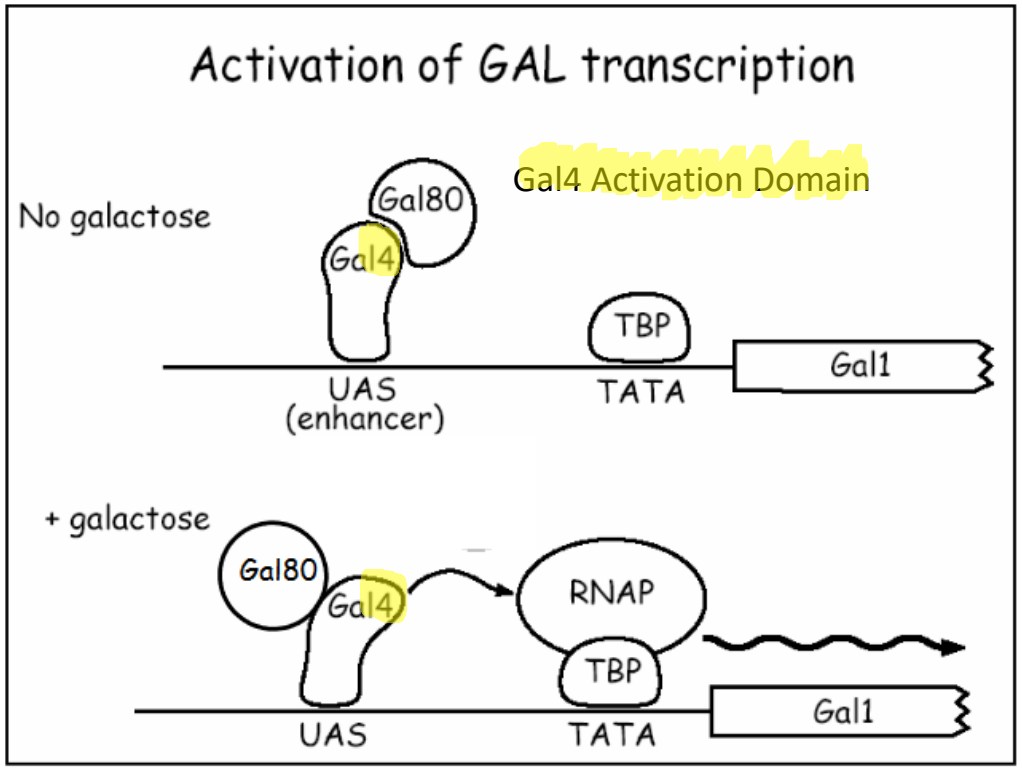
Exemple de motifs régulateur:
Le motif de recrutement du facteur de transcription Gal4

Details for GAL4 from Macisaac et. al

TF	Logo
GAL4 Matrices: PFM PCM PWM Data Source: Macisaac et. al	

Position Count Matrix for Macisaac et. al GAL4:

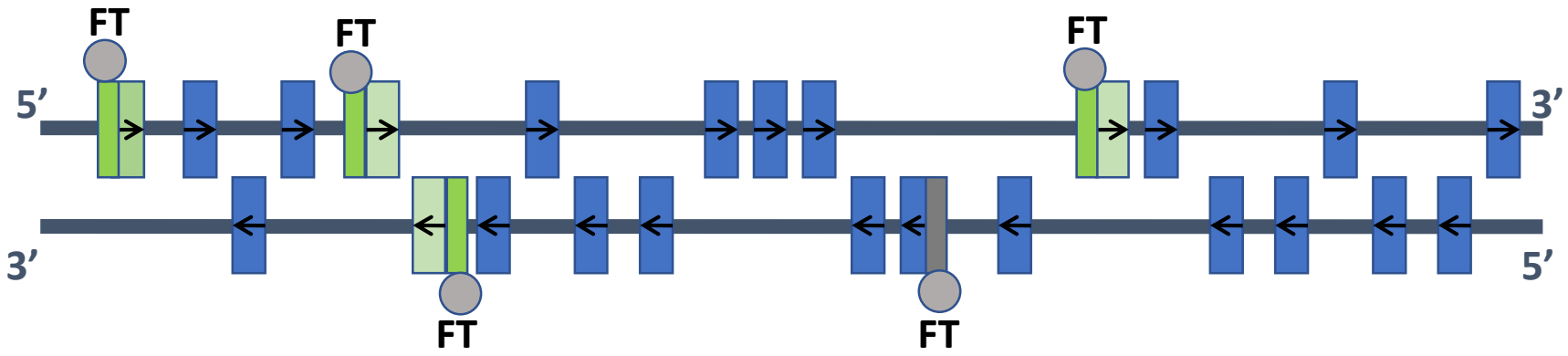
A |	0	0	5	64	18	19	57	14	64	11	5	6	3	18	0	0	0	57
C |	94	1	4	12	42	14	7	56	0	36	19	18	67	42	98	99	0	9
G |	0	98	88	10	37	18	29	21	1	31	0	47	21	18	1	0	95	17
T |	5	0	1	13	1	47	5	7	34	20	74	28	9	20	0	0	4	15



→ Utilisation d'un outil en ligne de recherche de motif:
http://rsat.sb-roscoff.fr/genome-scale-dna-pattern.cgi
Motif code IUPAC: CGGNNNNNNNNNNCCG
189 occurrence sur 6604 CDS (2%) dont 13 dans des gènes du métabolisme du galactose

Répartition des motifs régulateurs dans les génome
et co-régultation des gènes impliqués dans un même processus

FT: Facteur de transcription



Projet: comment découvrir les motifs régulateurs de FT dans un génome?