DÉCHIFFREMENT(S):
des hiéroglyphes à l'ADN

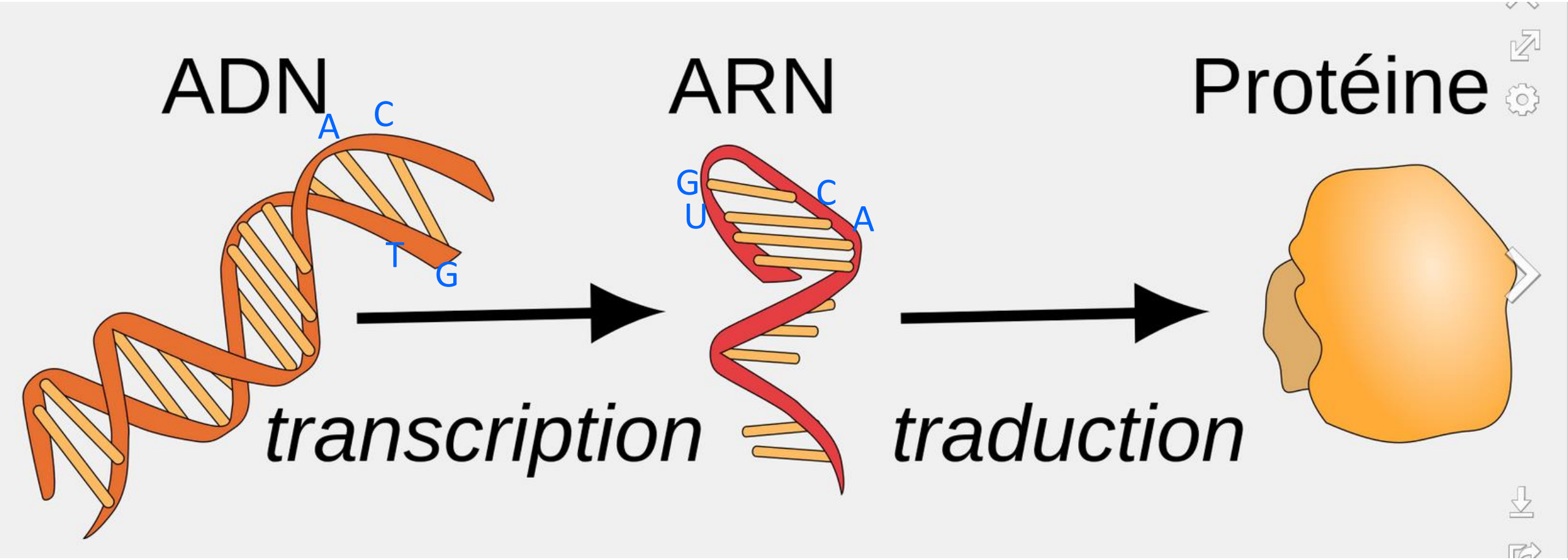
Que comprenons-nous du Code Génétique en 2022 ?

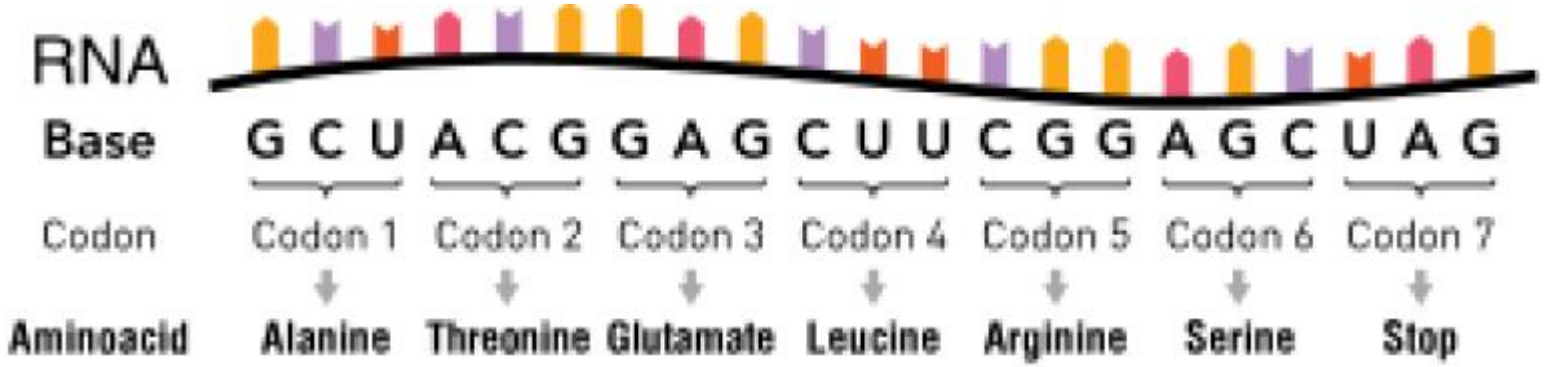
Jean Weissenbach


Ceci n'est pas le résultat
d'un déchiffrement


L'ADN est un HYPERTEXTE

Le dogme central de la biologie moléculaire





Comparaison de séquences protéiques

```
Query:   RYKELTEQQMPGALPPECTPNMDGPHARSVRREQSLHSFHTLFCRRCFKYDRFLH
         +YKELTEQQ+PGALPPECTPN+DGP+A+SV+REQSLHSFHTLFCRRCFKYD FLH
Subject: KYKELTEQQLPGALPPECTPNIDGPNAKSVQREQSLHSFHTLFCRRCFKYDCFLH

Query:   LLFQLFLALSDLKQLRILHTDLKPDNVMLVD--EKELKIKLMDFGLALLTHEAKT--GTI
         +L Q+  AL  LK L ++H DLKP+N+MLVD   +  ++K++DFG A  +H +KT   T
Subject: ILQQVATALKKLKSLGLIHADLKPENIMLVDPVRQPYRVKVIDFGSA--SHVSKTVCSTY
```

1. Base	2. Base: U	2. Base: C	2. Base: A	2. Base: G
U	UUU Phenylalanin (Phe)	UCU Serin (Ser)	UAU Tyrosin (Tyr)	UGU Cystein (Cys)
	UUC Phenylalanin (Phe)	UCC Serin (Ser)	UAC Tyrosin (Tyr)	UGC Cystein (Cys)
	UUA Leucin (Leu)	UCA Serin (Ser)	UAA *Stop*	UGA *Stop**
	UUG Leucin (Leu)	UCG Serin (Ser)	UAG *Stop*	UGG Tryptophan (Trp)
C	CUU Leucin (Leu)	CCU Prolin (Pro)	CAU Histidin (His)	CGU Arginin (Arg)
	CUC Leucin (Leu)	CCC Prolin (Pro)	CAC Histidin (His)	CGC Arginin (Arg)
	CUA Leucin (Leu)	CCA Prolin (Pro)	CAA Glutamin (Gln)	CGA Arginin (Arg)
	CUG Leucin (Leu)	CCG Prolin (Pro)	CAG Glutamin (Gln)	CGG Arginin (Arg)
A	AUU Isoleucin (Ile)	ACU Threonin (Thr)	AAU Asparagin (Asn)	AGU Serin (Ser)
	AUC Isoleucin (Ile)	ACC Threonin (Thr)	AAC Asparagin (Asn)	AGC Serin (Ser)
	AUA Isoleucin (Ile)	ACA Threonin (Thr)	AAA Lysin (Lys)	AGA Arginin (Arg)
	AUG Methionin (Met)*	ACG Threonin (Thr)	AAG Lysin (Lys)	AGG Arginin (Arg)
G	GUU Valin (Val)	GCU Alanin (Ala)	GAU Asparaginsäure (Asp)	GGU Glycin (Gly)
	GUC Valin (Val)	GCC Alanin (Ala)	GAC Asparaginsäure (Asp)	GGC Glycin (Gly)
	GUA Valin (Val)	GCA Alanin (Ala)	GAA Glutaminsäure (Glu)	GGA Glycin (Gly)
	GUG Valin (Val)	GCG Alanin (Ala)	GAG Glutaminsäure (Glu)	GGG Glycin (Gly)

Remarques sur le code génétique

Le code ne s’est pas constitué de manière aléatoire

Il permet de minimiser les erreurs de réplication, transcription ou traduction

Une fois constitué il s’est maintenu inchangé

Il est universel

Robustesse du code génétique aux erreurs

L’organisation du code génétique va assurer une minimalisation des erreurs

erreur en position 3 souvent sans conséquence (wobble)
erreur en position 2 changement majeur d’aminoacide
erreur en position 1 changement d’aminoacide avec effet pouvant être modéré

1. Base	2. Base: U	2. Base: C	2. Base: A	2. Base: G
U	UUU Phenylalanin (Phe)	UCU Serin (Ser)	UAU Tyrosin (Tyr)	UGU Cystein (Cys)
	UUC Phenylalanin (Phe)	UCC Serin (Ser)	UAC Tyrosin (Tyr)	UGC Cystein (Cys)
	UUA Leucin (Leu)	UCA Serin (Ser)	UAA *Stop*	UGA *Stop**
	UUG Leucin (Leu)	UCG Serin (Ser)	UAG *Stop*	UGG Tryptophan (Trp)
C	CUU Leucin (Leu)	CCU Prolin (Pro)	CAU Histidin (His)	CGU Arginin (Arg)
	CUC Leucin (Leu)	CCC Prolin (Pro)	CAC Histidin (His)	CGC Arginin (Arg)
	CUA Leucin (Leu)	CCA Prolin (Pro)	CAA Glutamin (Gln)	CGA Arginin (Arg)
	CUG Leucin (Leu)	CCG Prolin (Pro)	CAG Glutamin (Gln)	CGG Arginin (Arg)
A	AUU Isoleucin (Ile)	ACU Threonin (Thr)	AAU Asparagin (Asn)	AGU Serin (Ser)
	AUC Isoleucin (Ile)	ACC Threonin (Thr)	AAC Asparagin (Asn)	AGC Serin (Ser)
	AUA Isoleucin (Ile)	ACA Threonin (Thr)	AAA Lysin (Lys)	AGA Arginin (Arg)
	AUG Methionin (Met)*	ACG Threonin (Thr)	AAG Lysin (Lys)	AGG Arginin (Arg)
G	GUU Valin (Val)	GCU Alanin (Ala)	GAU Asparaginsäure (Asp)	GGU Glycin (Gly)
	GUC Valin (Val)	GCC Alanin (Ala)	GAC Asparaginsäure (Asp)	GGC Glycin (Gly)
	GUA Valin (Val)	GCA Alanin (Ala)	GAA Glutaminsäure (Glu)	GGA Glycin (Gly)
	GUG Valin (Val)	GCG Alanin (Ala)	GAG Glutaminsäure (Glu)	GGG Glycin (Gly)

Questions en suspens

Origine et évolution du code génétique (codons = triplets)

Pourquoi ces aminoacides et pas d'autres ? Pourquoi juste 20

Universalité

Robustesse aux erreurs, non-maximale

Premiers pas du séquençage des acides nucléiques

Les techniques de séquençage de l'ARN furent développées en premier

- 1965 Holley: yeast $tRNA^{Ala}$

- 1967 Sanger: ^{32}P RNA *E. coli* 5S rRNA

- 1969 Sanger 57 nucléotides de l'ARN codant la capside du bactériophage R17

Nucleotide Sequence from the Coat Protein Cistron of R17 Bacteriophage RNA

by
J. M. ADAMS*
P. G. N. JEPPESEN
F. SANGER
B. G. BARRELL

MRC Laboratory of Molecular Biology, Cambridge

Abstract

The sequence of fifty-seven nucleotides in the coat protein cistron of phage R17 RNA directly confirms the genetic code, shows that the code used by the phage is degenerate

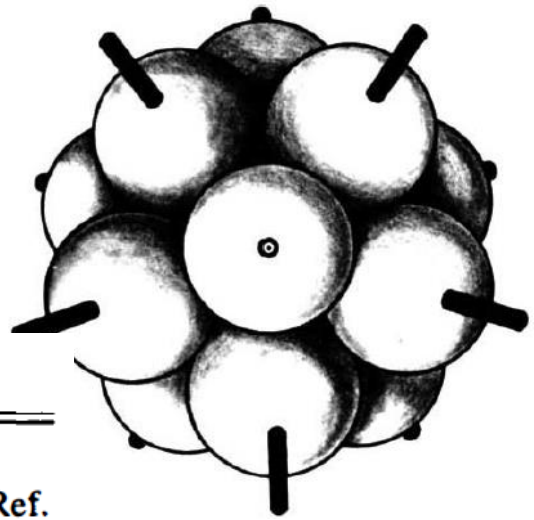
Nature, 1969

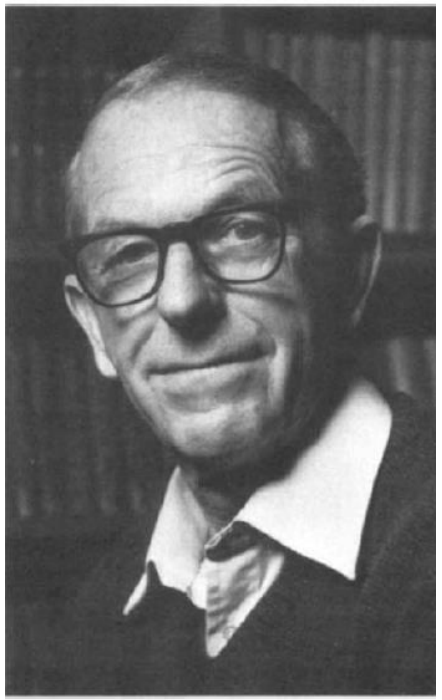
Nucleotide sequence of bacteriophage ΦX174 DNA

F. Sanger, G. M. Air*, B. G. Barrell, N. L. Brown†, A. R. Coulson, J. C. Fiddes, C. A. Hutchison III‡, P. M. Slocombe§ & M. Smith‖

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

ΦX174 sequenced





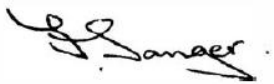


Table 1 The progress in sequencing

Year	Protein	RNA	DNA	Number of residues	Ref.
1935	Insulin			1	4
1945	Insulin			2	3
1947	Gramicidin S			5	16
1949	Insulin			9	12
1955	Insulin			51	22
1960	Ribonuclease			120	25
1965		$tRNA_{Ala}$		75	32
1967		5S RNA		120	35
1968			Bacteriophage λ	12	45
1978			Bacteriophage φX 174	5,386	61
1981			Mitochondria	16,569	58
1982			Bacteriophage λ	48,502	54
1984			Epstein-Barr virus	172,282	64

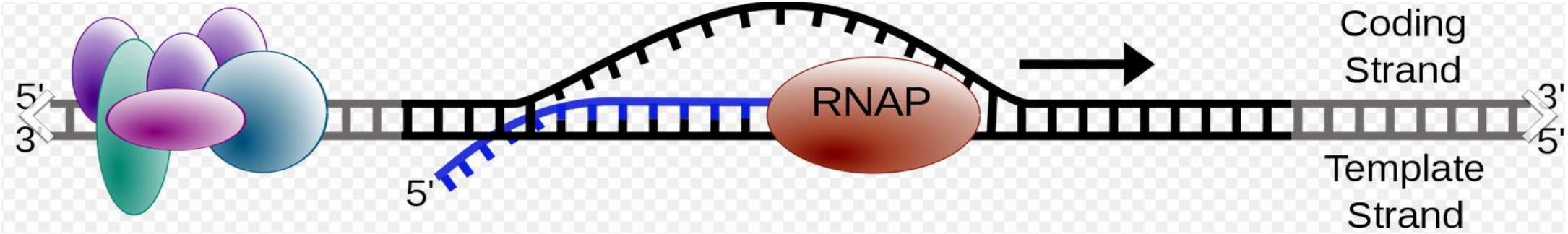
Les débuts des programmes informatiques d'analyse de séquence

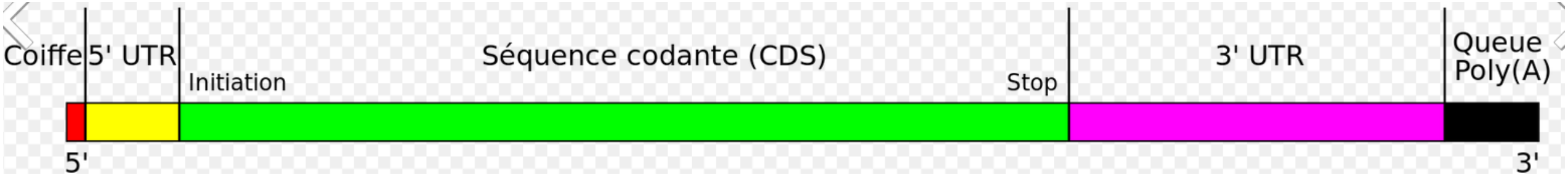
Ces premiers programmes comparent et alignent des séquences 2 par 2 et facilitent l'assemblage des fragments séquencés

1970 Needleman Wunsch ; 197- Staden ; 197- 198- Smith – Waterman ; 198- Lipman et col.

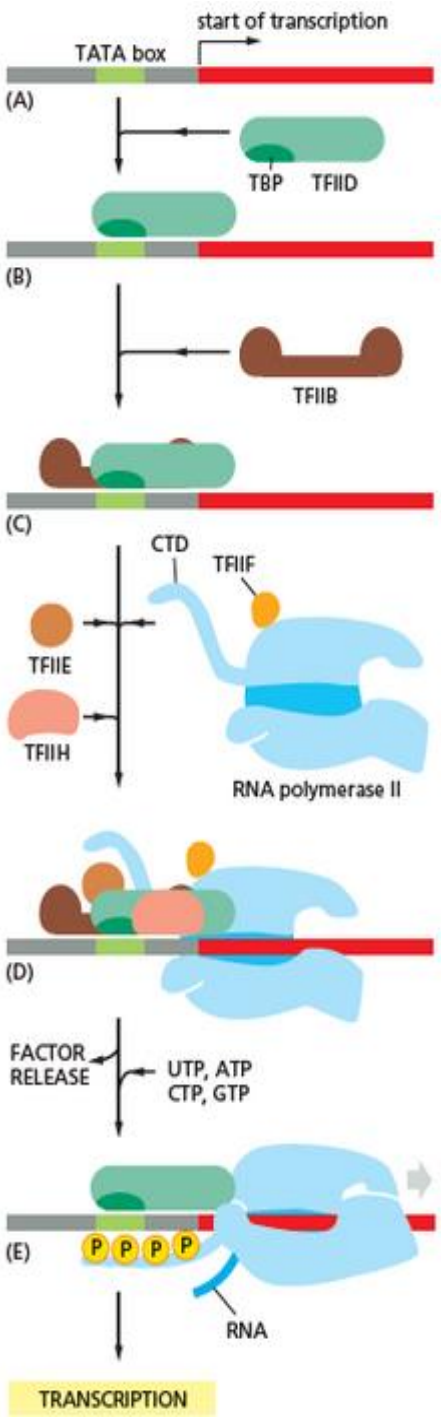
Les programmes qui suivront dans les années 80-90 rechercheront les différents types de gènes sur les séquences d'ADN (protéines, rARN, tARN)

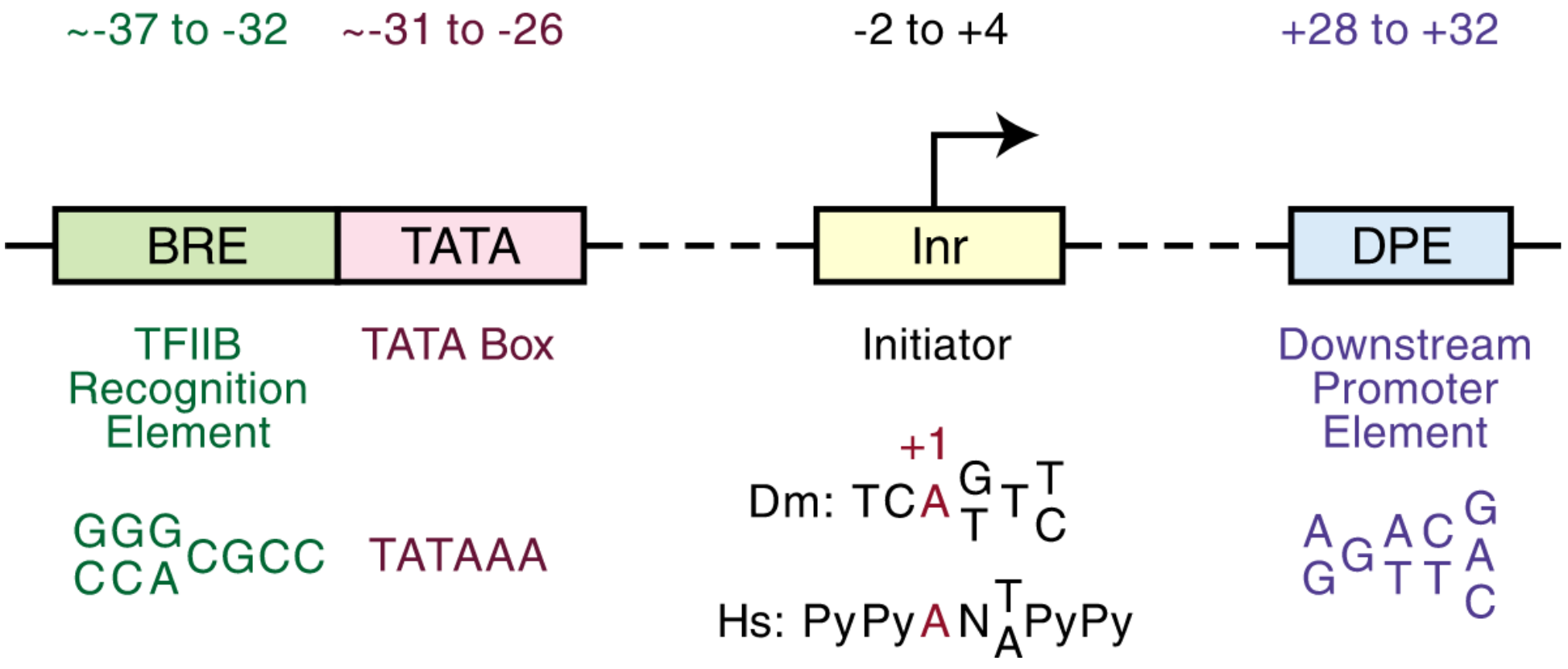
Transcription (gène eucaryote)



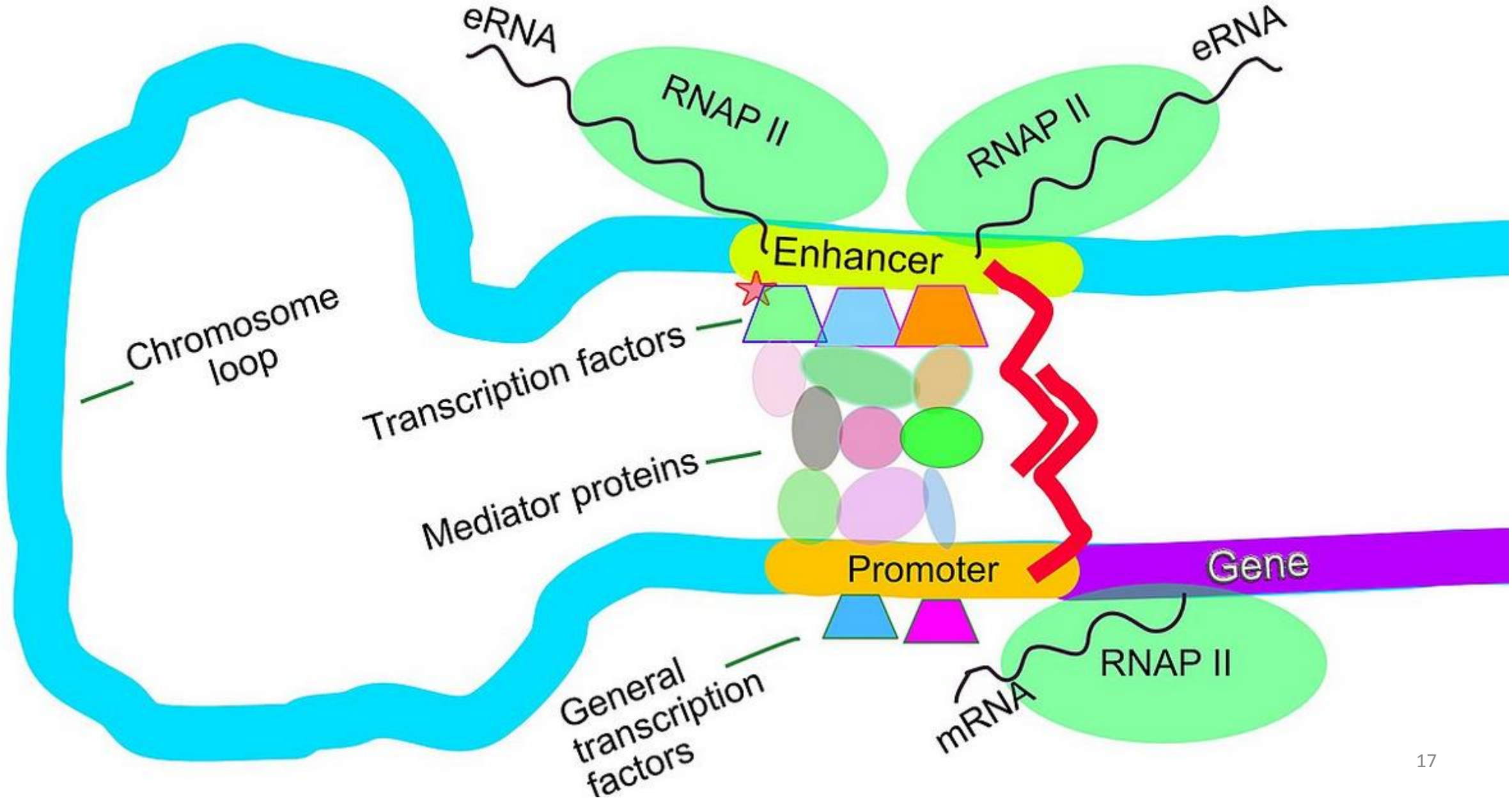


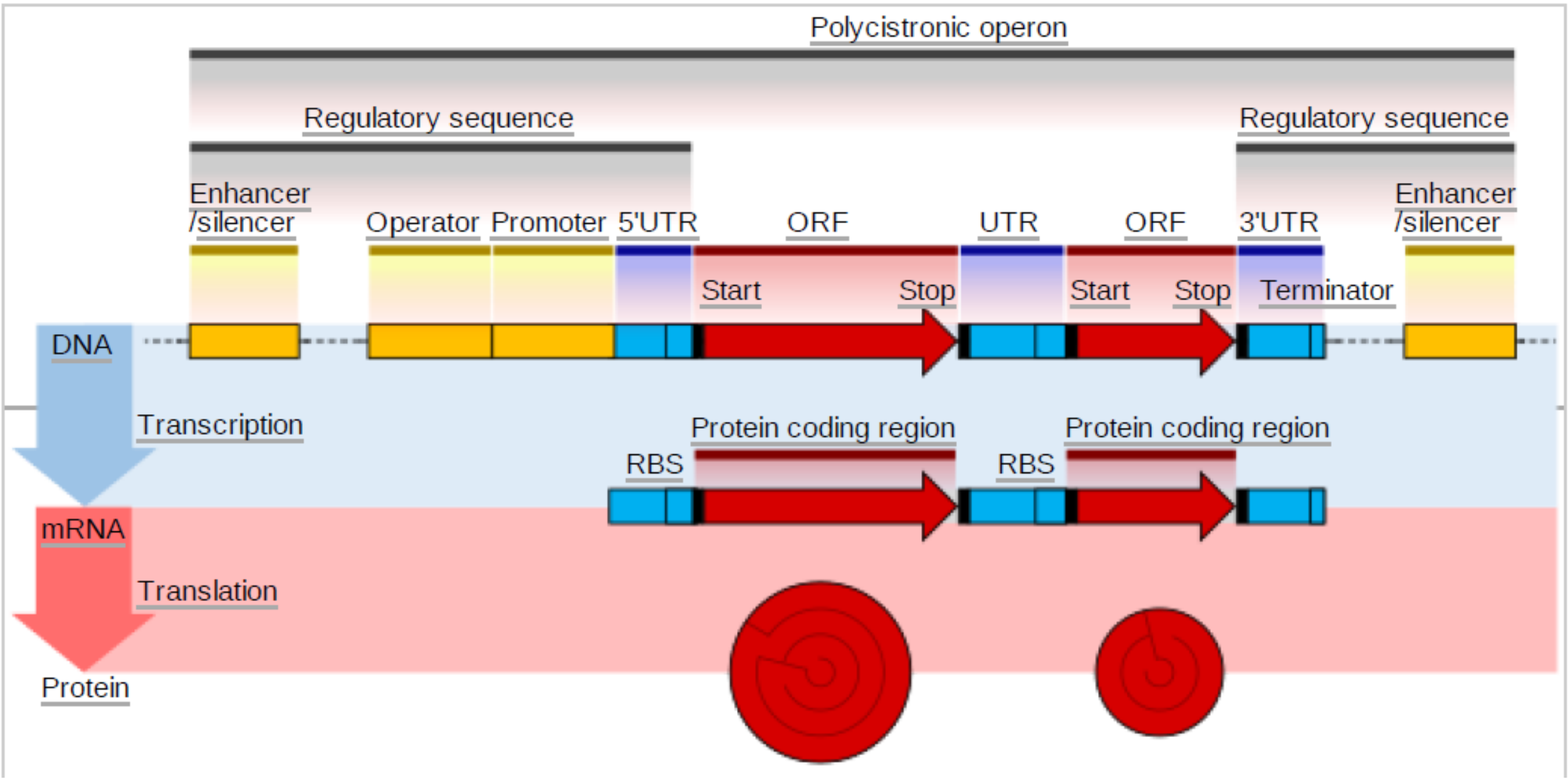
Signatures d'éléments de régulation de la transcription d'un gène eucaryote





Transcription et sa régulation





Shafee T, Lowe R (2017). "Eukaryotic and prokaryotic gene structure". *WikiJournal of Medicine* **4** (1)

Venter et al. 1995
Science

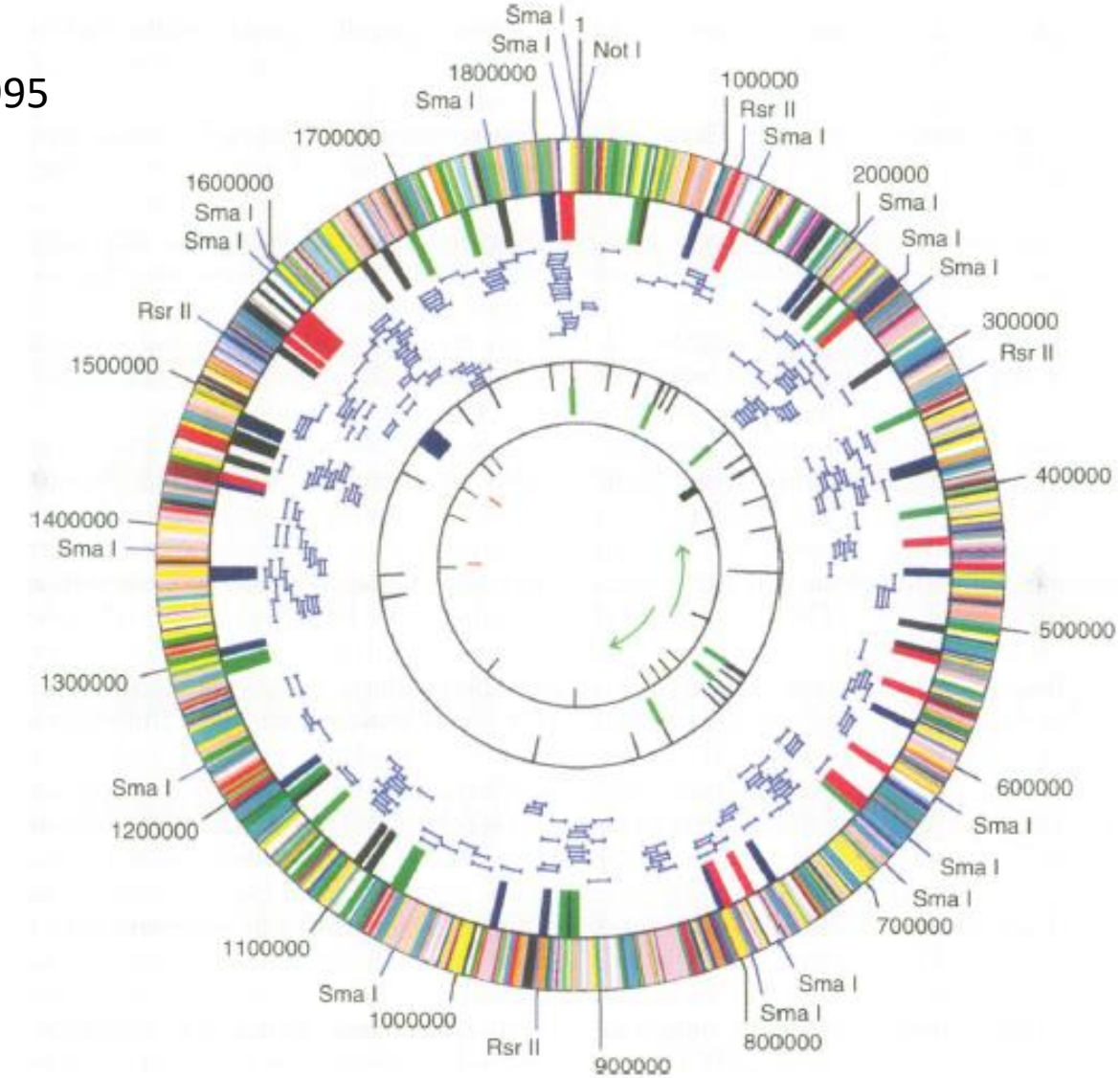
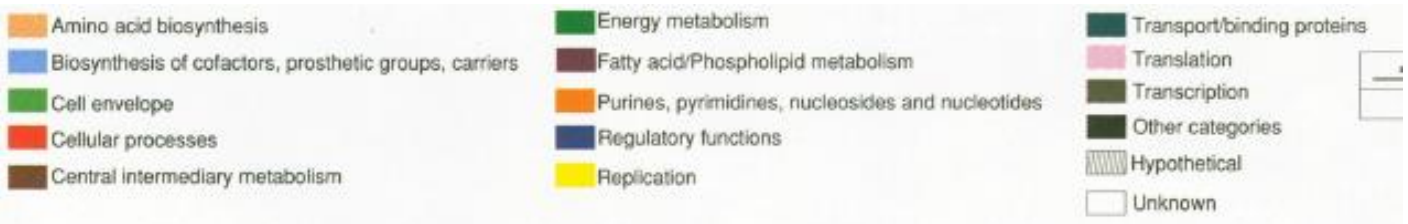
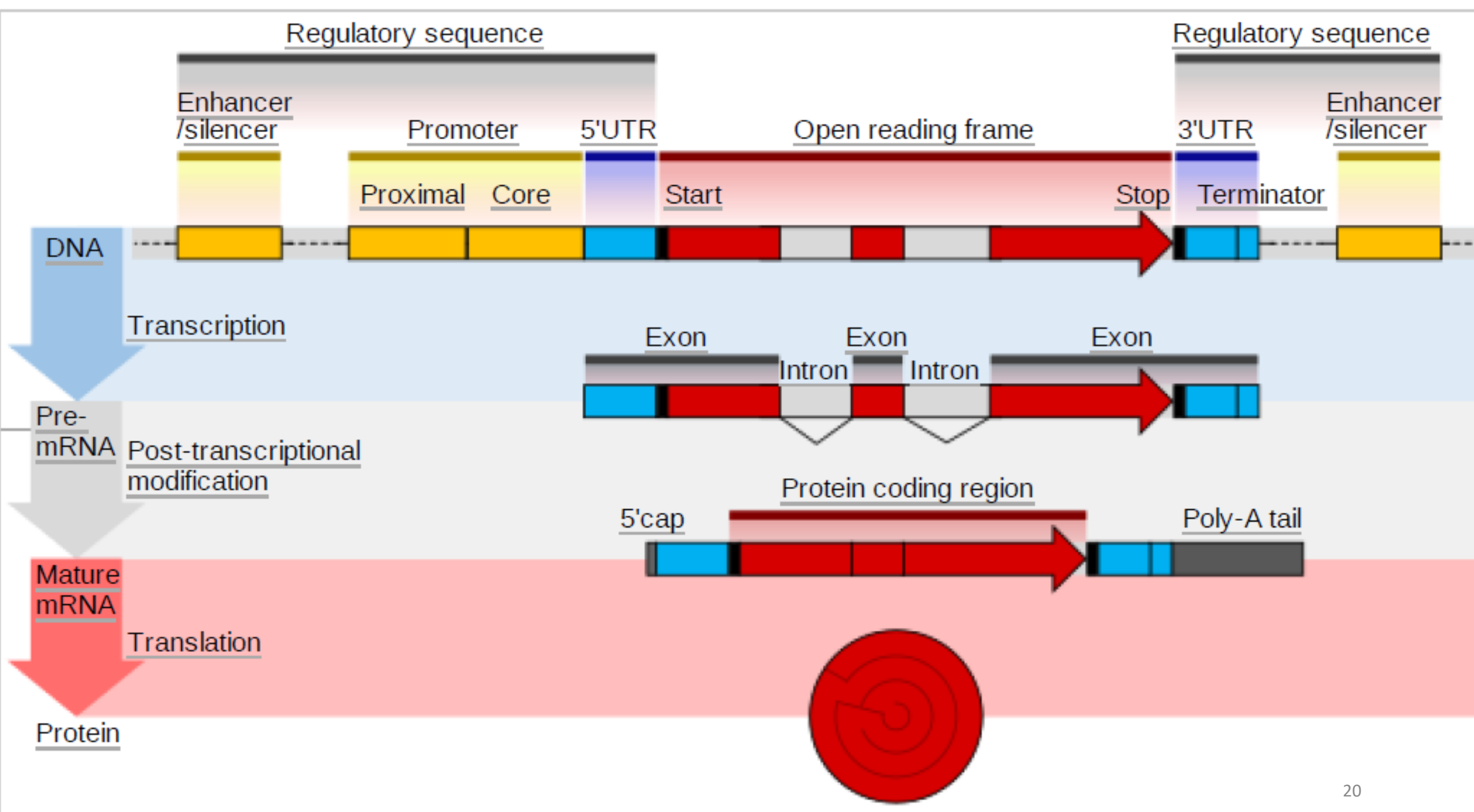
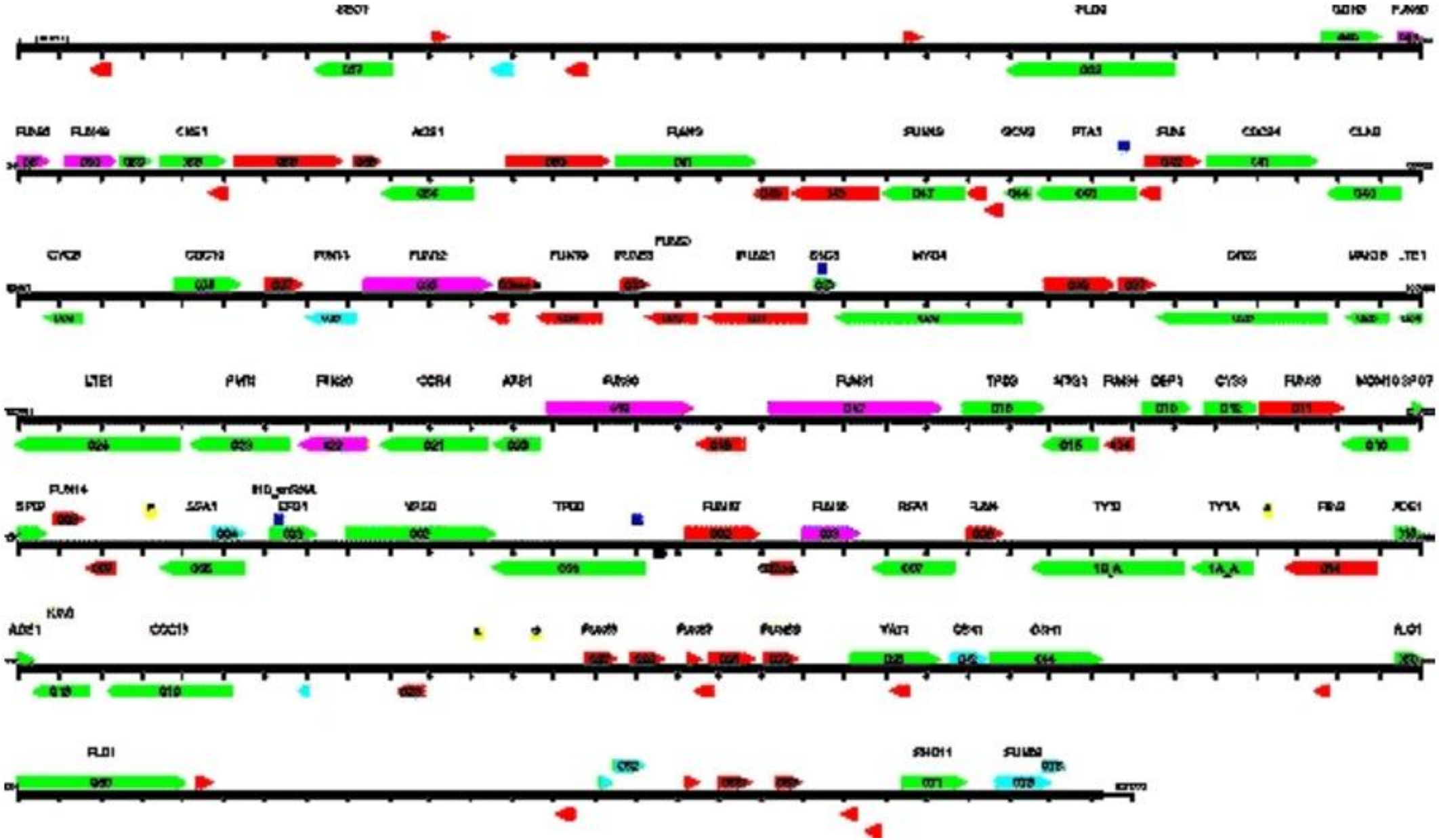


Fig. 1. A circular representation of the *H. influenzae* Rd chromosome illustrating the location of each

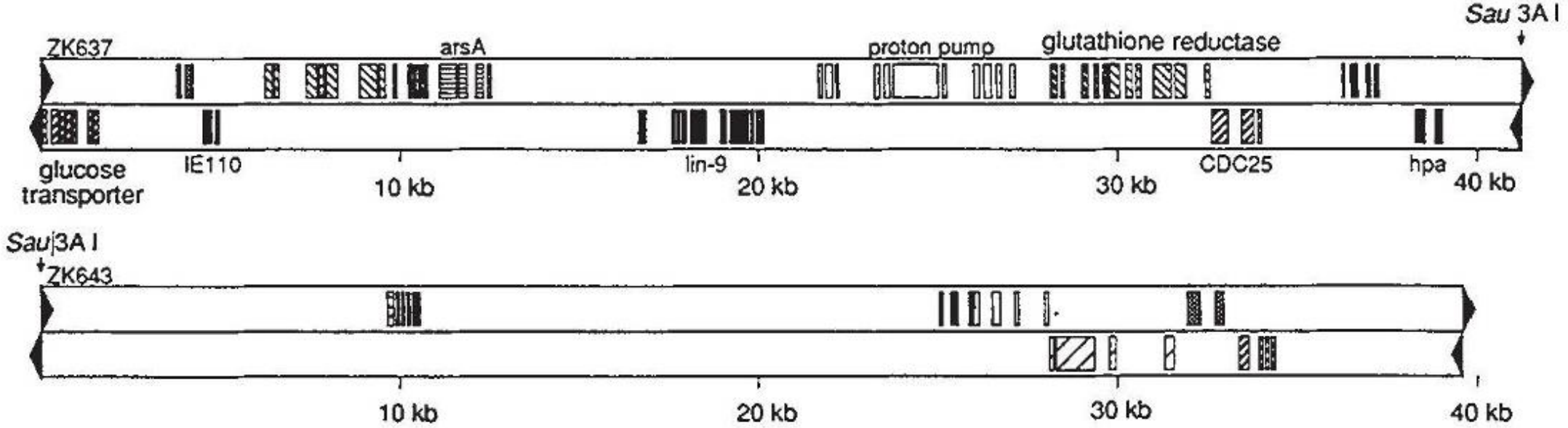


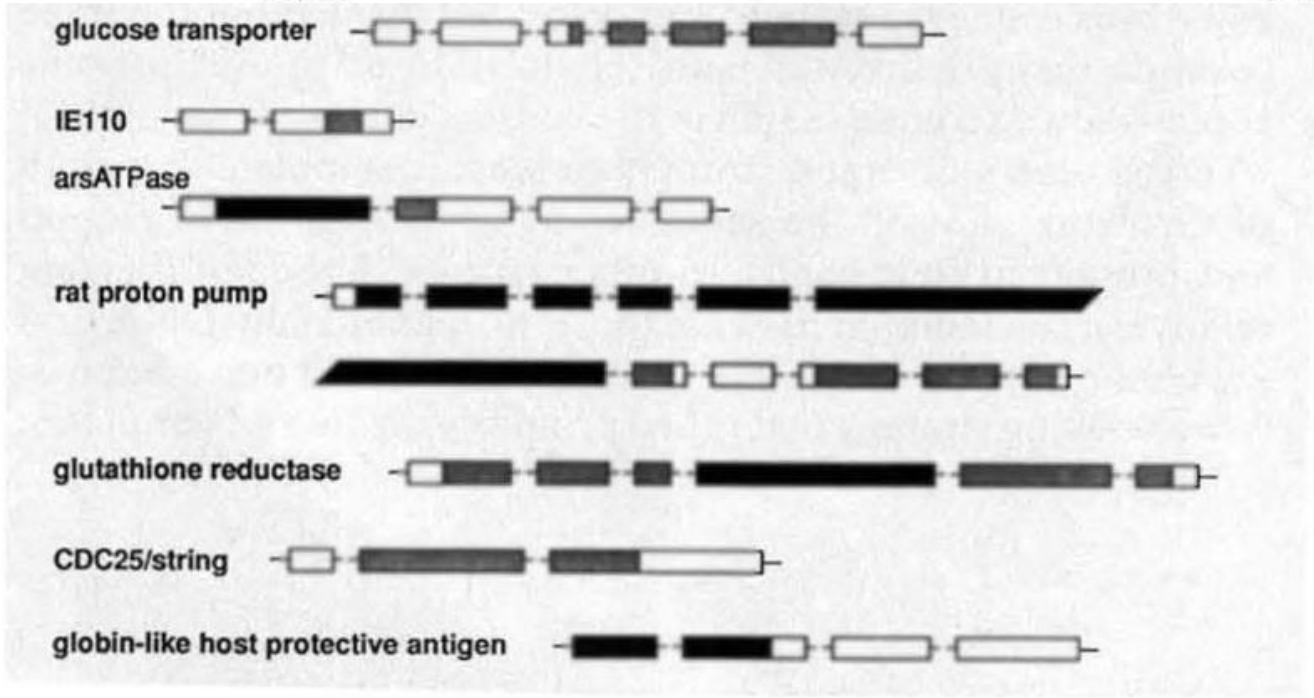

Regulatory sequence
Regulatory sequence
Enhancer /silencer
Promoter
5'UTR
Open reading frame
3'UTR
Enhancer /silencer
Proximal
Core
Start
Stop
Terminator
DNA
Transcription
Exon
Exon
Exon
Intron
Intron
Pre-mRNA
Post-transcriptional modification
Protein coding region
5'cap
Poly-A tail
Mature mRNA
Translation
Protein

Séquence d'un chromosome de levure

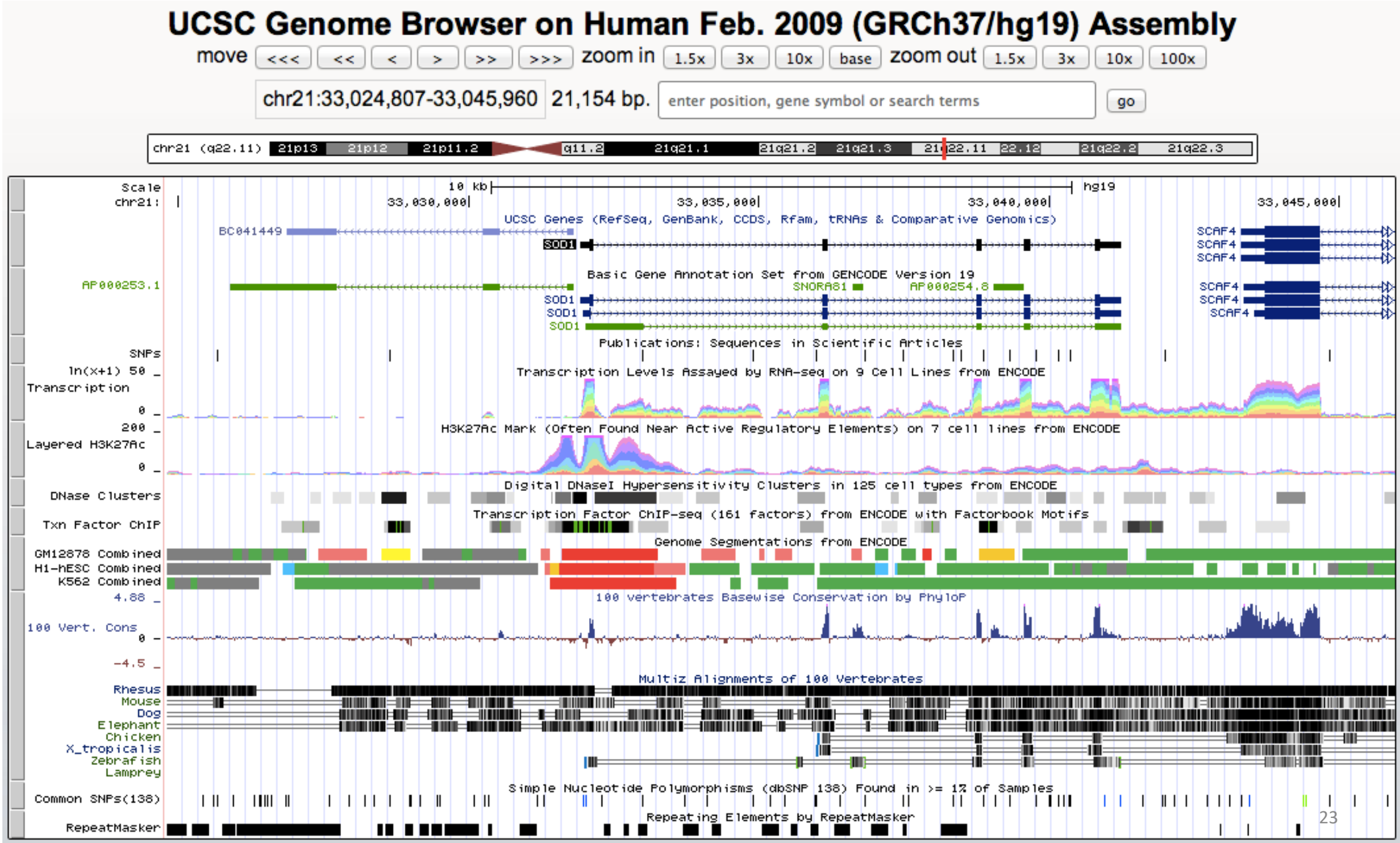


Séquence de fragments d'un chromosome de *C. elegans*

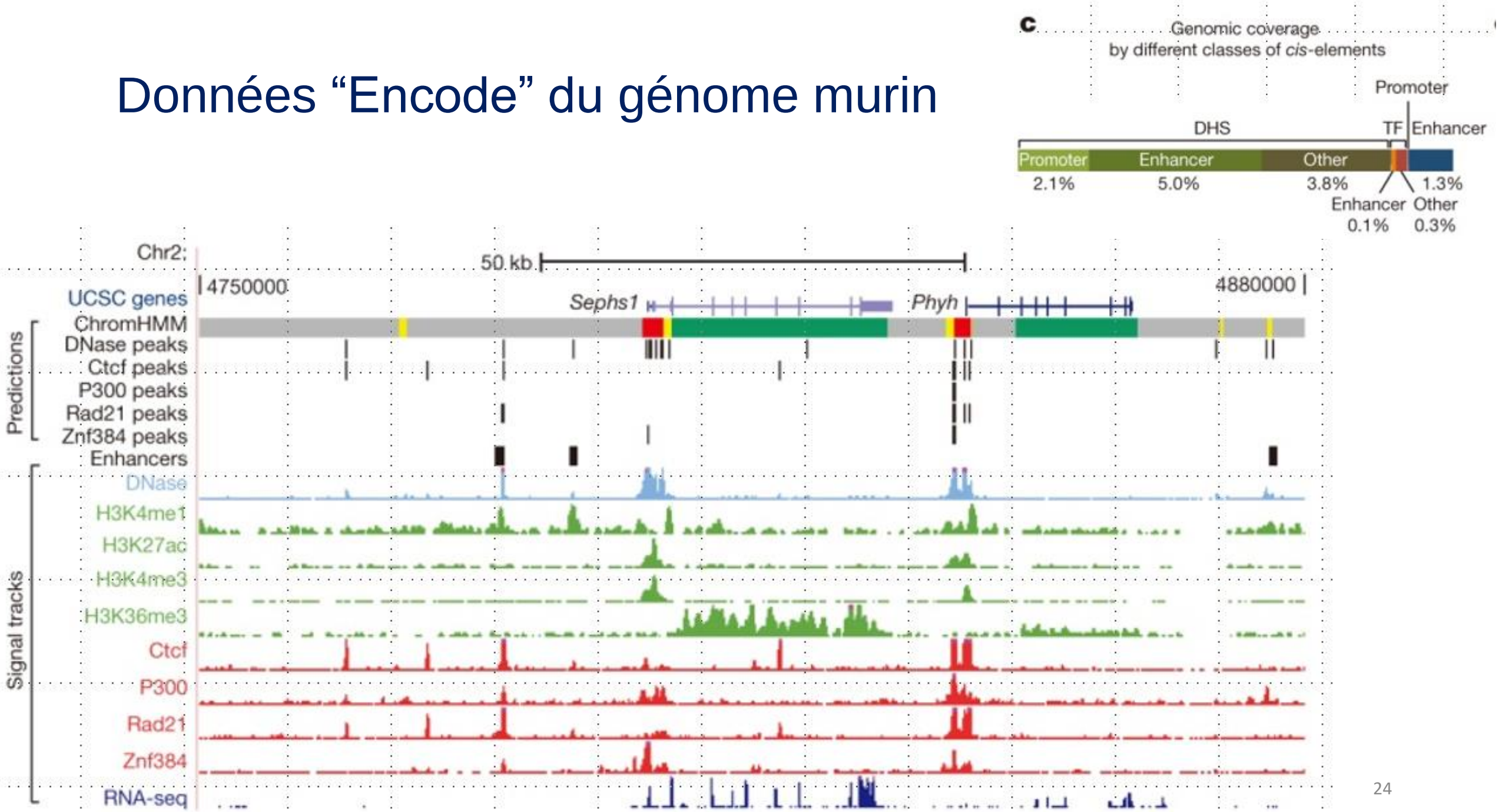


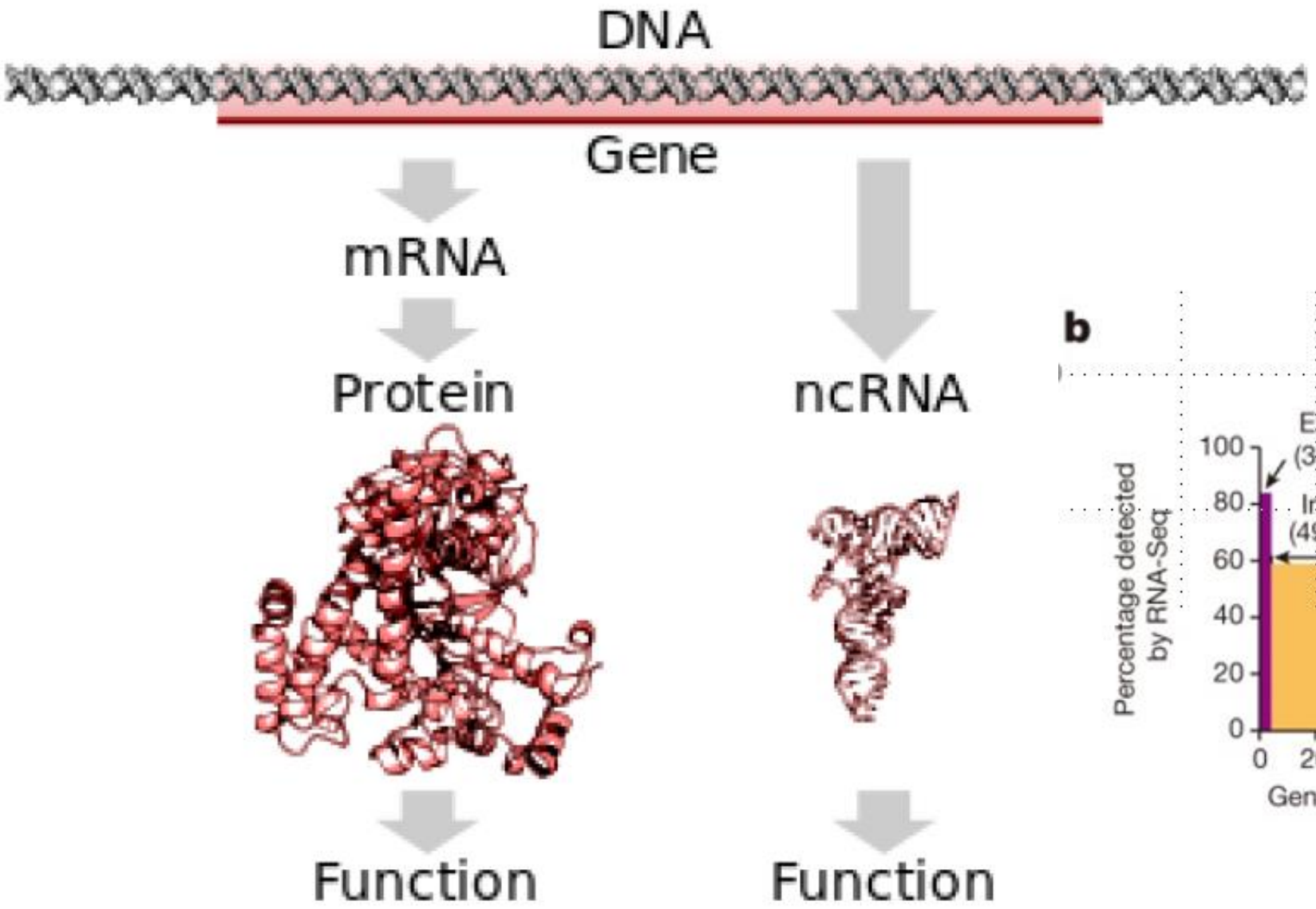


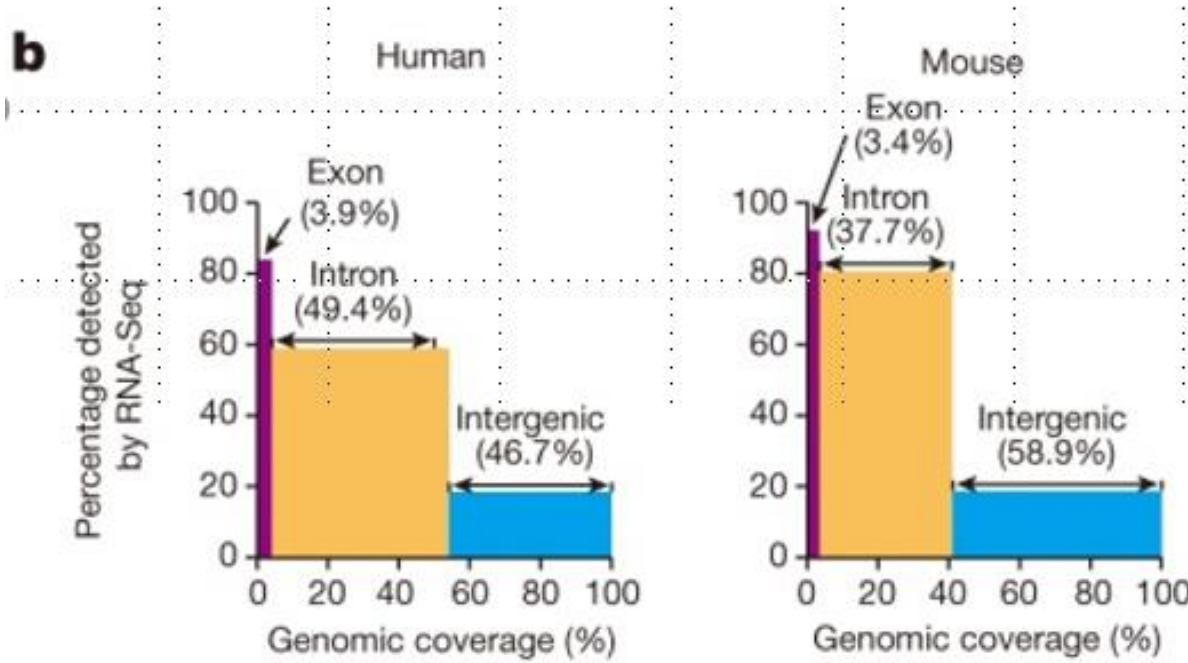
Sulston et al. Nature 1992


UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
chr21:33,024,807-33,045,960 21,154 bp.
enter position, gene symbol or search terms
go
chr21 (q22.11) 21p13 21p12 21p11.2 q11.2 21q21.1 21q21.2 21q21.3 21q22.11 22.12 21q22.2 21q22.3
Scale
chr21:
10 kb
hg19
33,030,000
33,035,000
33,040,000
33,045,000
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
BC041449
SOD1
SCAF4
Basic Gene Annotation Set from GENCODE Version 19
AP000253.1
SNORA81
AP000254.8
SNPs
Publications: Sequences in Scientific Articles
ln(x+1) 50
Transcription
0
Transcription Levels Assayed by RNA-seq on 9 Cell Lines from ENCODE
200
Layered H3K27Ac
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
DNase Clusters
Digital DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE
Txn Factor ChIP
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
GM12878 Combined
H1-hESC Combined
K562 Combined
Genome Segmentations from ENCODE
4.88
100 Vert. Cons
100 vertebrates Basewise Conservation by PhyloP
-4.5
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
Common SNPs(138)
Simple Nucleotide Polymorphisms (dbSNP 138) Found in >= 1% of Samples
RepeatMasker
Repeating Elements by RepeatMasker

Données “Encode” du génome murin







Protein coding genes are transcribed to an mRNA intermediate, then translated to a functional protein. RNA-coding genes are transcribed to a functional non-coding RNA

Le monde hétéroclite des ARN non codants (ncRNA)

ARN classés d'après leur taille, localisation, fonction

- long non-coding RNA (lncRNAs) (>200 nucléotides)

- micro RNA (miRNAs)

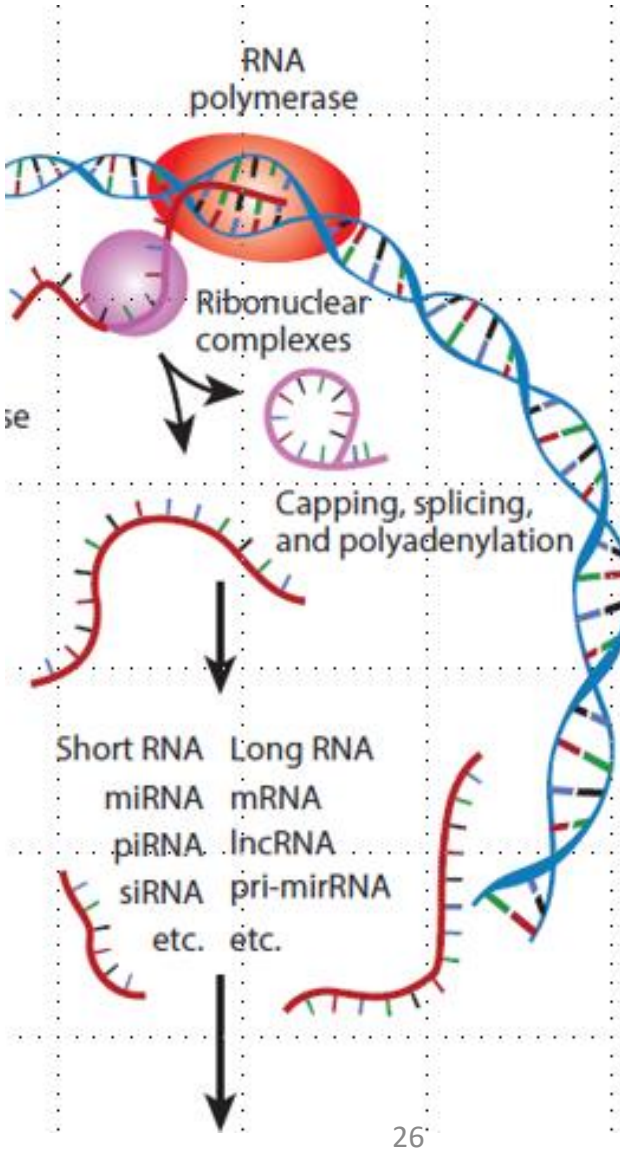
- small interfering RNAs (siRNA)

- PIWI-interacting RNAs (piRNAs)

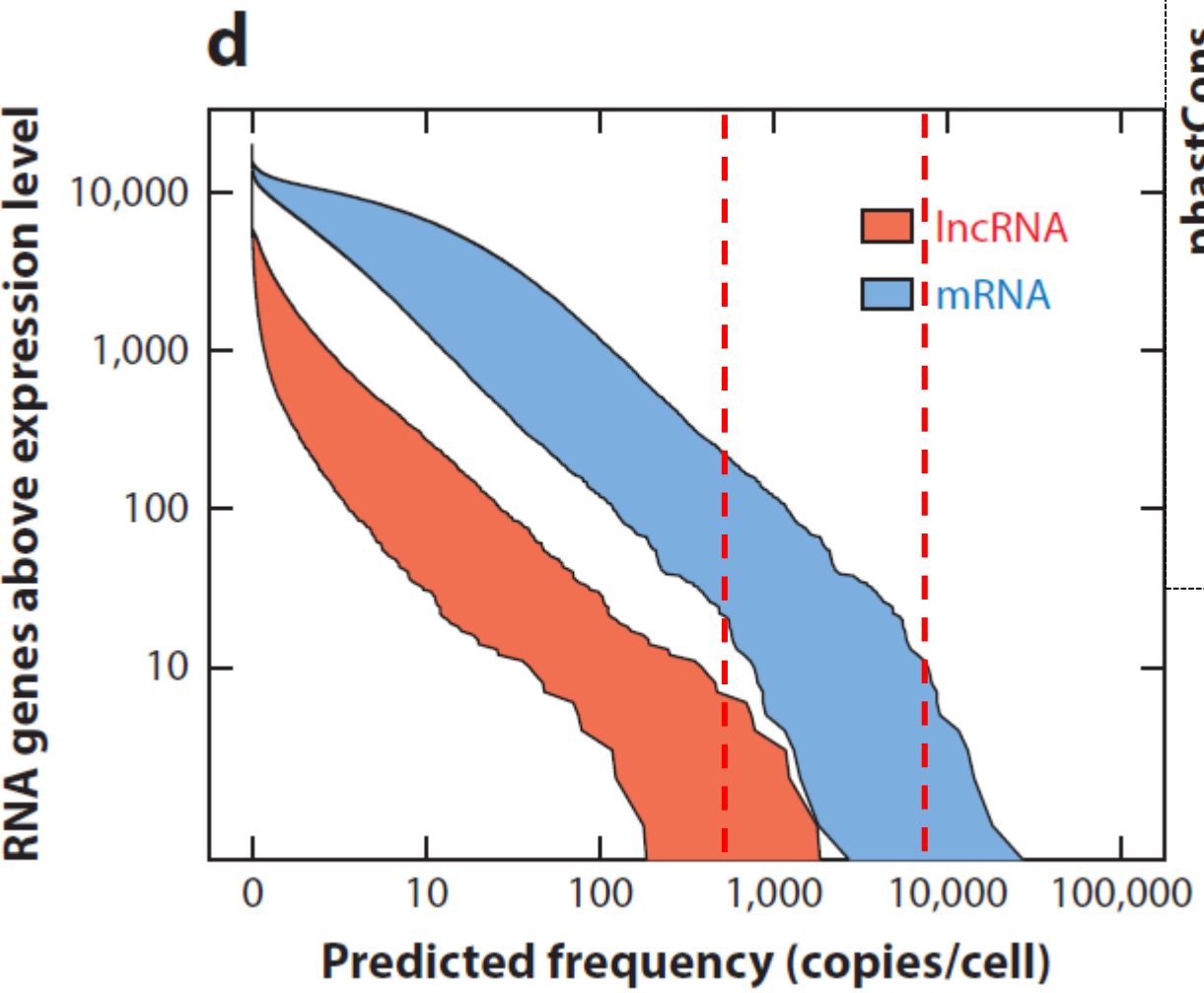
régulation de l'expression

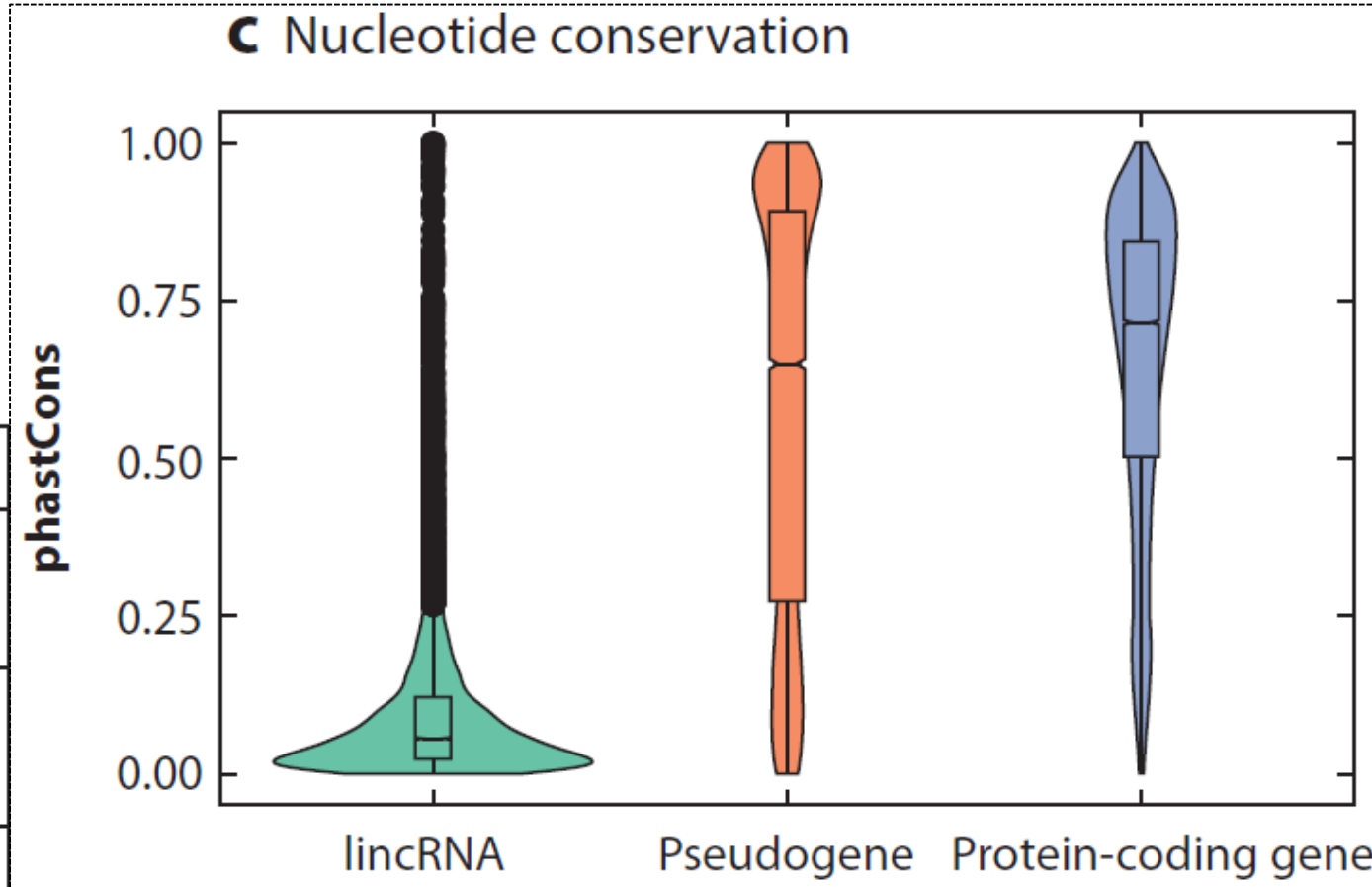
- small nucleolar RNAs (snoRNAs)

modifications d'ARN



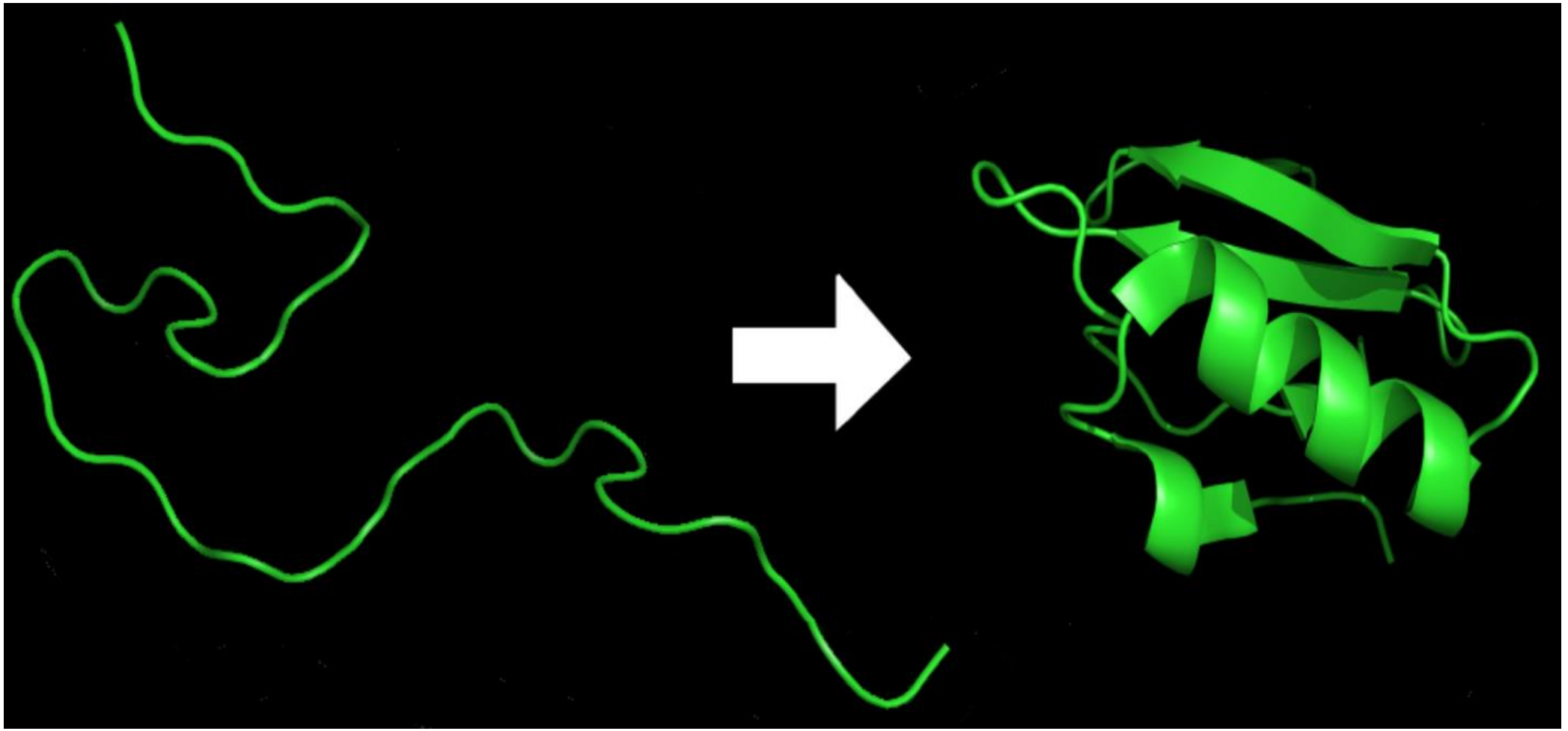
ARN non codants (ncRNA) et ARN codants





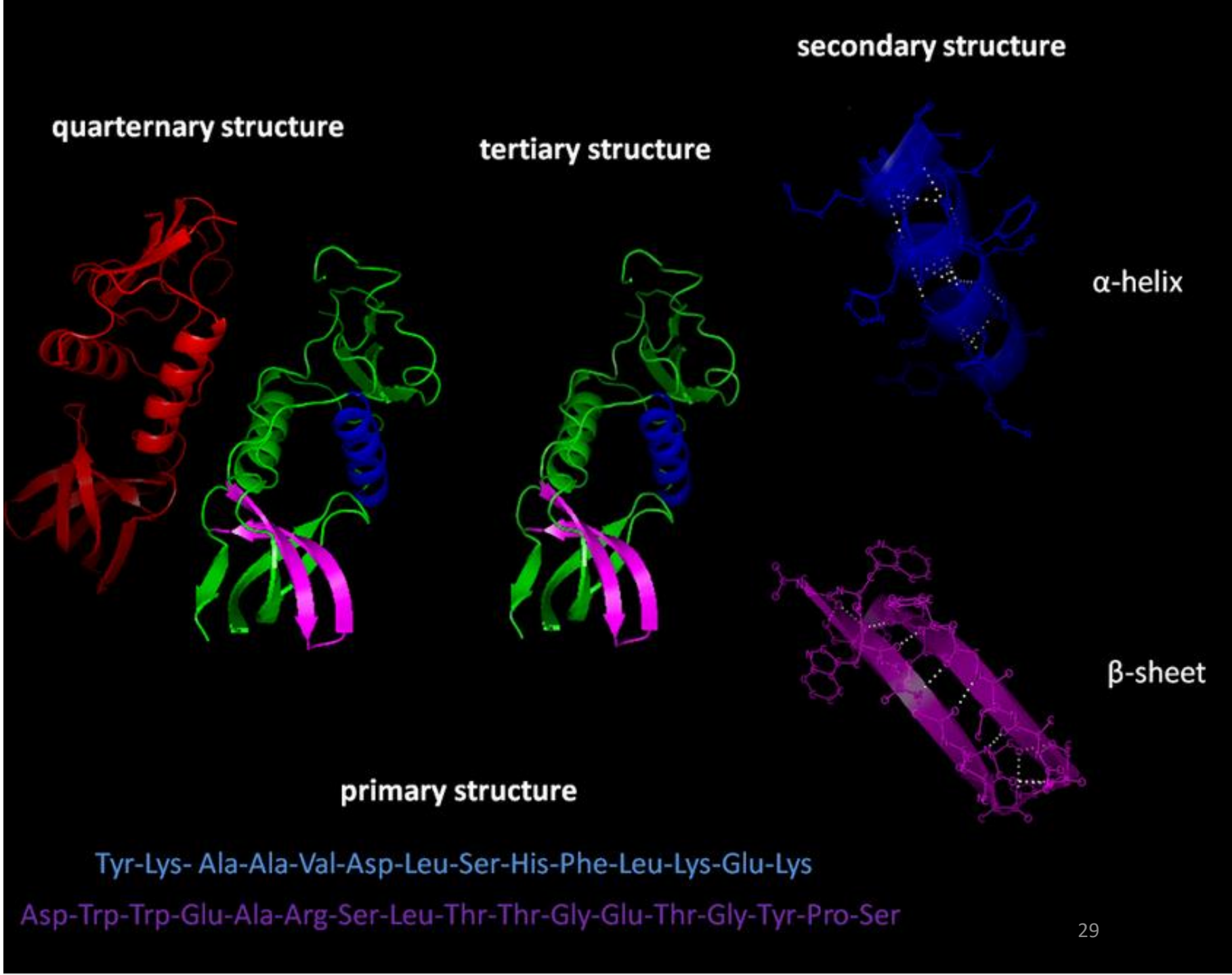
- La majeure partie des ncRNA sont des introns excisés en cours de maturation des mRNA
- Quelques lncRNA régulateurs bien caractérisés
- Peu de génétique associée aux lncRNA
- Bruit de fond de la machinerie transcriptionnelle ?

Repliement des protéines



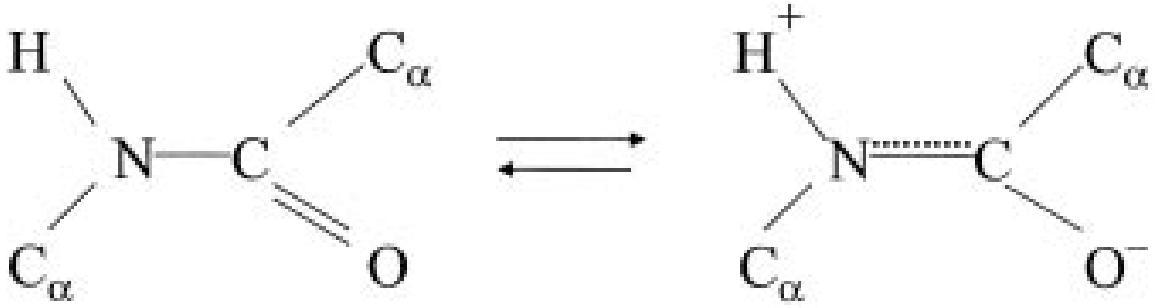
Les protéines sont fonctionnelles lorsqu'elles ont adopté leur conformation 3D dictée par leur structure primaire (leur séquence d'aminoacides)

Repliement des protéines

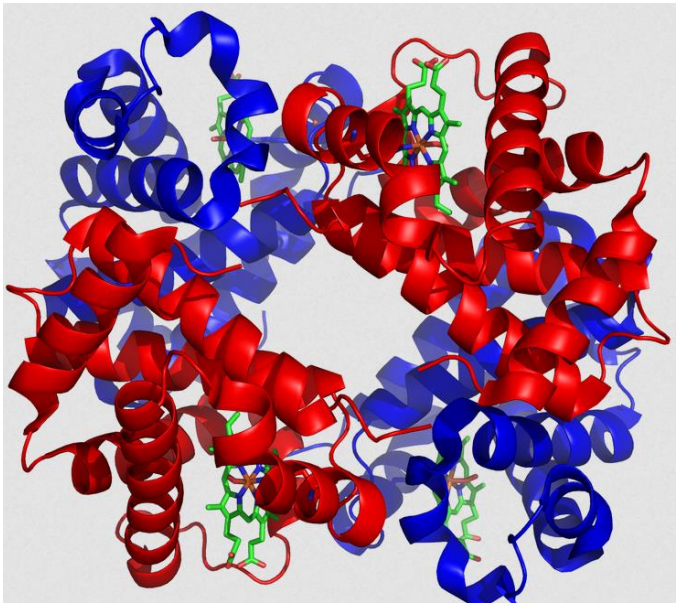


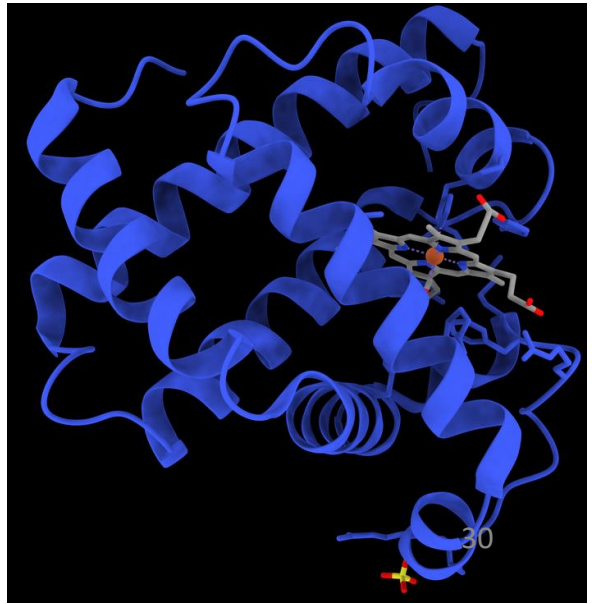
Strucrure tridimensionnelle des protéines

1951 - des association locales d'aminoacides sont proposées : les hélices alpha et les feuillets beta qui vont se révéler les
2 motifs structuraux de loin
 les plus fréquents dans les protéines
et leur servant d'ossature



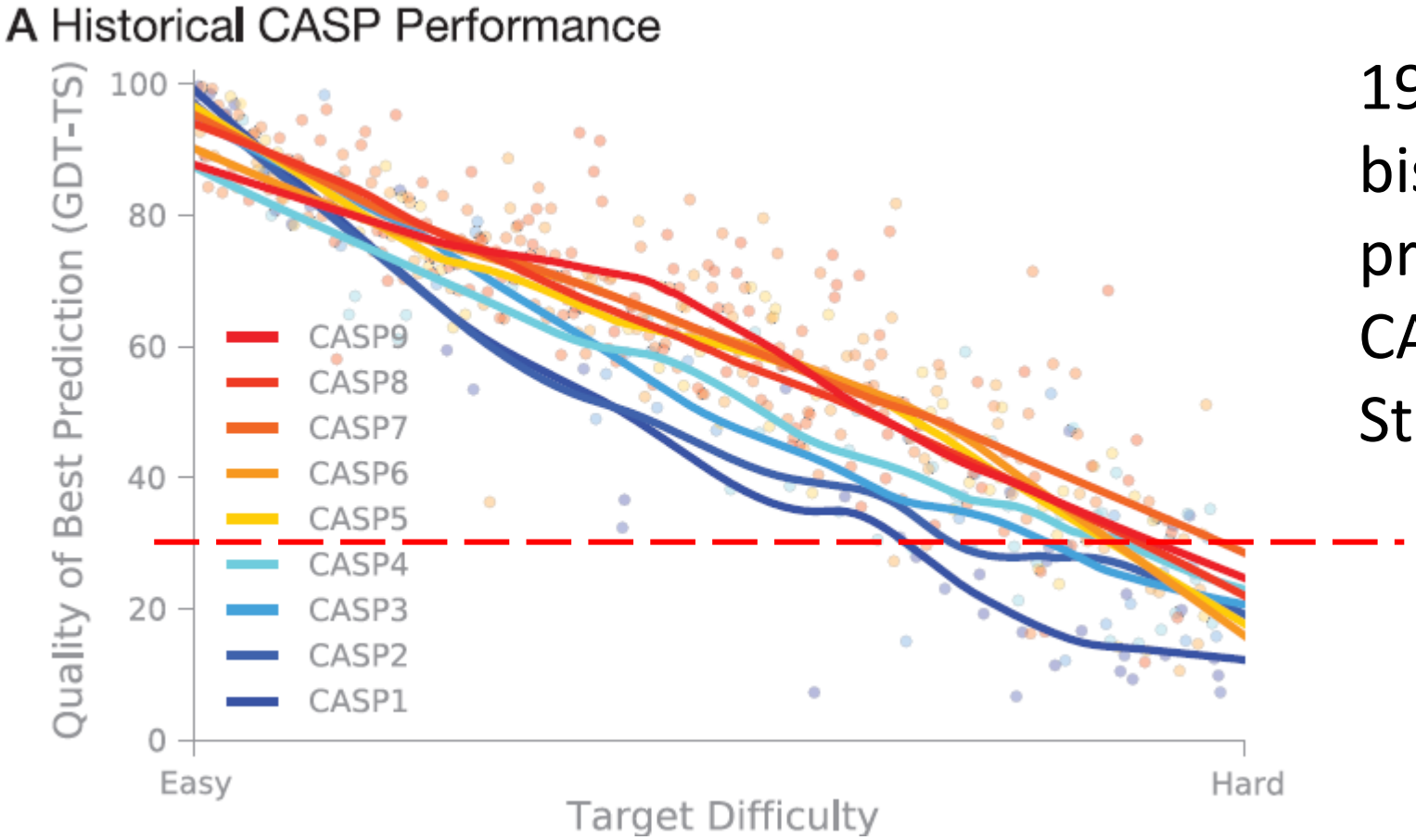
1959 - premières structures 3D de protéines (hémoglobine et myoglobine)
À partir des images de diffraction des rayons X sur des cristaux des protéines





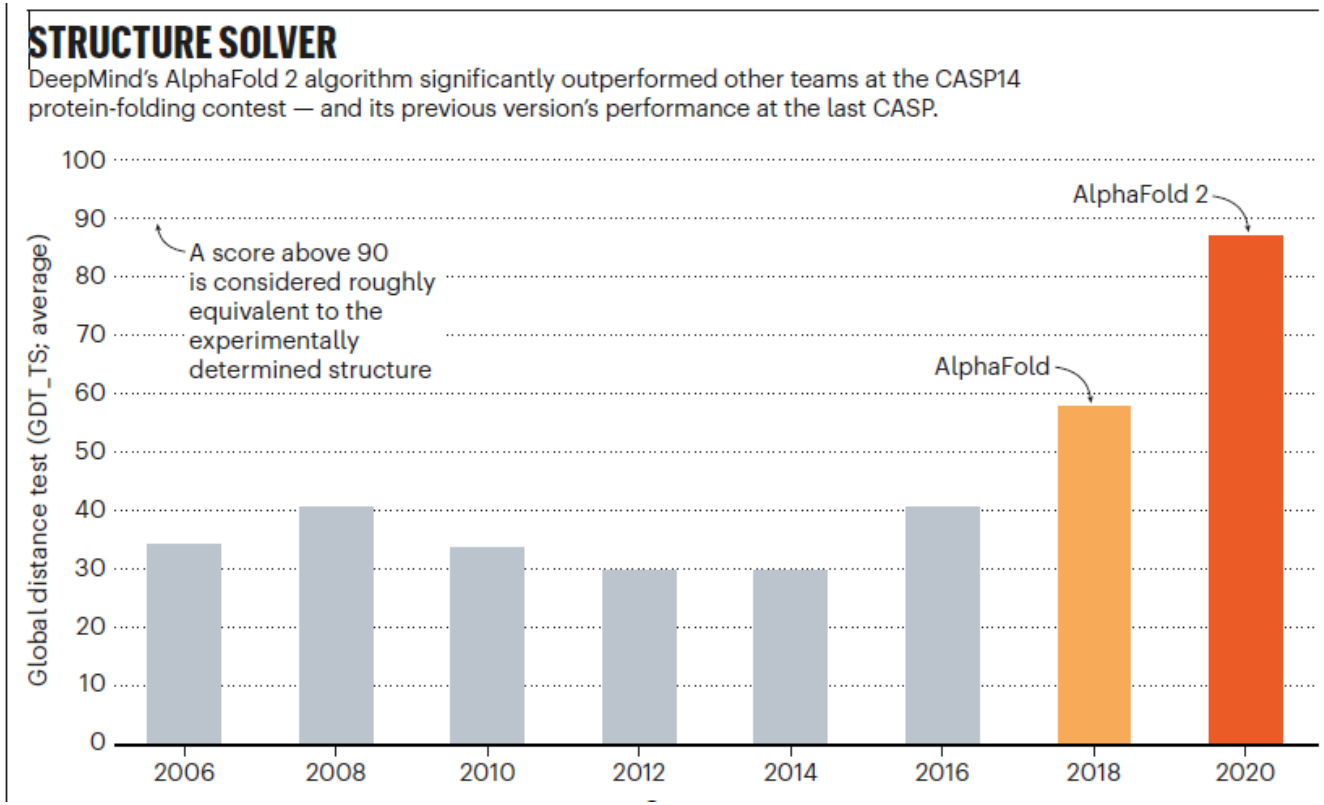
Prédire le repliement des protéines

60 ans de tentatives de prédiction de la structure 3D à partir de la séquence d'aminoacides avec notamment l'utilisation de plus en plus systématique de méthodes informatiques



1994 création d'un concours bisannuel de programmes de prédiction de structure 3D CASP (Critical Assessment of Structure Prediction)

Prédire la structure 3D des protéines

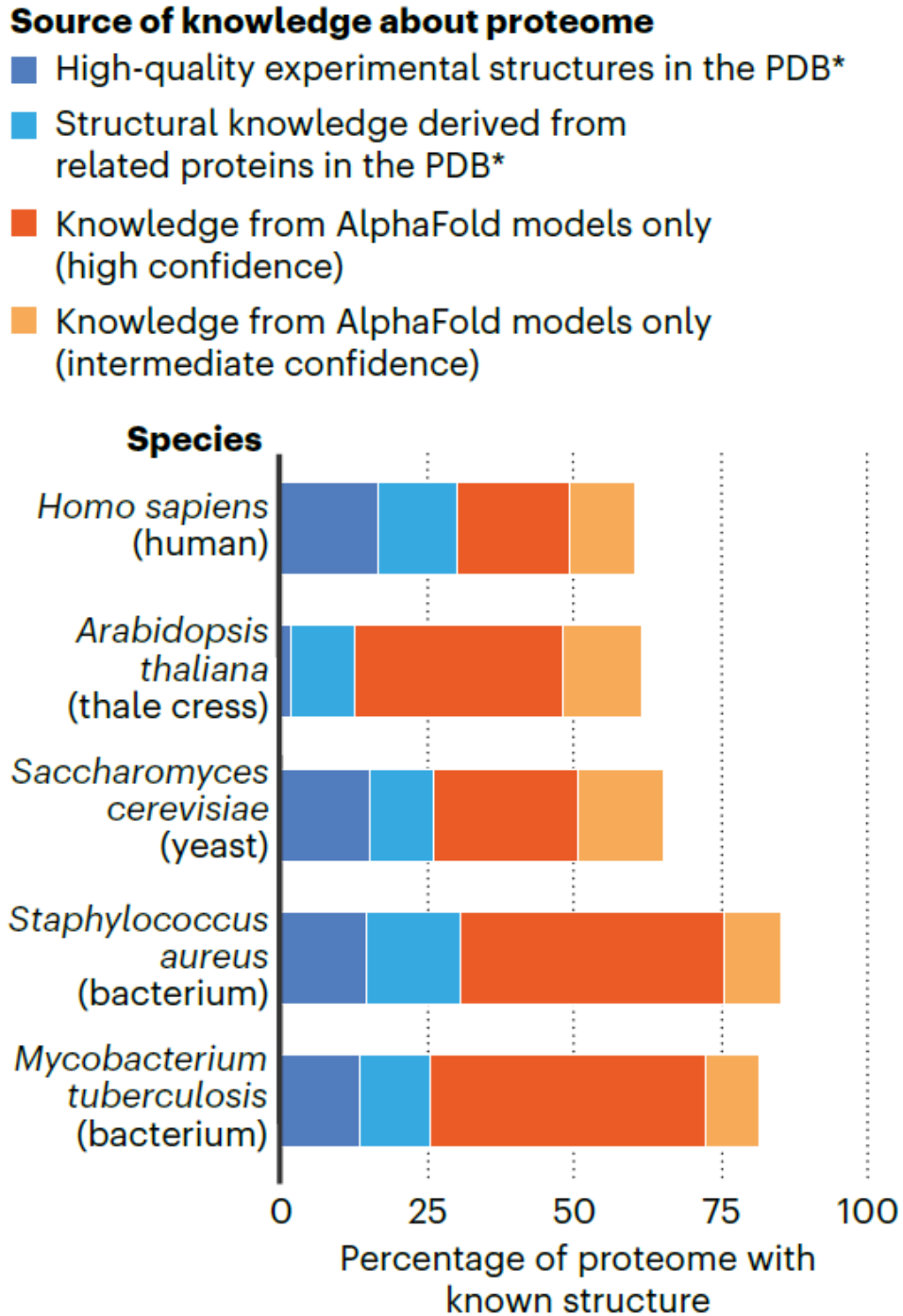


AlphaFold2 : algorithme s'appuyant sur l'intelligence artificielle : réseaux neuronaux, apprentissage automatique

AlphaFold2 a bénéficié d'un corpus d'apprentissage considérable : ensemble des structures 3D issues des approches expérimentales de cristallographie et déposées dans PDB (Protein Data Base)

AlphaFold2 ne fait pas (encore ?) tout

- Effet des mutations sur la structure 3D
- Dynamique des changements de conformation
- Fonction biologique


Source of knowledge about proteome
High-quality experimental structures in the PDB*
Structural knowledge derived from related proteins in the PDB*
Knowledge from AlphaFold models only (high confidence)
Knowledge from AlphaFold models only (intermediate confidence)
Species
Homo sapiens (human)
Arabidopsis thaliana (thale cress)
Saccharomyces cerevisiae (yeast)
Staphylococcus aureus (bacterium)
Mycobacterium tuberculosis (bacterium)
0
25
50
75
100
Percentage of proteome with known structure

80 ans de déchiffrement

Depuis l'identification de l'ADN comme support de l'information génétique et des protéines comme principales molécules effectrices des systèmes biologiques au début des années 40, des physiciens, chimistes, biochimistes, biologistes moléculaires, informaticiens etc. ont

- Établi la correspondance entre les codes des nucléotides et des aminoacides
- Déterminé la structure chimique détaillée (séquence) d'ADN, ARN et protéines
- Réalisé l'inventaire des protéines de dizaines de milliers d'organismes
- Établi expérimentalement la structure 3D de nombreuses protéines issus de ces inventaires
- Déchiffré de nombreux sites sur l'ADN contrôlant de l'expression des gènes
- Identifié de nombreuses molécules régulatrices de cette expression (protéines, ncRNA)
- Prédit la structure 3D de protéines à l'aide de méthodes de l'intelligence artificielle

Dynamique du génome et de ses produits

Mécanismes naturels ayant lieu *in vivo*

- Variants de séquence (mutations)
- Transfert horizontal
- Réparation/édition (lésions)
- Modifications post transcription
- Modifications post traduction

Vers des néo-gènes et néo-génomes ?

Démarches expérimentales

- Modifications du génome et des gènes :
  Génie génétique, ADN recombinant/CRISPR-cas9

- Modifications du code génétique :
  aminoacides et nucléotides non canoniques, xénobiologie

- Sachant lire l'ADN, sait-on l'écrire ?
  génomes synthétiques,
  protéines artificielles