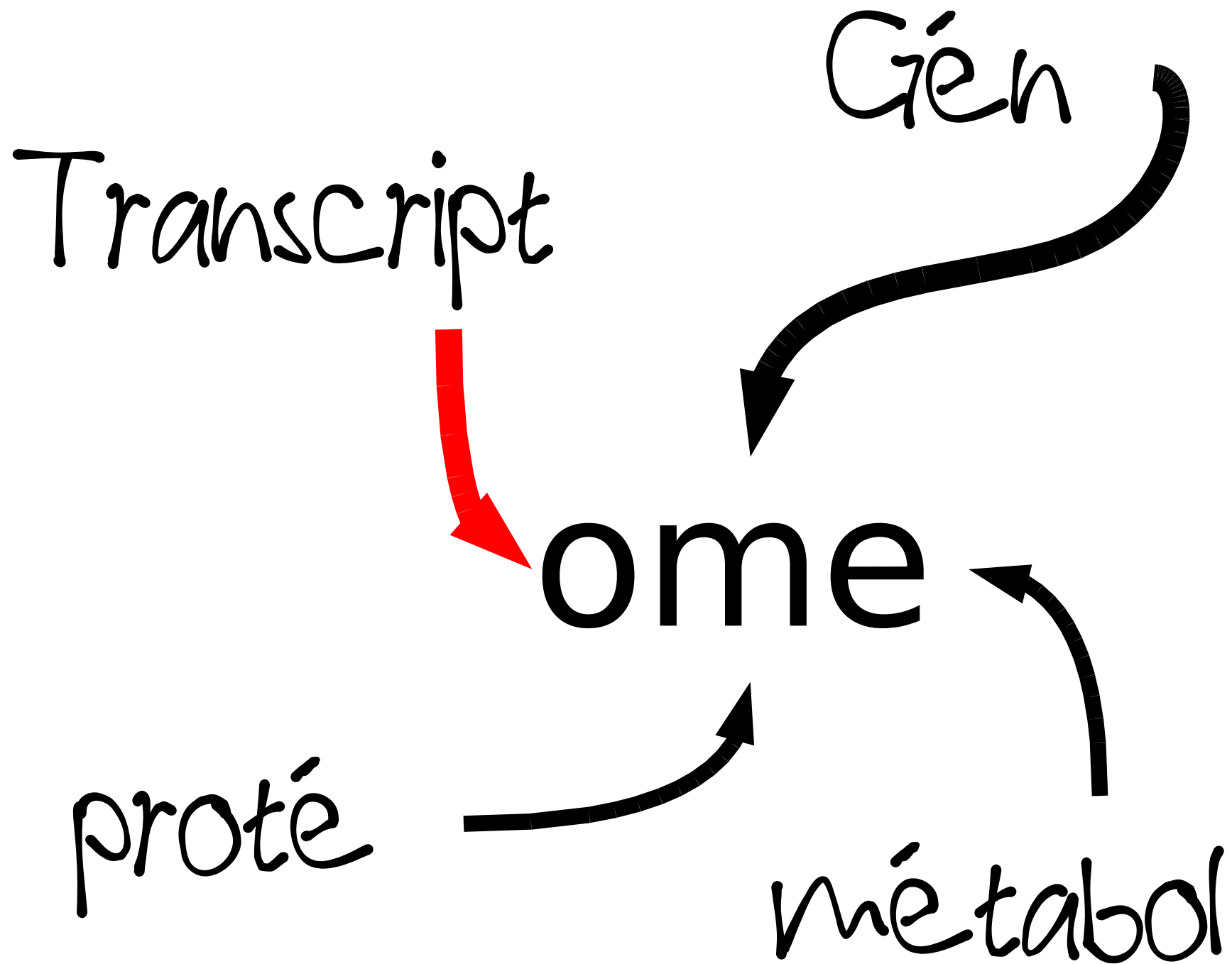




Le monde des "omes"

génom

ADN

transcriptome

ARN

protéome

Protéines

métabolome

Oxydation
des acides gras

glycolyse

Cycle de Krebs

Enzyme

ATP

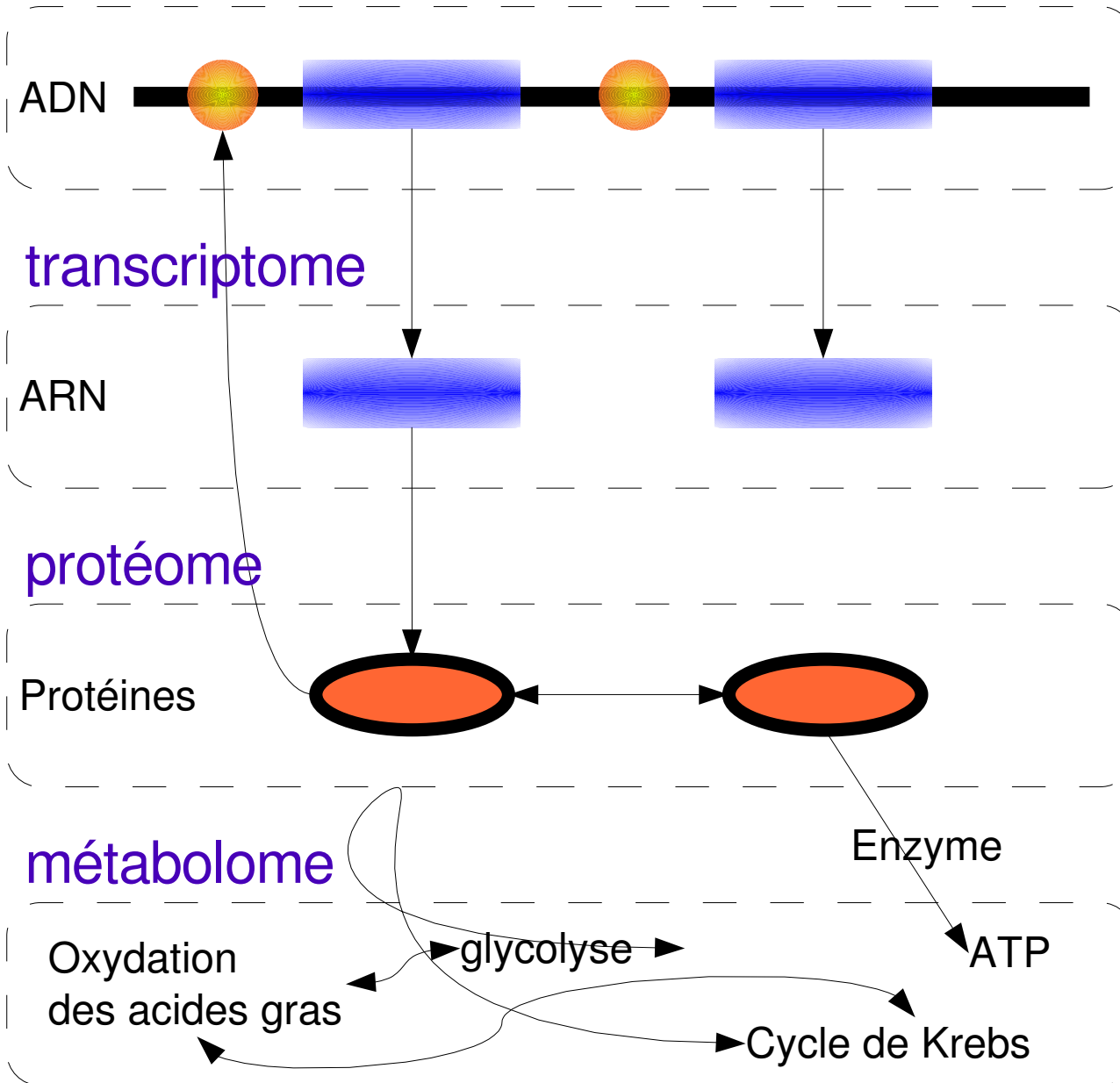
Séquençage de l'ADN

Séquence
Structure ARN
Evolution

Expression

Structure des protéines
Evolution

Métabolisme





↑
Système écologique

Population

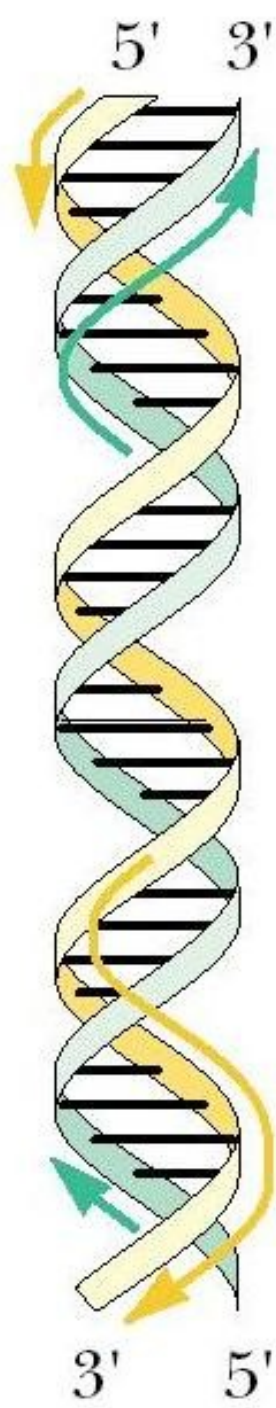
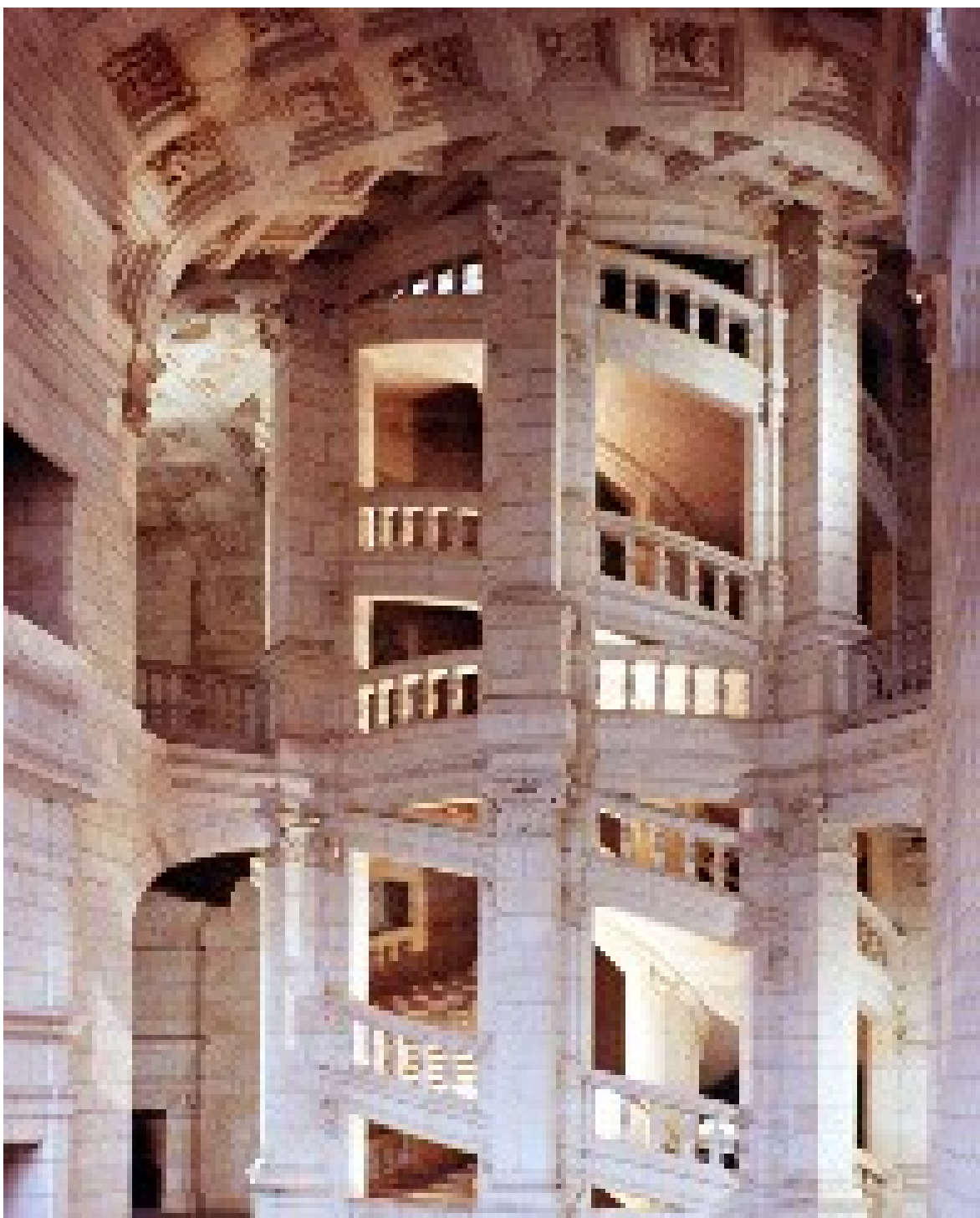
Organismes

Cellule

Molécules — [Assemblages
Polymères
Monomères



Acides nucléiques
Protéines



acides nucléiques

Outils bioinformatiques

- ★ identifier les gènes
- ★ prédire leur fonction
- ★ appréhender les signaux de régulation, les relations et interactions entre objets biologiques

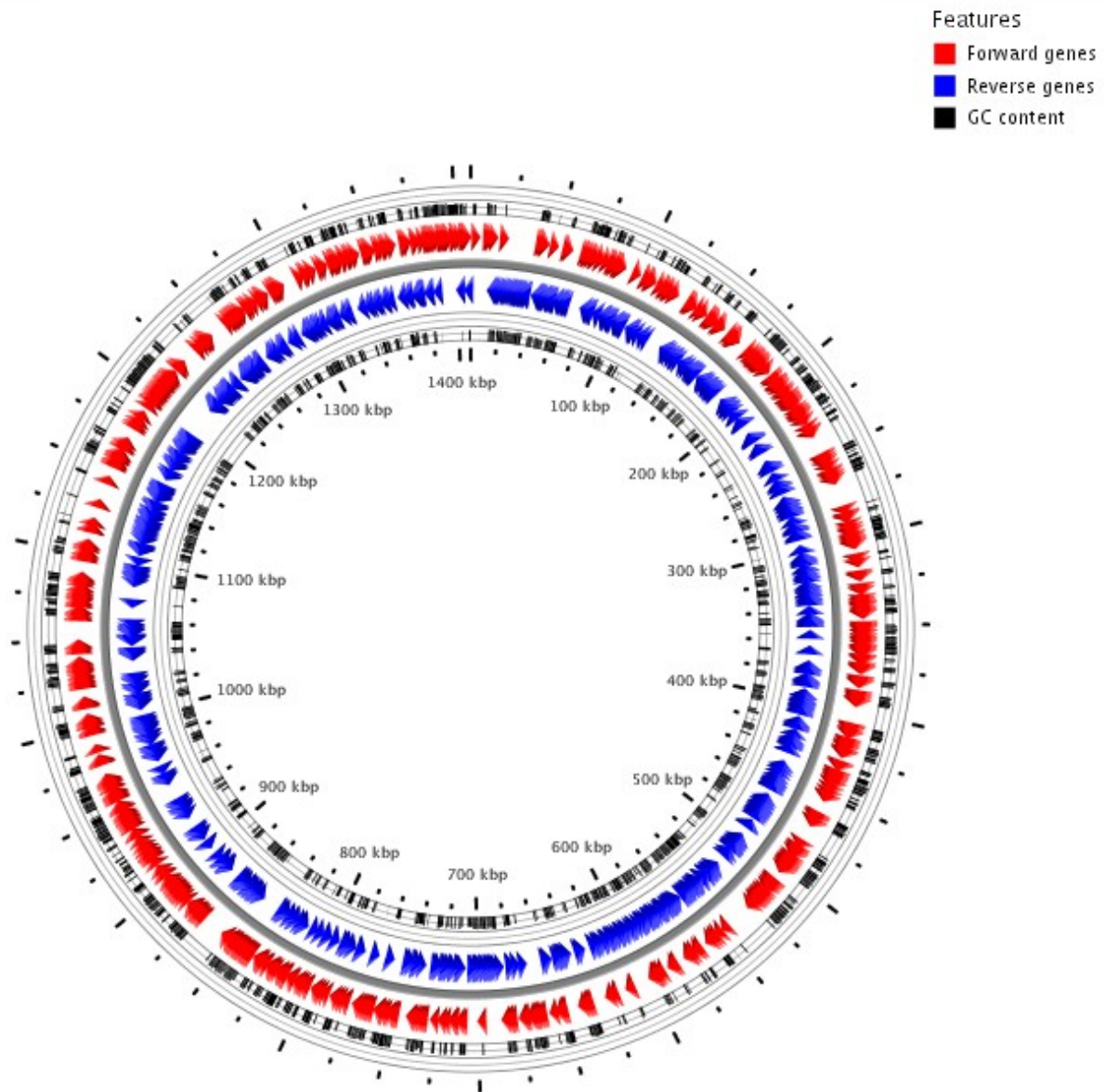
Structure du génome des procaryotes

- ★ Unique chromosome circulaire

- ★ Plasmide (structures chromosomiques d'ADN circulaires libres ds le cytoplasme à réplication indépendante).

- ★ Séquence de chaque gène continue: pas d'introns

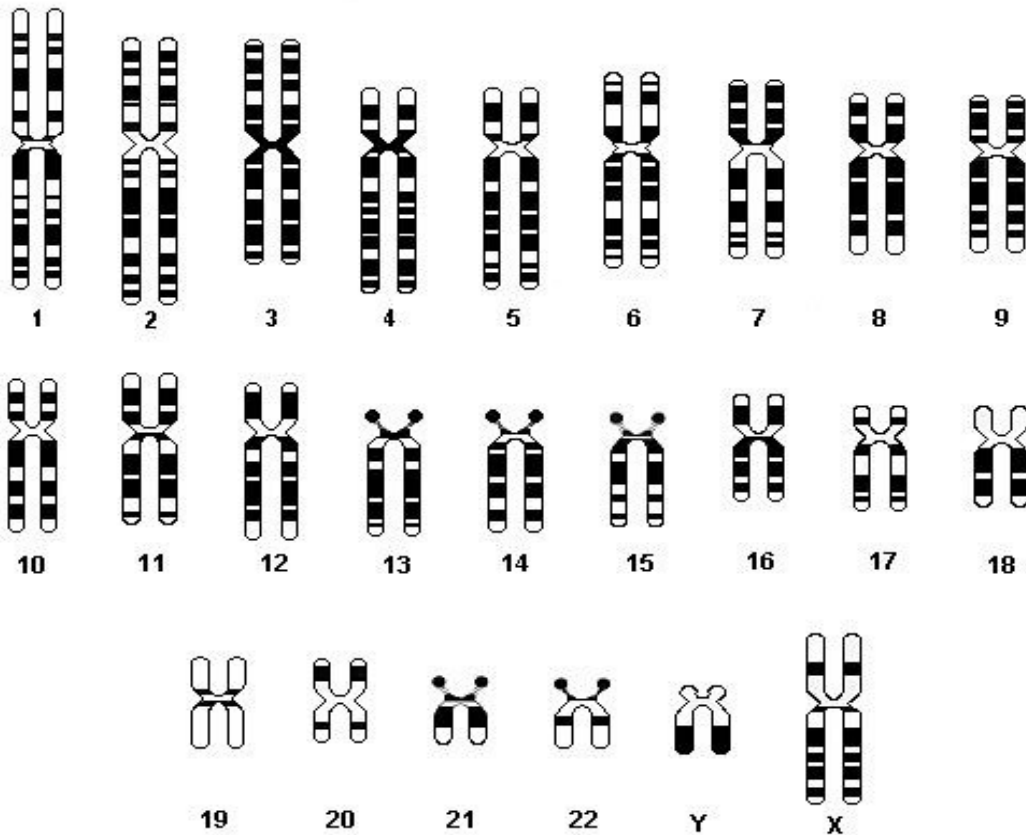
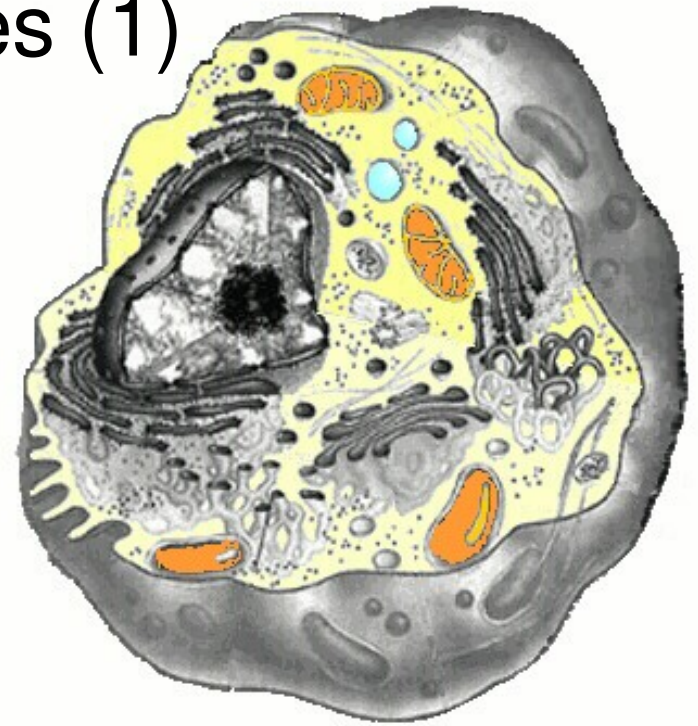
- ★ 85 à 90% de régions codantes



Structure du génome des eucaryotes (1)

- Noyau isolant le génome: génome nucléaire

- ★ Organites ayant leur propre génome (ADN circulaire)



- ★ Nombre de chromosomes constant pour une espèce

- ★ Un chromosome est constitué d'un ADN ininterrompu + diverses protéines

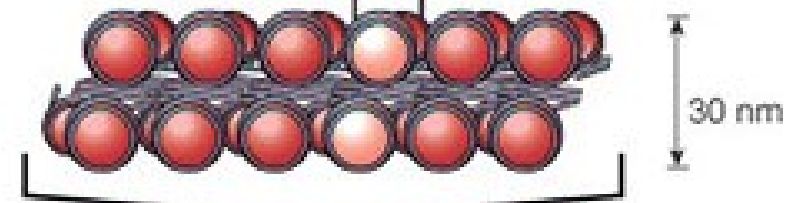
Short region of
DNA double helix



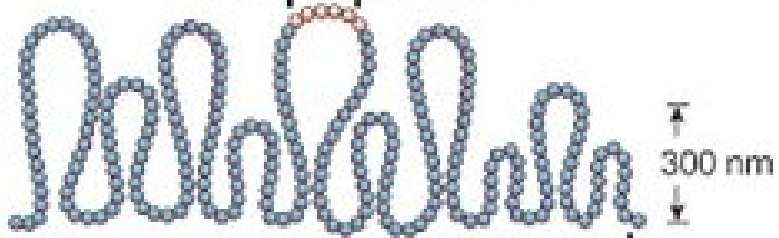
"Beads on a string"
form of chromatin



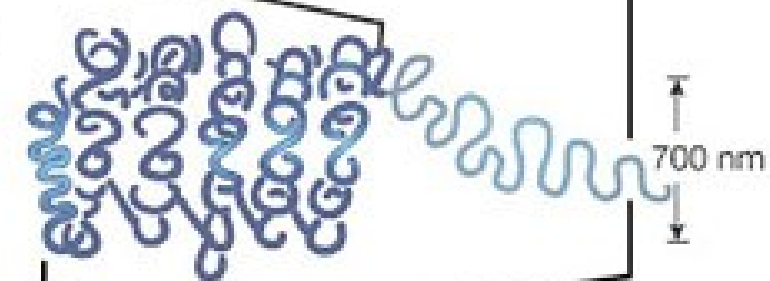
30-nm chromatin
fibre of packed
nucleosomes



Section of
chromosome in an
extended form



Condensed section
of chromosome



Entire mitotic
chromosome



Séquençage de génom



Historique du séquençage

- **1943-1953**: ADN support de l'information génétique
- **1977**: Techniques modernes de séquençage
- **1981**: Premières banques de données informatisées (GenBank, EMBL, PIR)
- **1990**: Début du projet génome humain
- **1995**: Premier génome complet d'un organisme cellulaire (*Haemophilus influenzae*)
- **2005**: 245 génomes complets annotés disponibles (eucaryotes et procaryotes)

Historique du séquençage

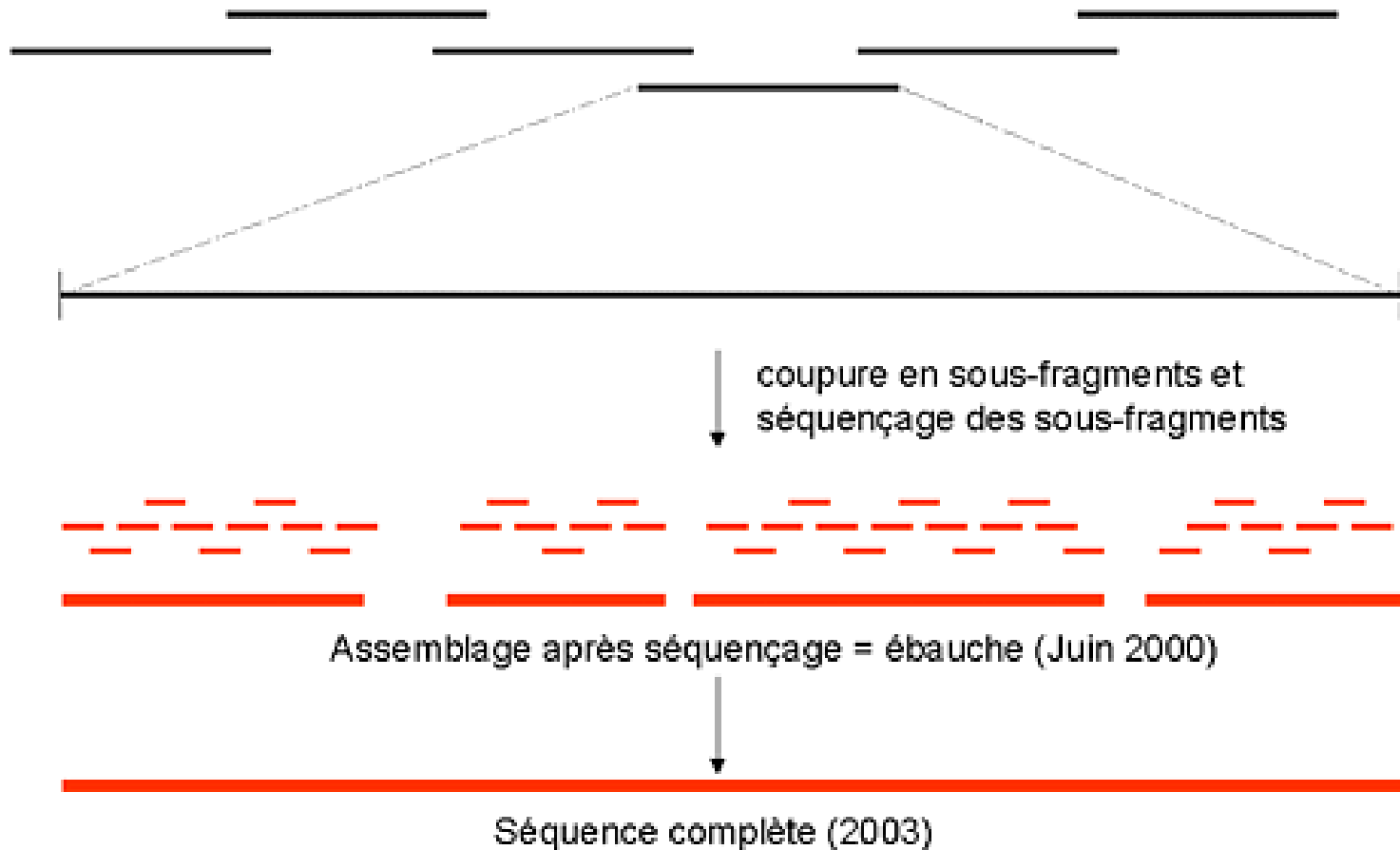
• <i>Amoeba dubia</i>	670,000,000,000
• <i>Homo sapiens</i>	3,400,000,000
• <i>Drosophila melanogaster</i>	180,000,000
• <i>Arabidopsis thaliana</i>	100,000,000
• <i>Caenorhabditis elegans</i>	100,000,000
• <i>Saccharomyces cerevisiae</i>	12,067,280
• <i>Escherichia coli</i>	4,639,221

Source : **DOGS - Database Of Genome Sizes**

http://www.cbs.dtu.dk/databases/DOGS/abbr_table.bysize.txt

Projet de séquençage du programme public international

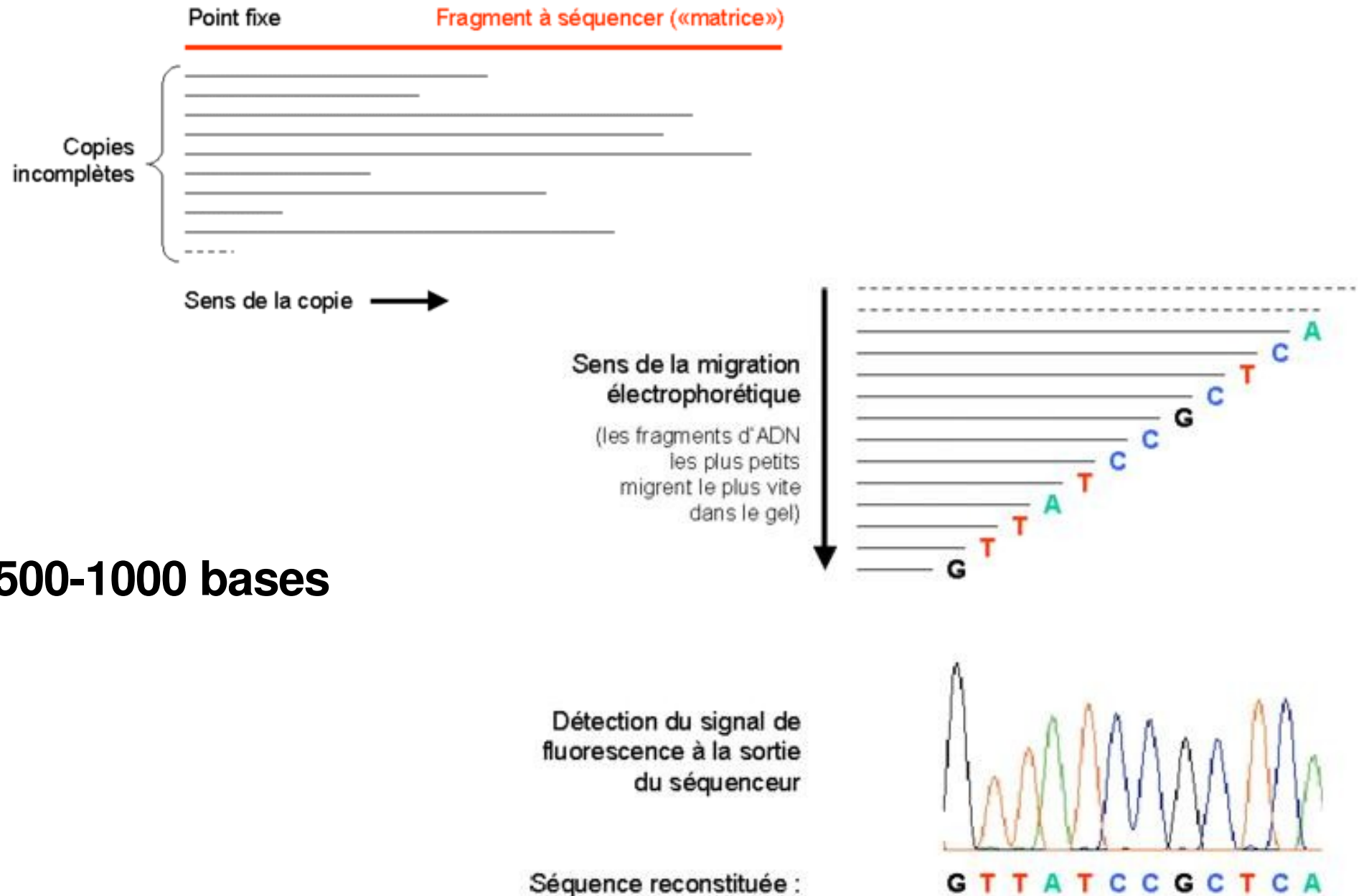
Grands fragments d'ADN ordonnés et chevauchants provenant d'une carte préalablement établie



Cartographie génétique

- **Automatisation du séquençage:**
- **Méthode de Sanger**
- **Séquençage haut débit**
- **Assemblage des contigs**

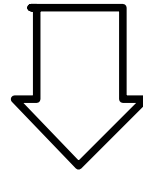
Méthode de Sanger



Séquençage de l'ADN: assemblage des contigs

Contig: Portion contiguë d'ADN résultant de l'assemblage d'un certain nombre de lectures (séquences) afin de reconstituer des enchaînements plus grands

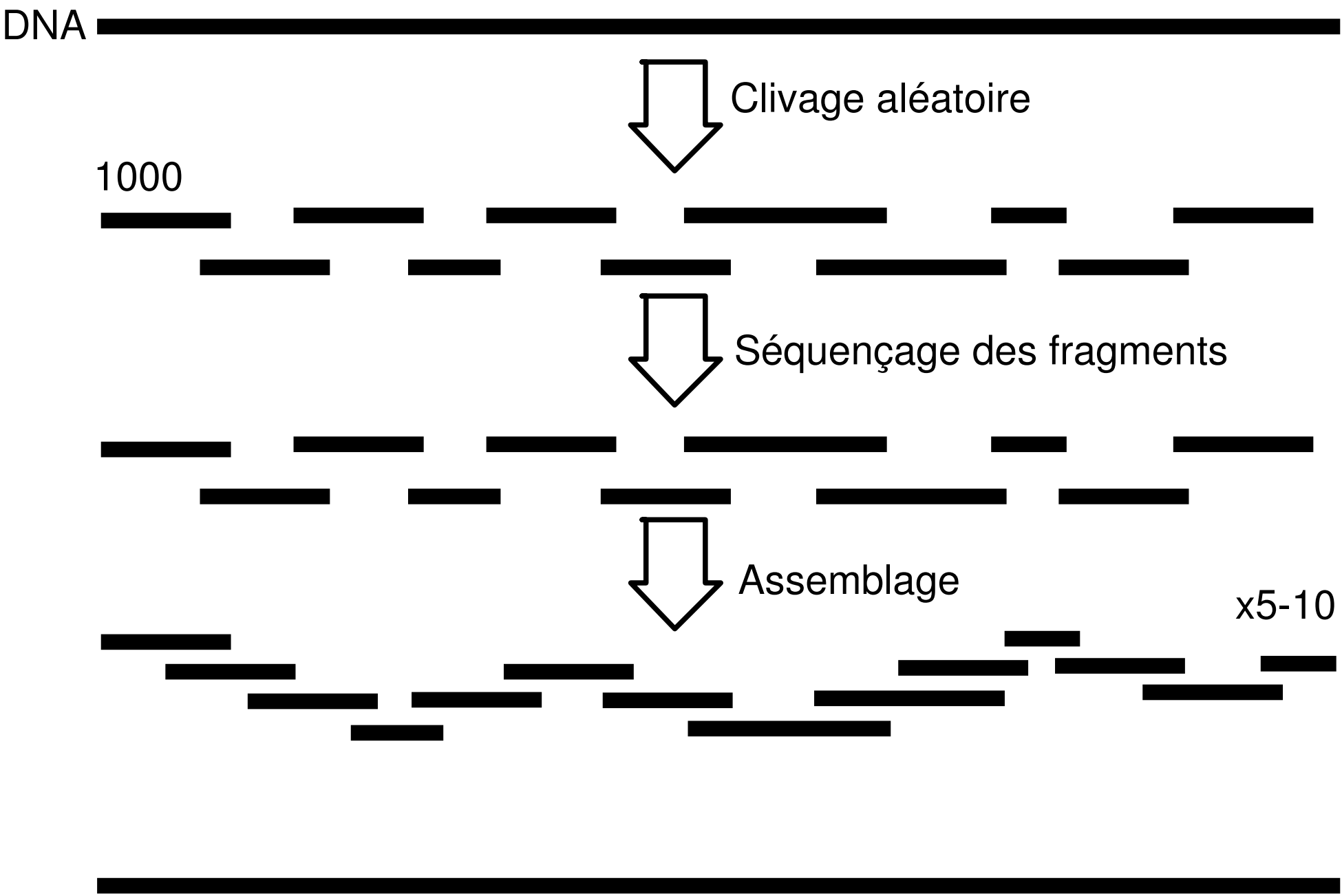
Fragments séquencés



Assemblage



Séquençage de l'ADN: assemblage de l'ADN



Séquençage de l'ADN: Exemple

ACCGT

CGTGC

TTAC

TACCGT

Séquençage de l'ADN: Exemple

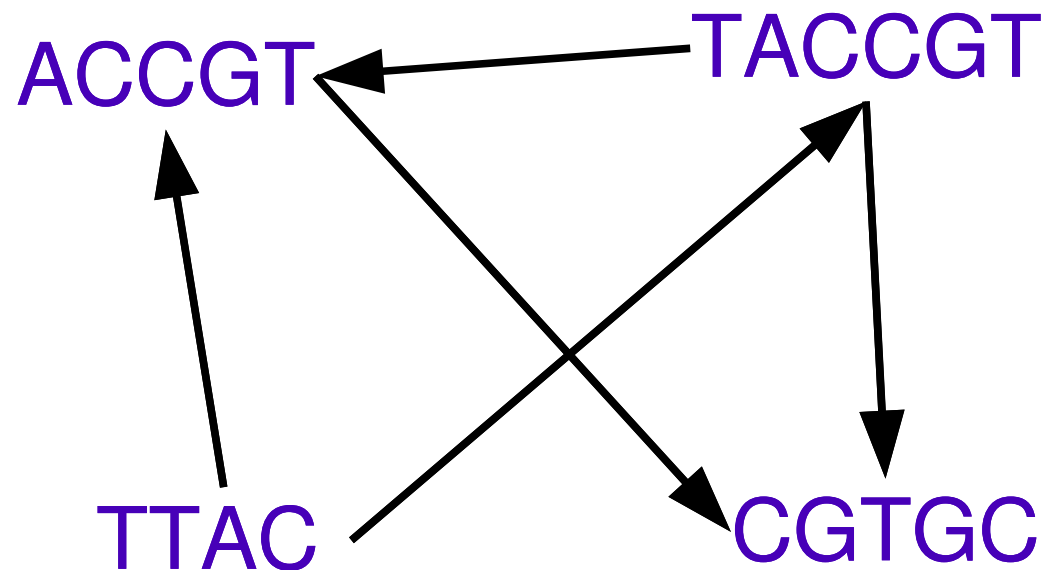
ACCGT
CGTGC
TTAC
TACCGT

ACCGT
CGTGC
TACCGT
TTAC

TTACCGTGC

Séquençage de l'ADN: Graphe hamiltonien

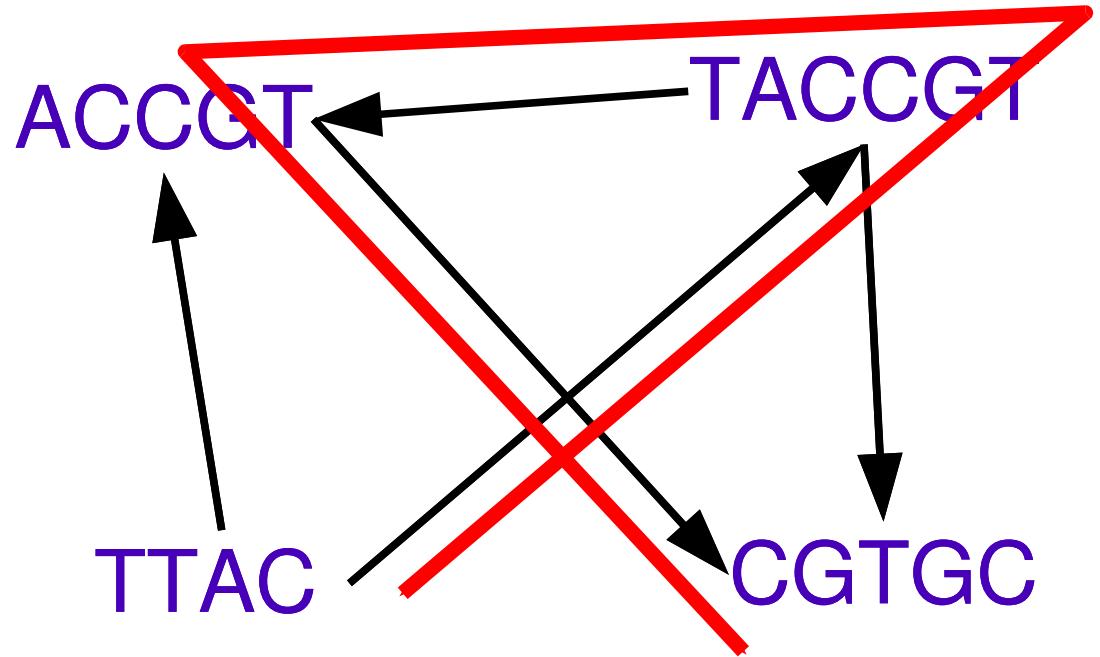
ACCGT
CGTGC
TTAC
TACCGT



Graphe hamiltonien: Traverser tous les noeuds du graphe

Séquençage de l'ADN: Graphe hamiltonien

ACCGT
CGTGC
TTAC
TACCGT



Problèmes réels:

- (i) erreurs de séquence
- (ii) orientation inconnue des fragments
- (iii) régions répétantes
- (iv) manque de chevauchements

TTACCGTGC

Cartographie génétique

- **Carte physique (ou chromosomique)**

couverture de chaque chromosome par un ensemble de grands fragments génomiques ordonnés sur la base de leurs chevauchements

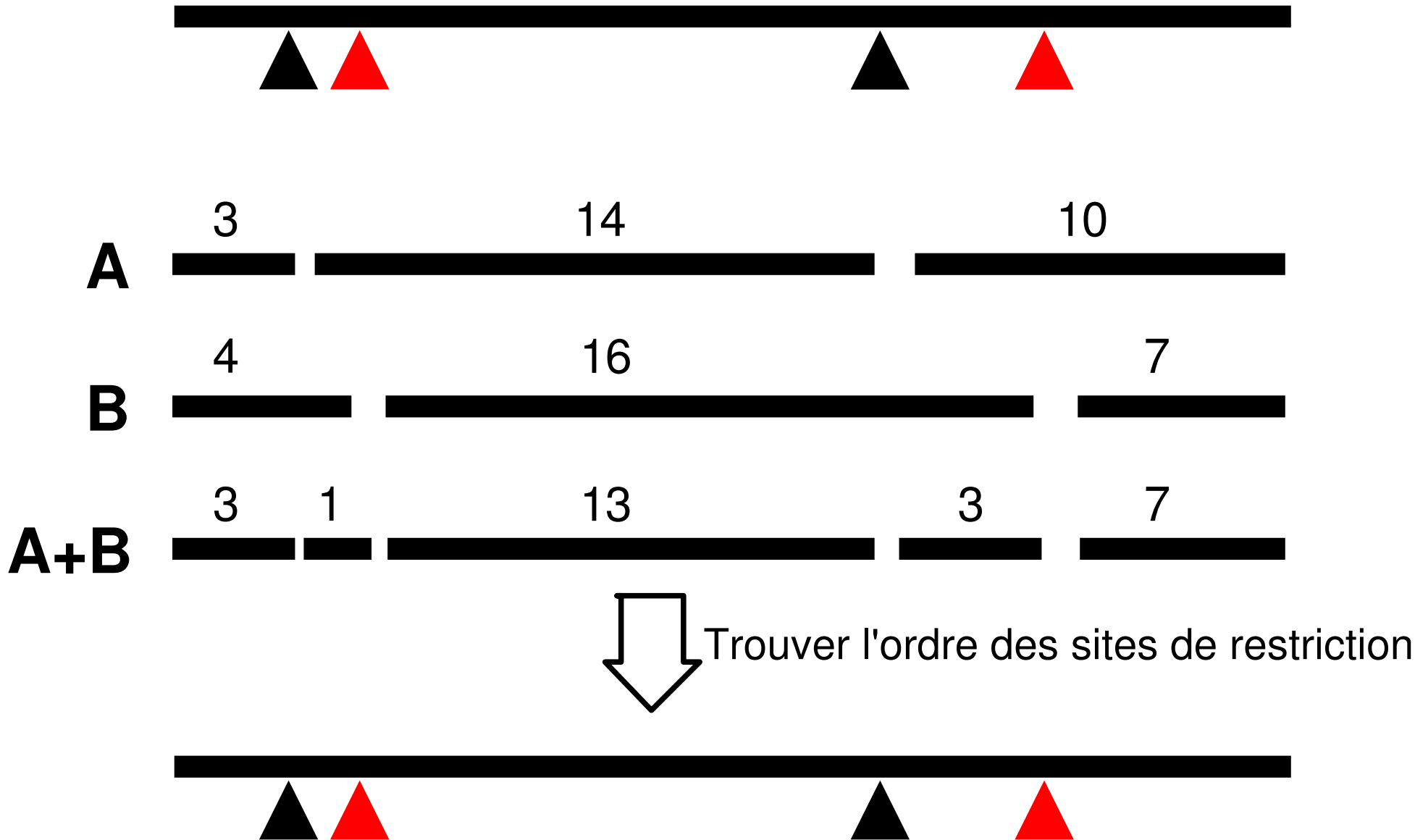
- **Carte de restriction**

arrangement linéaire des sites de reconnaissance des enzymes de restriction le long des chromosomes

- **Carte de liaison**

ensemble de marqueurs dont on détermine les positions relatives le long des chromosomes

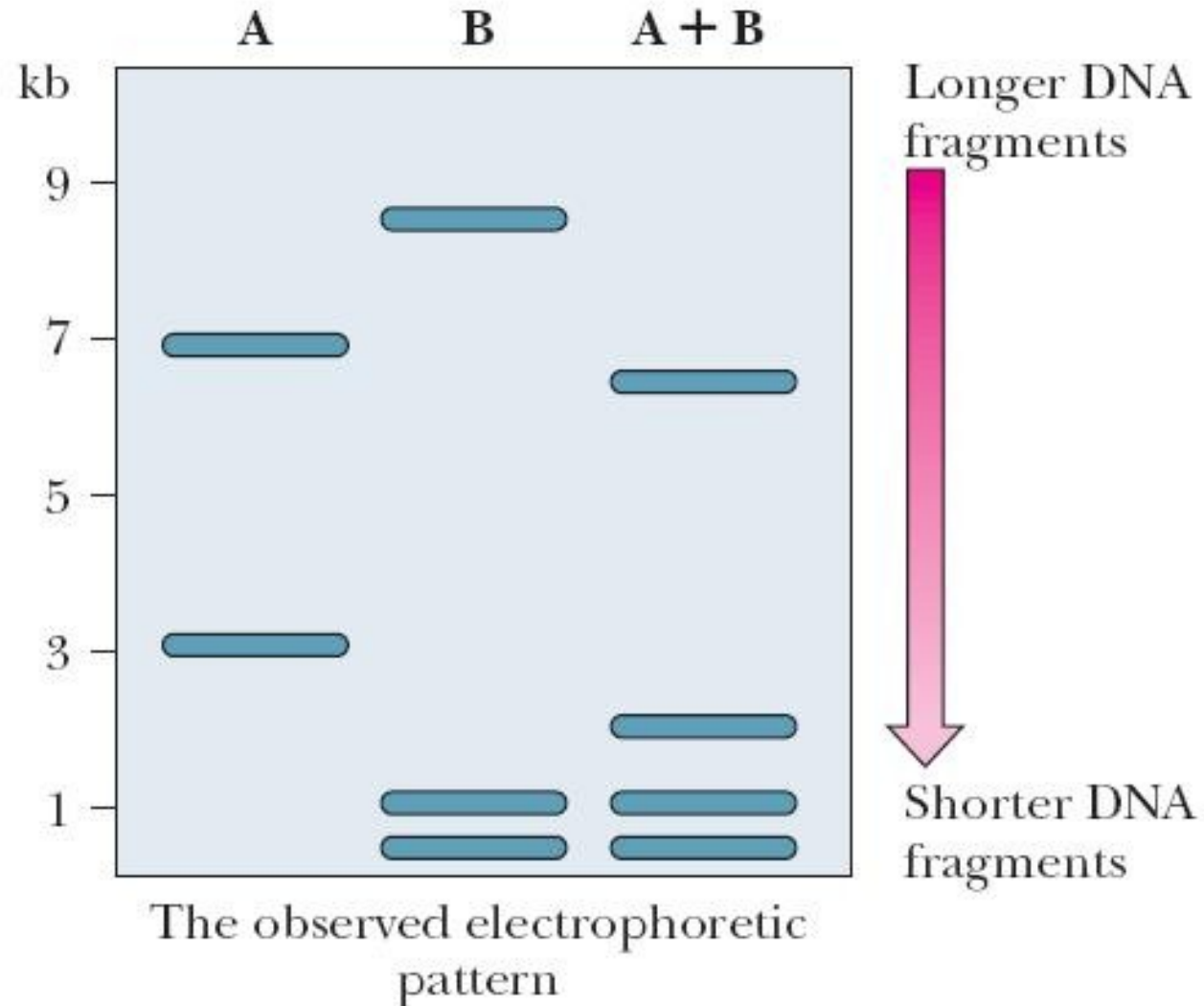
Carte de restriction



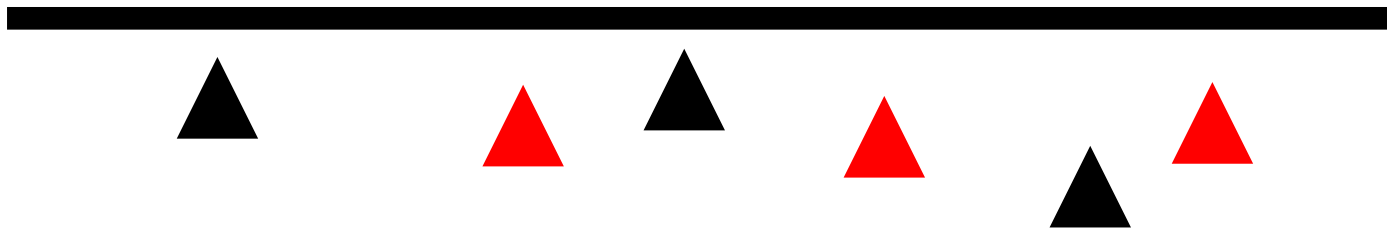
Pas de séquençage, juste la taille des fragments

Carte de restriction: Exemple

Treatment of a linear 10kb DNA molecule with endonucleases gave the following results:

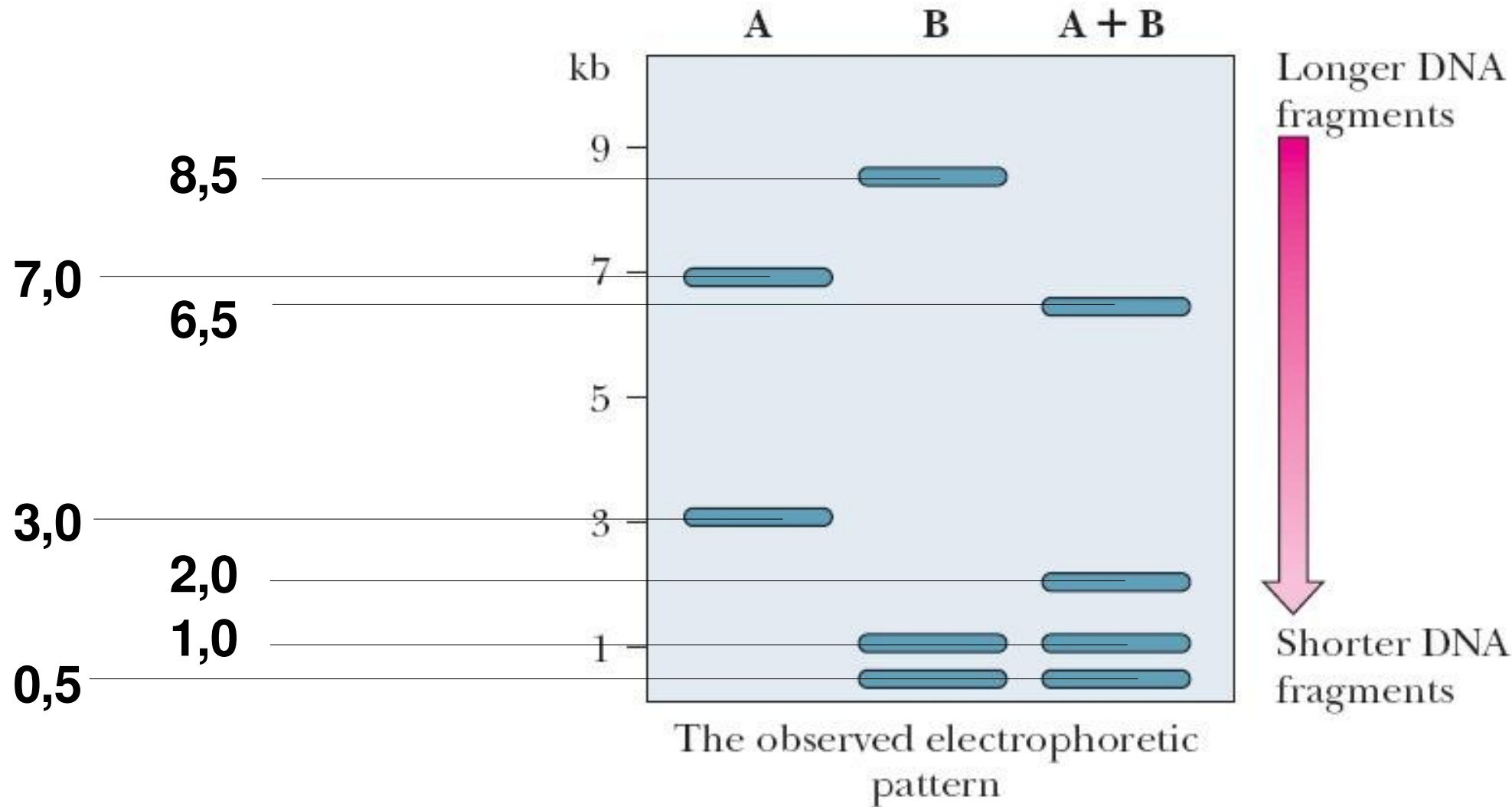


Carte de restriction: Exemple

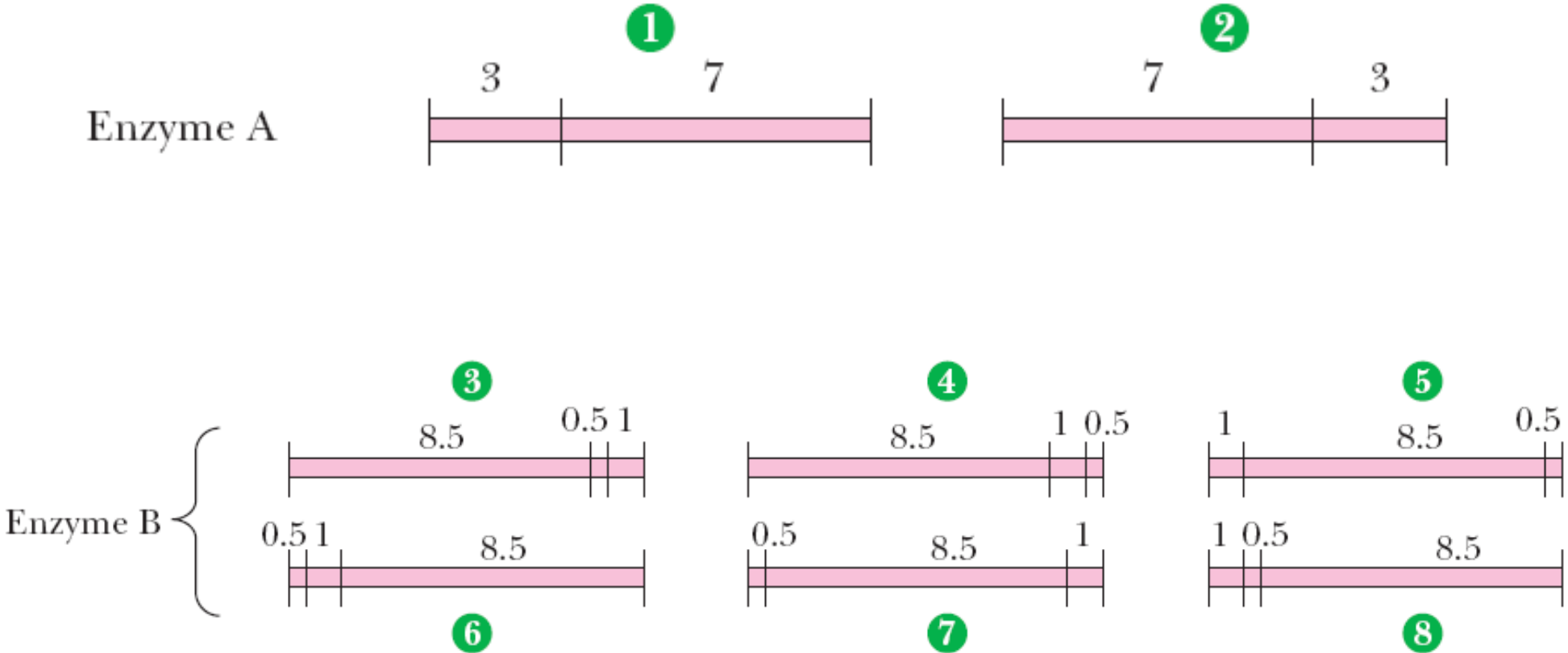


Carte de restriction: Exemple

Treatment of a linear 10kb DNA molecule with endonucleases gave the following results:

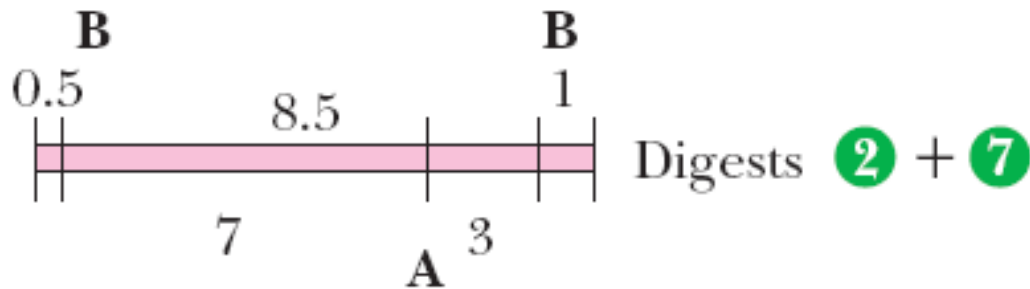
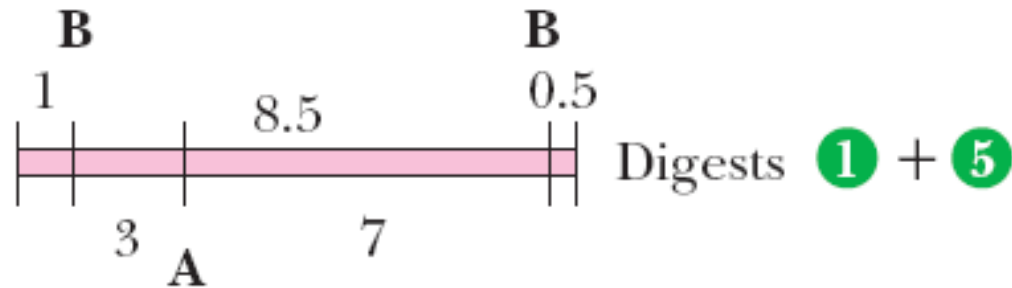


Carte de restriction: Exemple

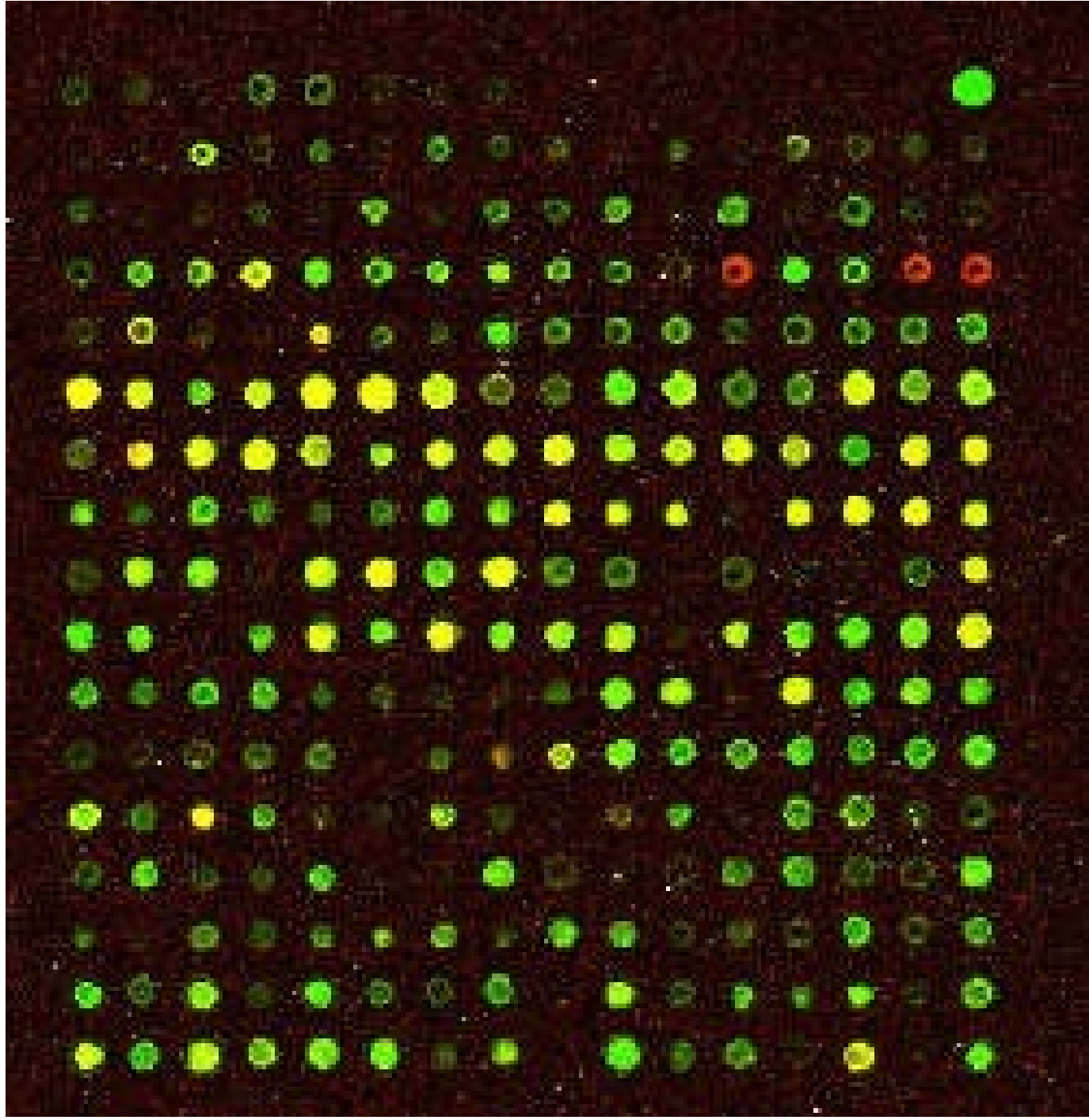


Carte de restriction: Exemple

The only combinations giving the observed **A + B** digests are ① + ⑤ and ② + ⑦



Transcriptome



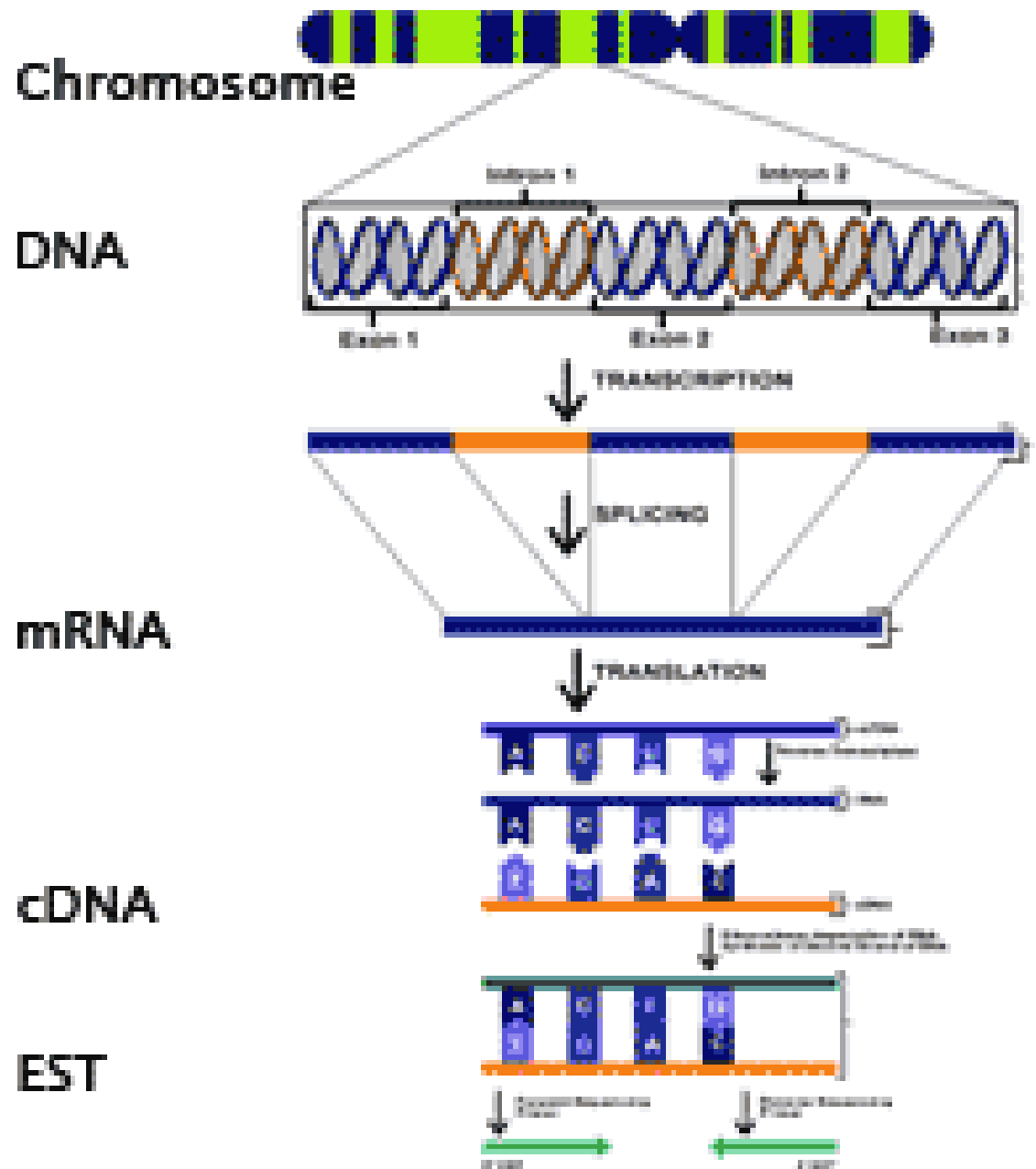
EST

Expressed Sequence Tag
(Etiquette de Séquence Transcrite)

Marqueurs de séquences exprimées.
(= étiquettes = tags).

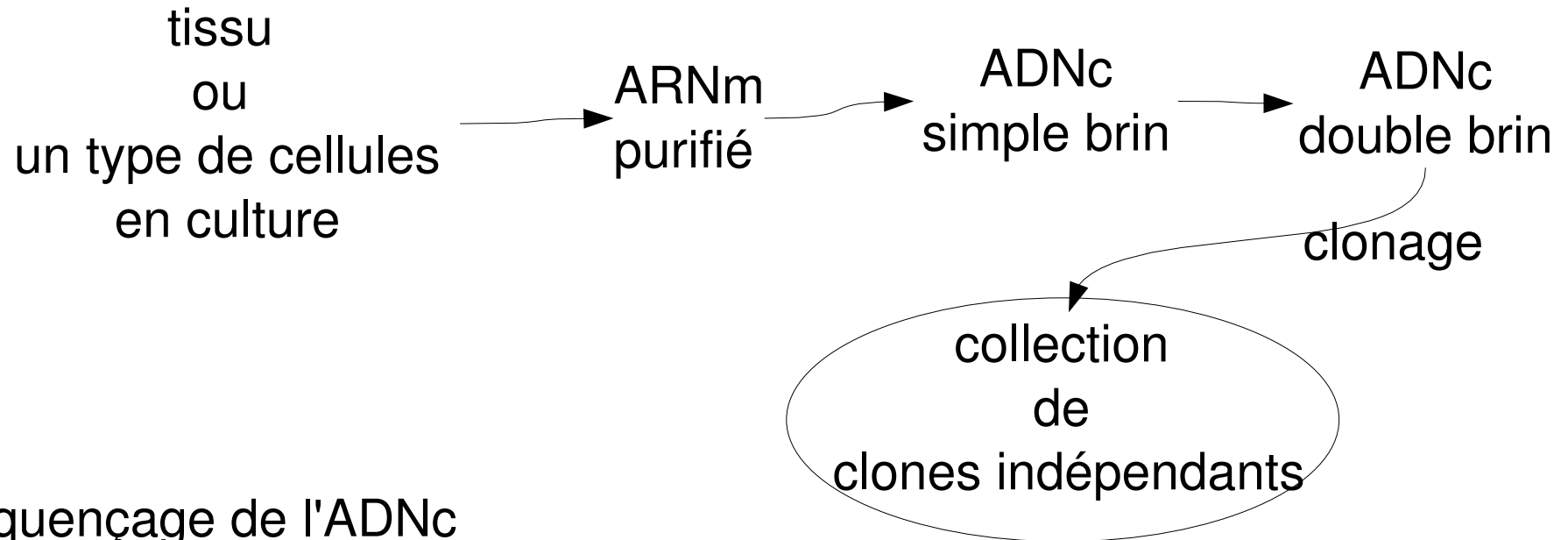
Séquence partielle
d'ADNc de quelques
centaines de nucléotides,
correspondant à une des
extrémités d'un ARNm et
qui lui est spécifique.

Comporte le plus souvent
de nombreuses erreurs.



Stratégie

Construire une "bibliothèque" d'ADNc



Séquençage de l'ADNc

EST :

Séquencage systématique
intensif

mais partiel (qqs 100aines de nucléotides)

taux important d'erreurs dans les séquences (environ 3%)

coût faible

Bases de données

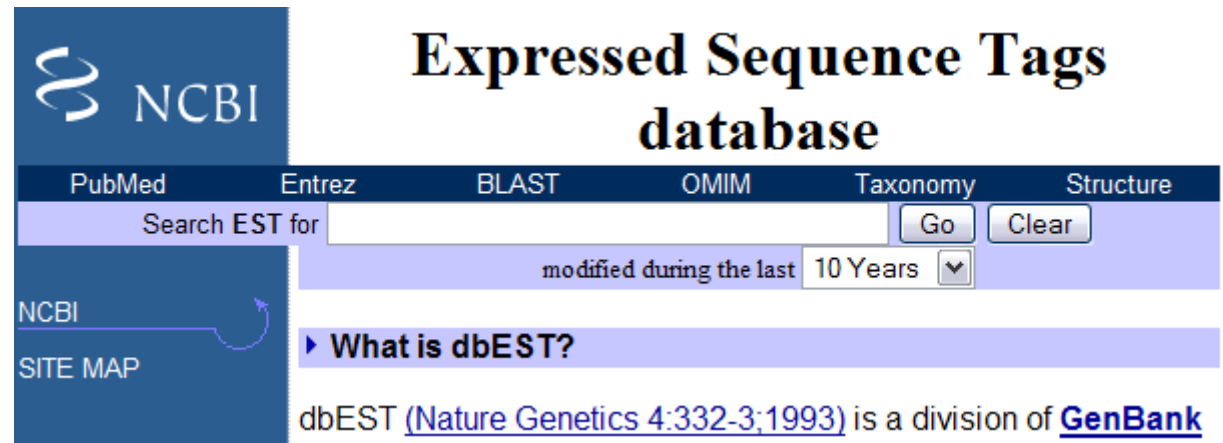


dbEST

banque publique
créée en 1993

centralise les séquences EST issues des différents projets de séquençage.

Forte redondance due au fait que ces séquences ont été obtenues à partir de clones pris au hasard dans de nombreuses banques d'ADNc établies à partir de divers tissus.





Des banques tentent de regrouper les séquences qui a priori proviennent du même gène.

- UNIGENE
- TIGR Genes Indices



UniGene - Mozilla Firefox

Fichier Edition Affichage Aller à Marque-pages Outils ?

[←](#) [→](#) [↺](#) [✕](#) [🏠](#) [🔗](#) http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene [OK](#) [G unigene](#)

[Hotmail](#) [Personnaliser les liens](#) [Windows Media](#) [Windows](#)




UniGene

ORGANIZED VIEW OF THE TRANSCRIPTOME

[My NCBI](#)
[\[Sign In\]](#) [\[Register\]](#)

All Databases PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books

Search UniGene for

NCBI

UniGene

Homepage

Query Tips

FAQ

DDD

Download UniGene

Related Resources

LocusLink

HomoloGene

dbEST

Trace Archive

BLAST

CGAP

UniGene is an experimental system for automatically partitioning GenBank sequences into a non-redundant set of gene-oriented clusters. Each UniGene cluster contains sequences that represent a unique gene, as well as related information such as the tissue types in which the gene has been expressed and map location.

Species	UniGene Entries
Chordata	
Mammalia	
Bos taurus	36,982
Canis familiaris	22,930
Homo sapiens	53,100
Macaca mulatta	4,701
Mus musculus	43,301
Ovis aries	1,714
Rattus norvegicus	38,678
Sus scrofa	32,069
Aves	
Gallus gallus	30,221
Amphibia	
Xenopus laevis	28,040
Xenopus tropicalis	32,769
Actinopterygii	
Danio rerio	31,681
Fundulus heteroclitus	3,154
Oncorhynchus mykiss	24,362
Oryzias latipes	9,848
Salmo salar	8,371
Takifugu rubripes	2,355
Ascidiacea	

Terminé



The Institute for Genomic Research

Home

Research Groups

Scientific Programs

Faculty Home

Database Home

Comprehensive Microbial Resources

Unfinished Microbial Genomes

Plant Genomics

Parasites Databases

Other Eukaryotic Projects

Gene Indices

TIGRFAMs

Fungal Databases

Fibrolitic Ruminal Bacteria

TIGR Microbial Observatories

Genome Properties Database

TGI Data

Gene Indices

What's New | Blast | TGI Software | EGO | Genomic Maps | Resourcer | FAQ

Animal Gene Indices

Plant Gene Indices

Protist Gene Indices

Fungal Gene Indices



Astatotilapia burtoni 1.0
05-20-04



Amblyomma variegatum 1.0
06-10-02



Aedes aegypti 4.0
08-24-04



Atlantic salmon 2.1
06-22-04



Brugia malayi 5.0
01-27-05



Boophilus microplus 1.0
01-12-04



Catfish 6.0
01-24-05



Cattle 11.0
09-28-04



Caenorhabditis elegans 9.0
09-22-04



Canis familiaris 6.0
01-17-05



Chicken 10.0
01-28-05



Ciona intestinalis 4.0
09-21-04



Drosophila 10.0
09-21-04



Fugu 2.0
09-27-04



Haplochromis chilotes 1.0
05-20-04



Haplochromis sp. 'red tail sheller' 1.0



Honey bee 4.0
05-04-04



Human 15.0
08-22-04

Ensembl Genome Browser - Mozilla Firefox

Fichier Edition Affichage Aller à Marque-pages Outils ?

http://www.ensembl.org/index.html OK ensembl

Hotmail Personnaliser les liens Windows Media Windows

Questions fréquemment posées à propos d... Ensembl Genome Browser

e! Ensembl Search all Ensembl: Anything Go

Ensembl v33 - Sep 2005 Help

Use Ensembl to...

- Run a BLAST search
- Search Ensembl
- Data mining [BioMart]
- Upload your own data
- Export data
- Download data

Docs and downloads

- Information
- What's New
- About Ensembl
- Ensembl data
- Software

Other links

- Home
- Sitemap
- Vega
- Pre! Pre Ensembl
- Trace server
- Archive! sites
- Stable Archive! link for this

browse a genome

Mammals

-  **Homo sapiens** [NCBI 35]
browse | what's new | Vega
-  **Pan troglodytes** [CHIMP1]
browse | what's new
-  **Macaca mulatta** [Mmul 0.1]
browse pre! site
-  **Mus musculus** [NCBI m34]
browse | what's new | Vega
-  **Rattus norvegicus** [RGSC 3.4]
browse | what's new
-  **Canis familiaris** [CanFam1.0]
browse | what's new | Vega
-  **Bos taurus** [Btau 1.0]
browse | what's new | pre!
[Btau 2.0]
-  **Monodelphis domestica** [MonDom2]
browse pre! site

Other chordates

-  **Gallus gallus** [WASHUC1]
browse | what's new
-  **Xenopus tropicalis** [JGI 3]
browse | what's new
-  **Danio rerio** [WTSI Zv5]
browse | what's new | Vega
-  **Takifugu rubripes** [Fugu 2.0]
browse | what's new
-  **Tetraodon nigroviridis** [TETRAODON 7]
browse | what's new
-  **Ciona intestinalis** [JGI 1.95]
browse | what's new

Other eukaryotes

-  **Drosophila melanogaster** [BGDP 4]
browse | what's new
-  **Anopheles gambiae** [MOZ 2]
browse | what's new
-  **Apis mellifera** [Amel 2.0]
browse | what's new
-  **Caenorhabditis elegans** [WS140]
browse | what's new
-  **Saccharomyces cerevisiae** [SGD]
browse | what's new

Rechercher : contig Occurrence suivante Occurrence précédente Surligner Respecter la casse Bas de la page atteint, la recherche se poursuit depuis le début

Transfert des données depuis www.ensembl.org...

Human (Homo sapiens) - Mozilla Firefox

Fichier Edition Affichage Aller à Marque-pages Outils ?

[←](#)
[→](#)
[↺](#)
[✕](#)
[🏠](#)
[http://www.ensembl.org/Homo_sapiens/index.html](#)
[OK](#)
[G ensembl](#)

[Hotmail](#)
[Personnaliser les liens](#)
[Windows Media](#)
[Windows](#)

[Questions fréquemment posées à propos d...](#)
[Human \(Homo sapiens\)](#)



Search e! Human: [Go](#)

e.g. [AC003985.2.1.104492](#), [RH9632](#), [ENSG00000139618](#)

Ensembl v33 - Sep 2005 [Help](#)

Use Ensembl to...

- Run a BLAST search
- Search Ensembl
- Data mining [BioMart]
- Upload your own data
- Export data
- Download data

Docs and downloads

- Information
- What's New
- About Ensembl
- Ensembl data
- Software

Select a species

- Mammals
- Other chordates
- Other eukaryotes

Other links

- Home
- Sitemap

Explore the Human genome

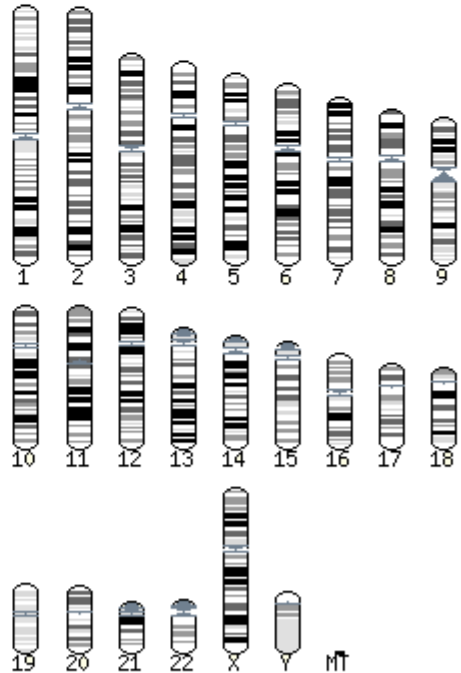
What's New in Ensembl 33

- New Human cDNAs**
Monthly update of the human cDNA database.
- Vega Human**
Updated Ensembl-Vega database for Homo sapiens.
- Affy mapping**
Some Affymetrix data was accidentally omitted from the last release of Ensembl (v32, July 2005). These data have been re-computed and added to the core databases of the affected species.
- Fixes to affy_feature**
The affy_feature table in the core database had no values in the 'mismatches' column in the last release; the correct values have been added.
- Gene descriptions**
The missing gene descriptions for ncRNAs have been added to the Human core database.

[More news...](#)

Karyotype

Click on a chromosome for a closer view

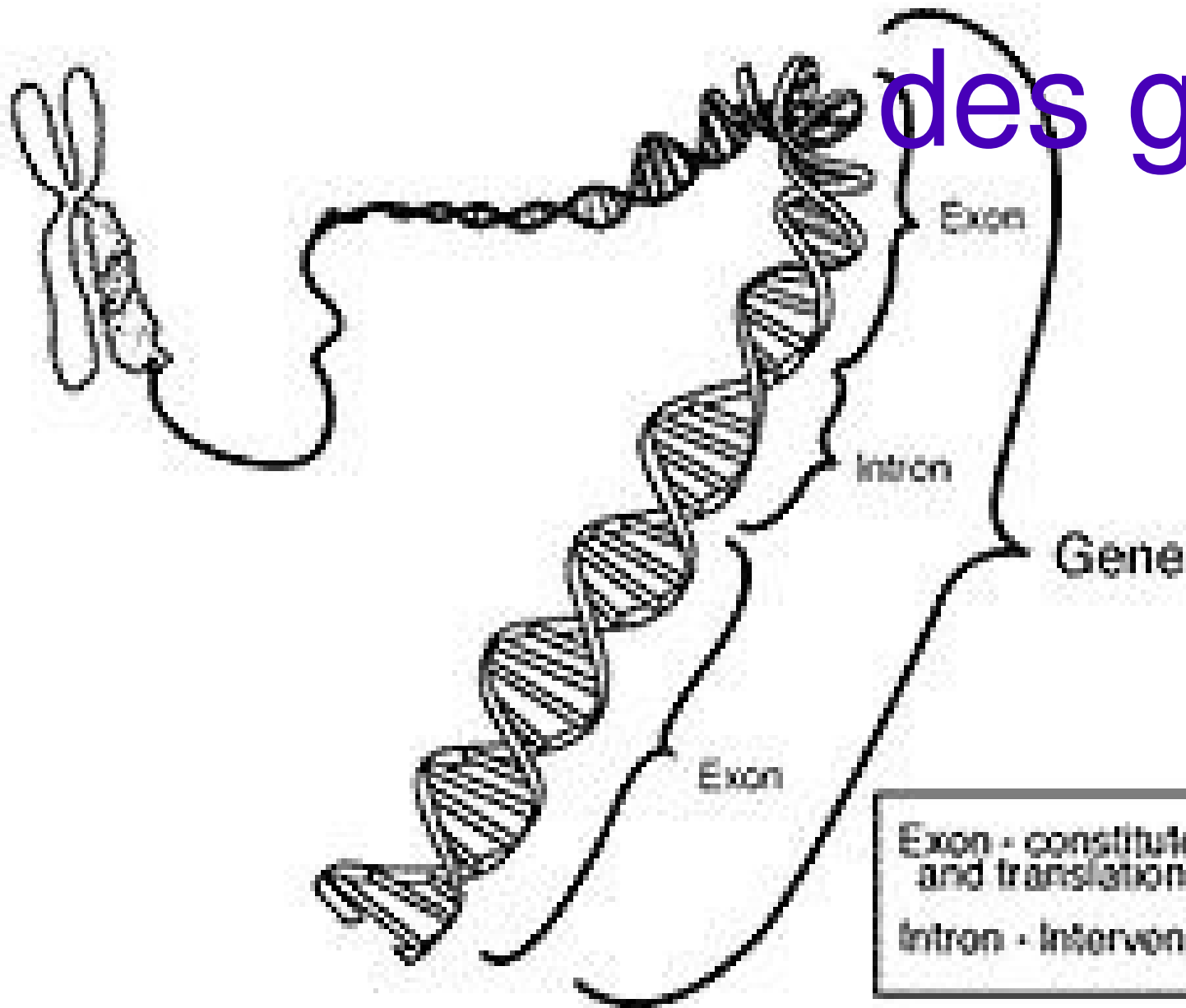


Jump directly to sequence position

Rechercher :
[Occurrence suivante](#)
[Occurrence précédente](#)
[Surligner](#)
[Respecter la casse](#)
[Bas de la page atteint, la recherche se poursuit depuis le début](#)

Terminé

Annotation des génomes



Exon - constitute the mRNA
and translation into protein
Intron - Intervening sequence

Structure des génomes

Règle de Chargaff: $A=T$ et $G=C$

Grande variation de la composition d'un génome à l'autre.

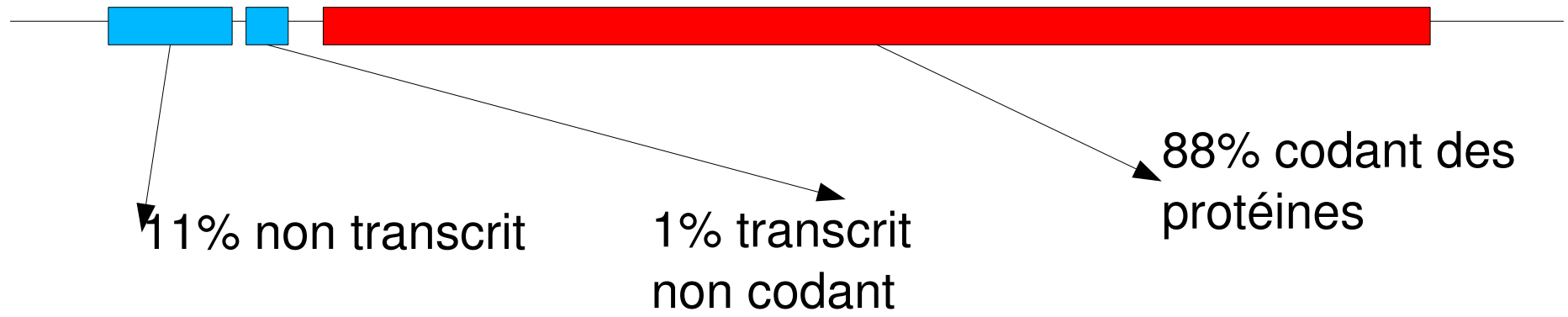
Ex: Archaeobactéries généralement riches en AT

Organisme	A	T	G	C
Humain	30.9	29.4	19.9	19.8
E. coli	24.7	23.6	26.0	25.7
Methanococcus	34.5	34.2	15.9	15.5

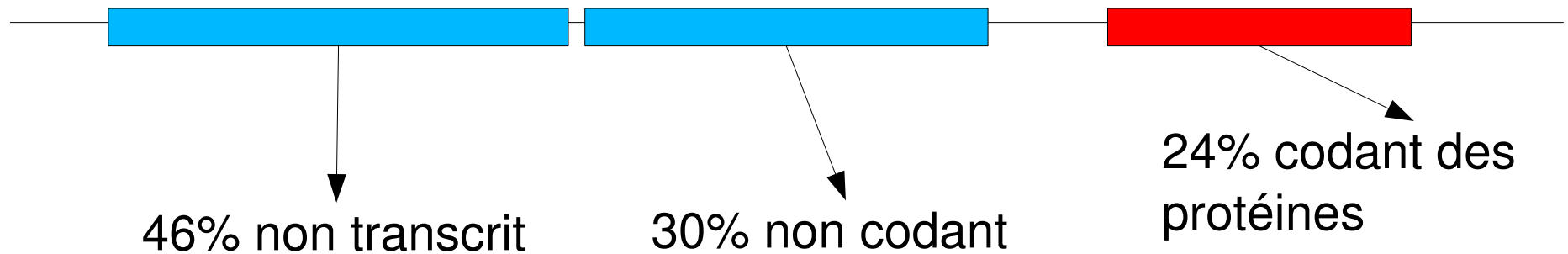
Régions transcrites plus riches en GC que les non transcrites

Structure du génome

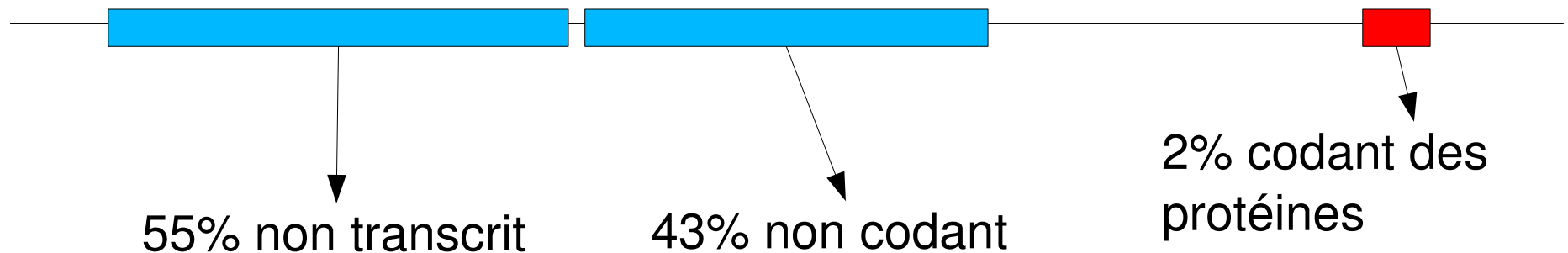
E. coli



C. elegans



Homme



Structure du génome

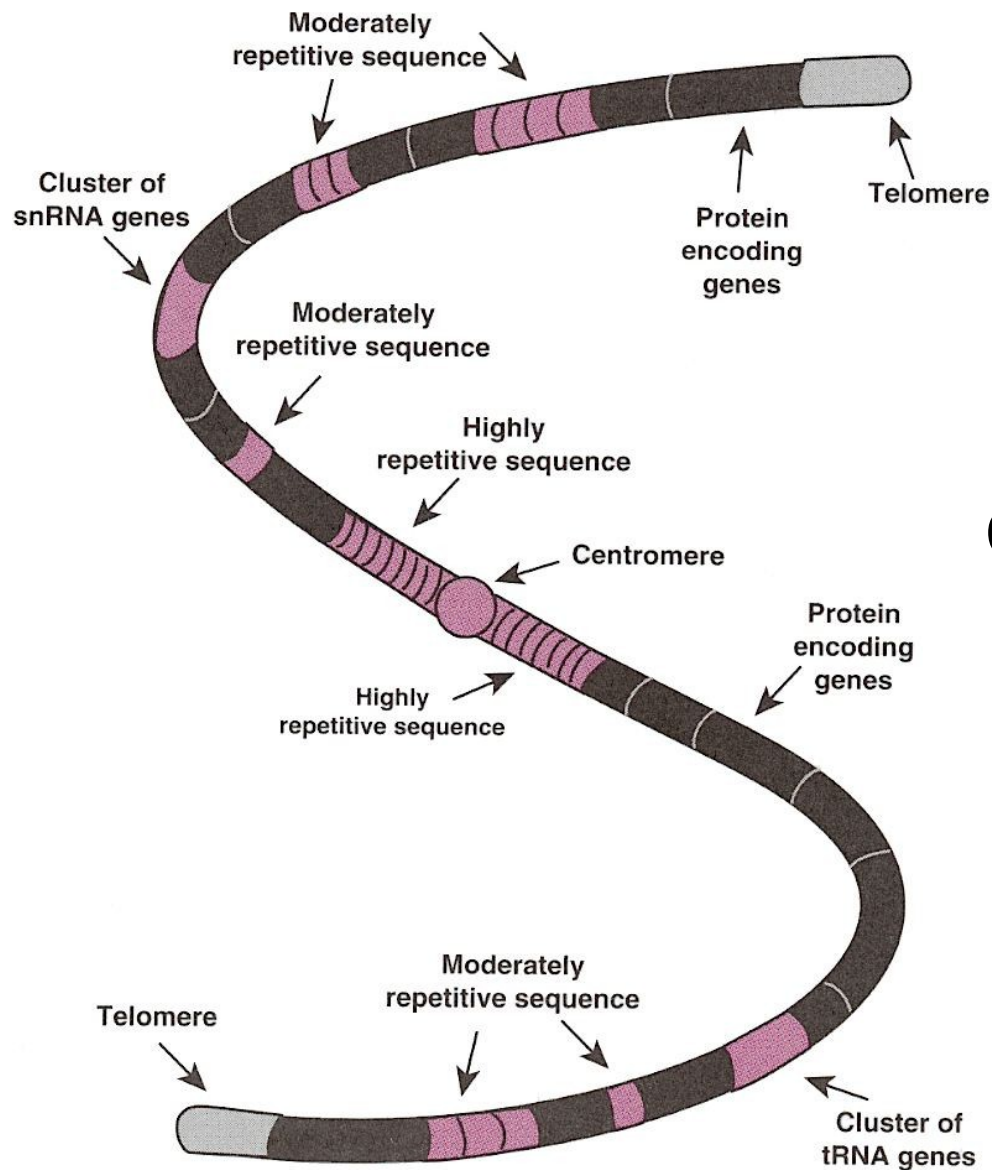
Procaryotes

- ★ Longueur moyenne 950 nt
- ★ densité en gènes : 95% génome transcrit (E.coli)
- ★ Gènes organisés en opérons. 600 opérons (E. coli)

Eucaryotes

- ★ Longueur moyenne 27 kb avec 8,8 introns
 - ★ exon moyen: 145bp
 - ★ Intron moyen : 3365bp
- ★ densité en gènes :
 - ★ 1 gène tous les 100kb (homme)
 - ★ 1 gène tous les 5-6kb (C. elegans)
 - ★ 1 gène tous les 2kb (S. cerevisiae)

STYLIZED EUKARYOTIC CHROMOSOME



Compartimentation
structurale
du génome eucaryote

Compartimentation structurale du génome

Organisation en classes de séquences répétées et uniques

- ★ Séquences uniques (60% génome humain).

 - ★ La plupart des gènes structuraux

- ★ Séquences localisées répétées en tandem

 - ★ ADN satellite (+10Mpb) au niveau des centromères et télomères

 - ★ Séquences minisatellites (blocs de 100 à 2000pb)

 - ★ Séquences microsatellites (< 100pb) STR:*Short Tandem Repeat*)

- ★ Séquences répétées dispersées (chez vertébrés)

 - ★ Séquences sine (*Short INterspersed Elements*). Séquences ALU

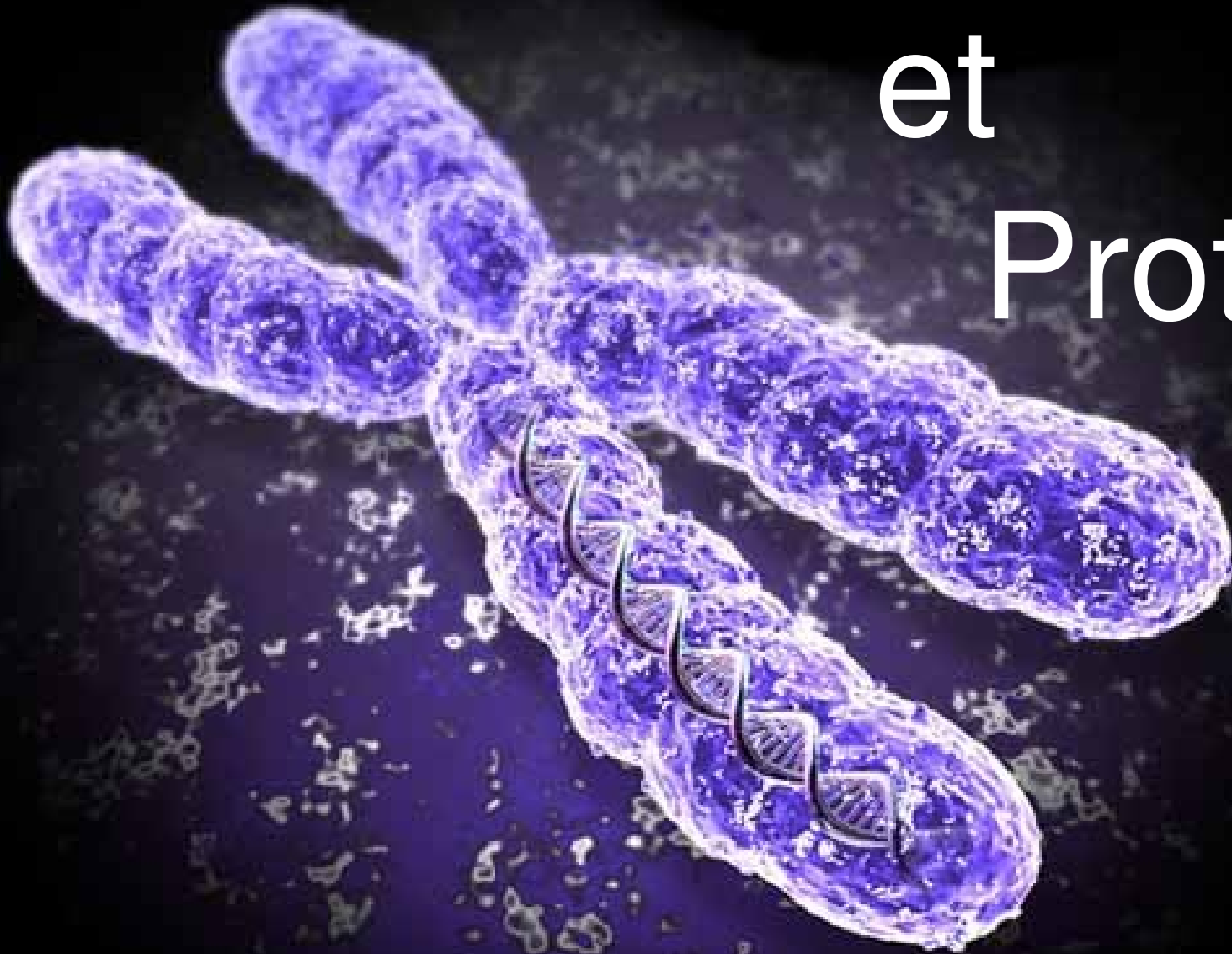
 - ★ Séquences line (Long *INterspersed Elements*)

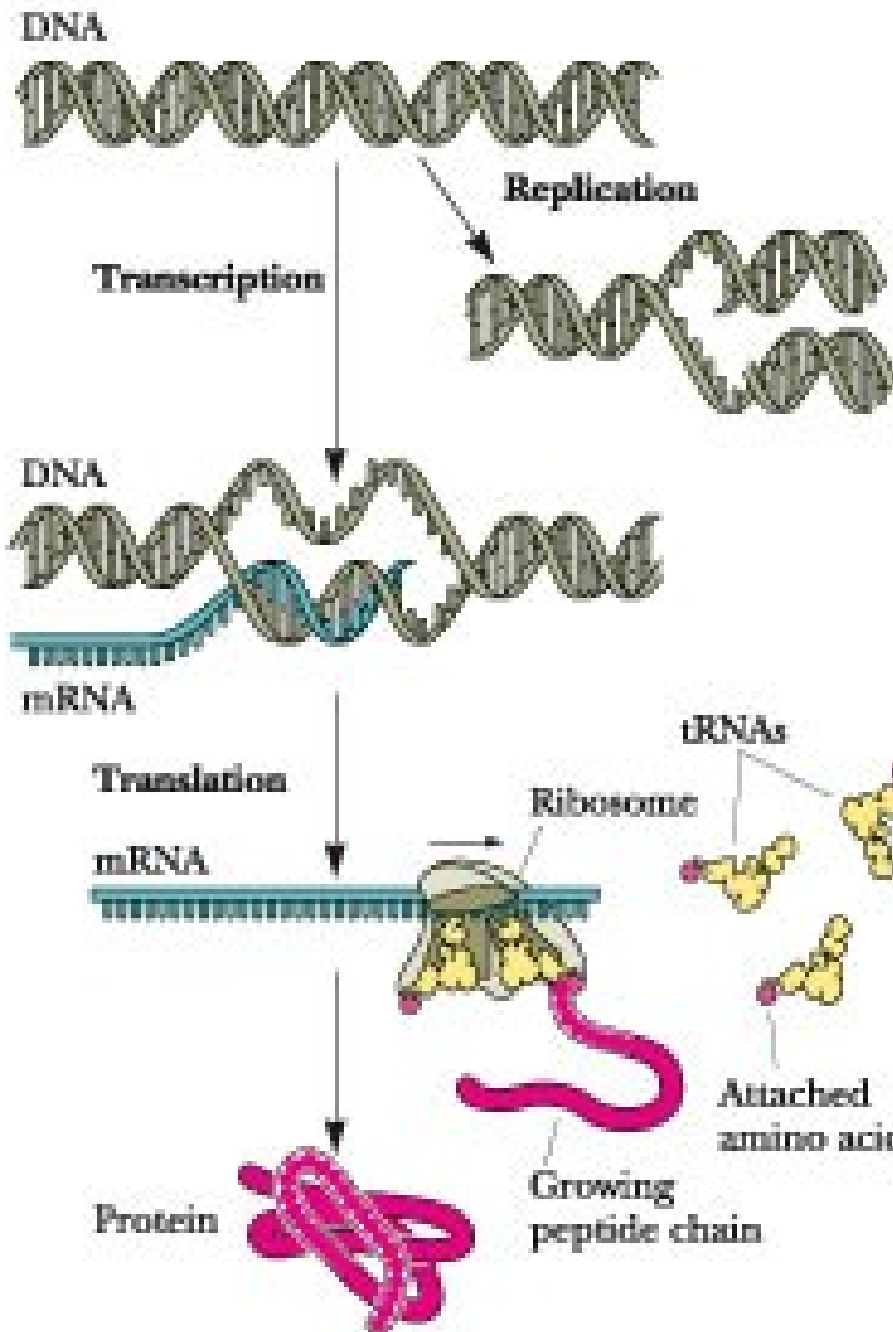
 - ★ Séquences microsatellites (< 100pb) STR:*Short Tandem Repeat*)

Organisation en bandes chromosomiques

Organisation en domaines de composition nucléotidique homogène (isochore)

Relation ADN-ARN et Protéines





Replication

DNA replication yields two DNA molecules identical to the original one, ensuring transmission of genetic information to daughter cells with exceptional fidelity.

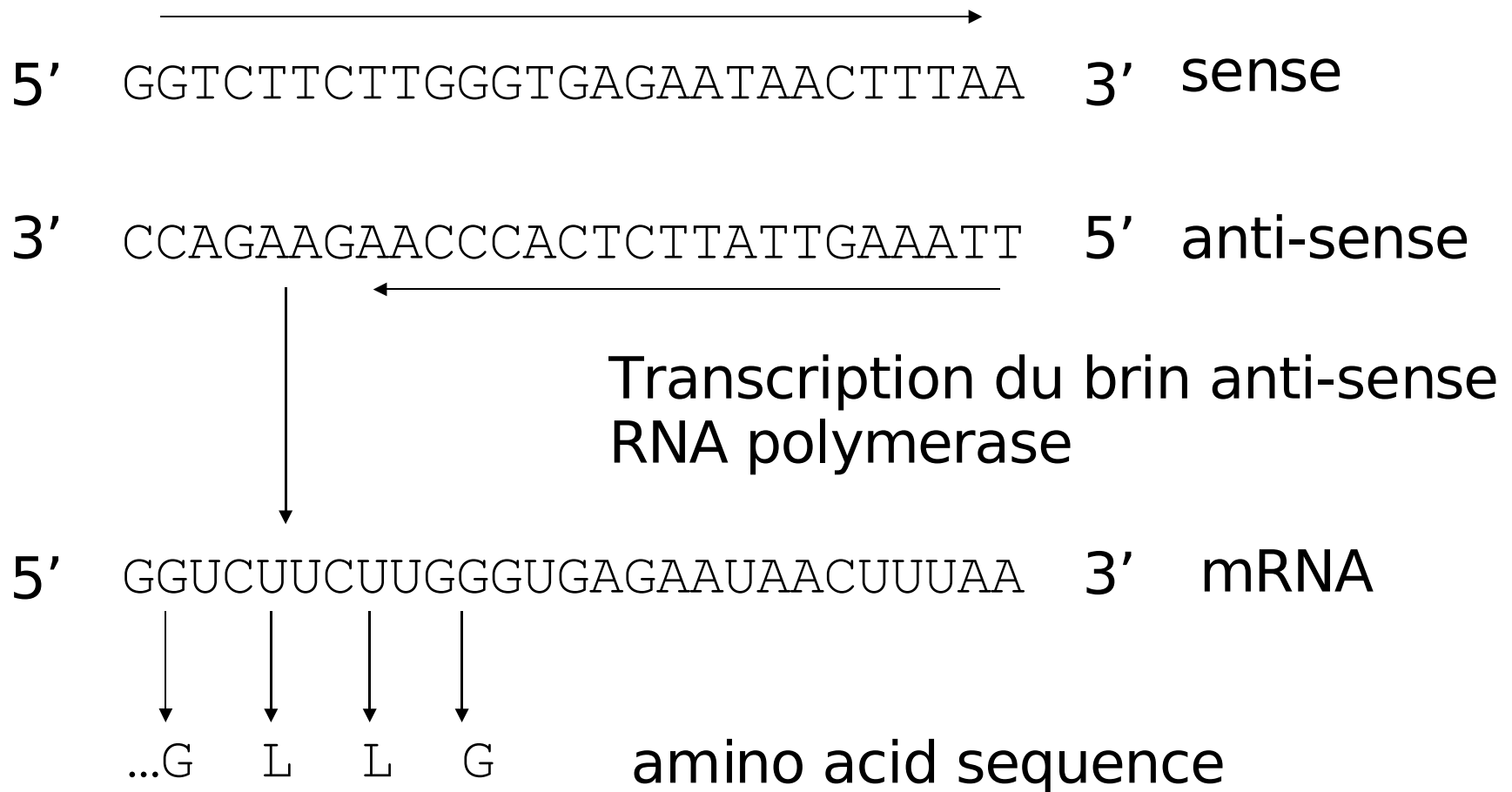
Transcription

The sequence of bases in DNA is recorded as a sequence of complementary bases in a single-stranded mRNA molecule.

Translation

Three-base codons on the mRNA corresponding to specific amino acids direct the sequence of building a protein. These codons are recognized by tRNAs (transfer RNAs) carrying the appropriate amino acids. Ribosomes are the "machinery" for protein synthesis.

Transcription et traduction



Code génétique

Triplet
(codon) → Acide aminé

4^3 → 20

- Dégénéré
- Universel (presque)

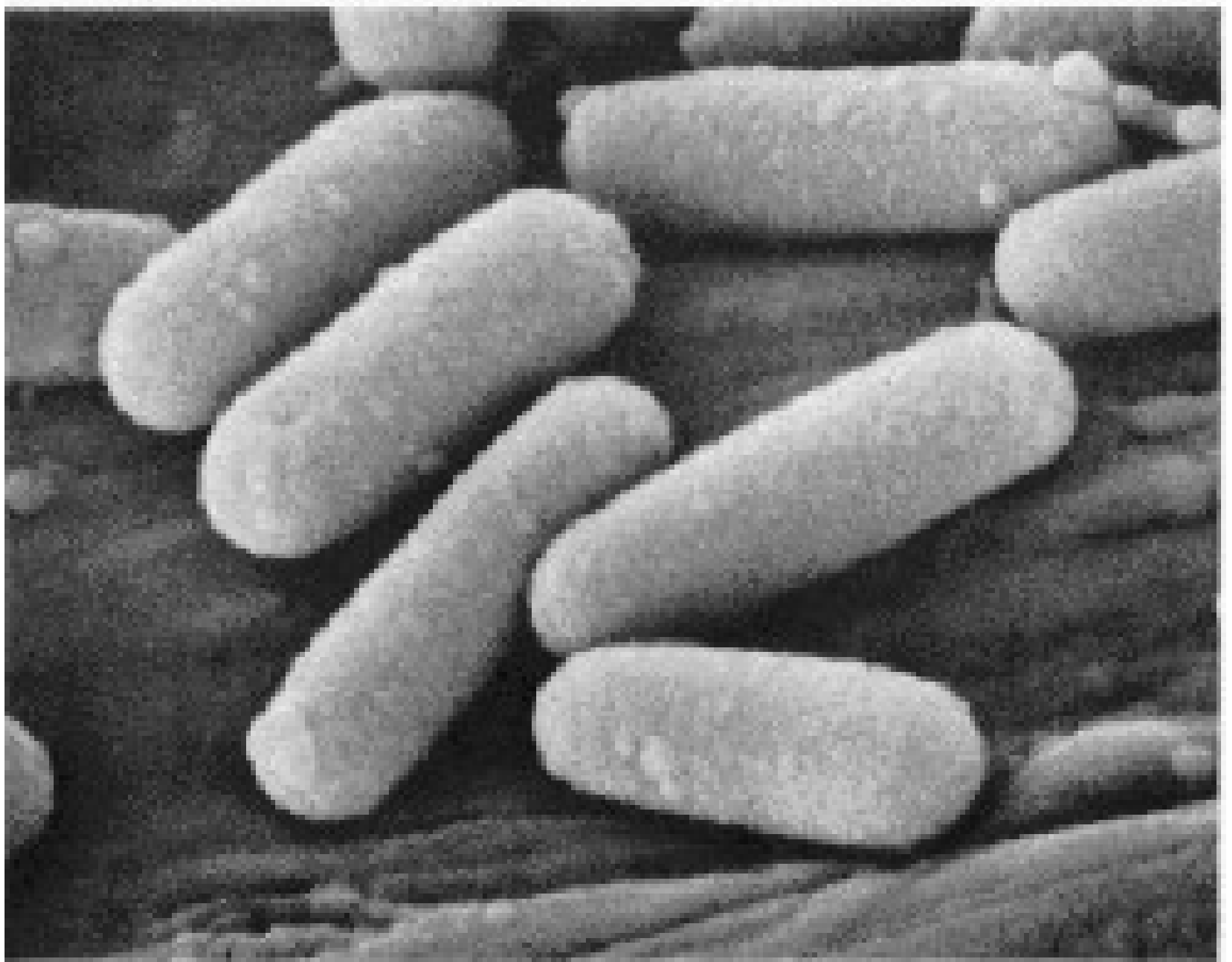
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Prédiction de structure d'un gène -ORF

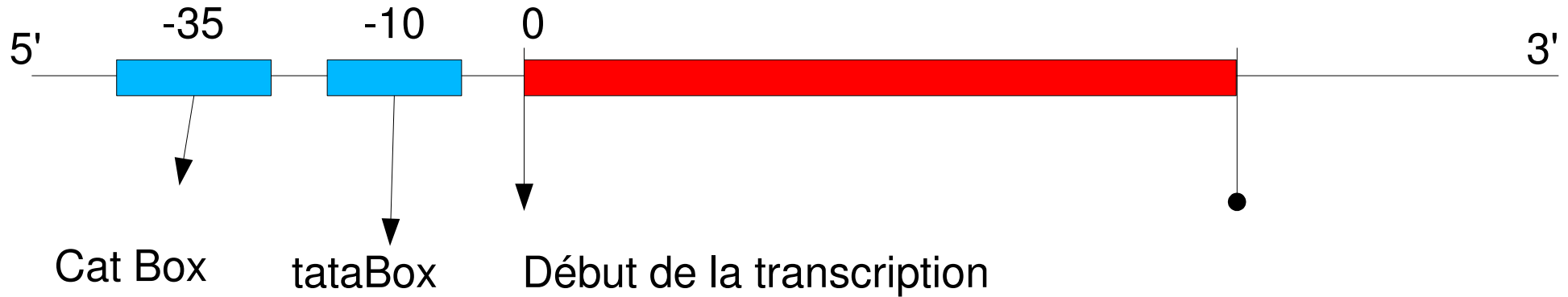
- ★ Recherche des différentes sous-séquences qui composent l'ADN

- ★ Phase ouverte potentielle

- ★ Sites signaux: promoteur, régions régulatrices, facteurs de transcription, sites de poly-adénylation, sites d'épissage, sites de début et de fin de traduction.



Structure du gène chez les Procaryotes



★ Début de la transcription: promoteur

★ 40 paires de nucléotides avant le début de la transcription

★ Séquence -35

★ Séquence -10 (TATA box)

★ Progression de la transcription: 5' -> 3'

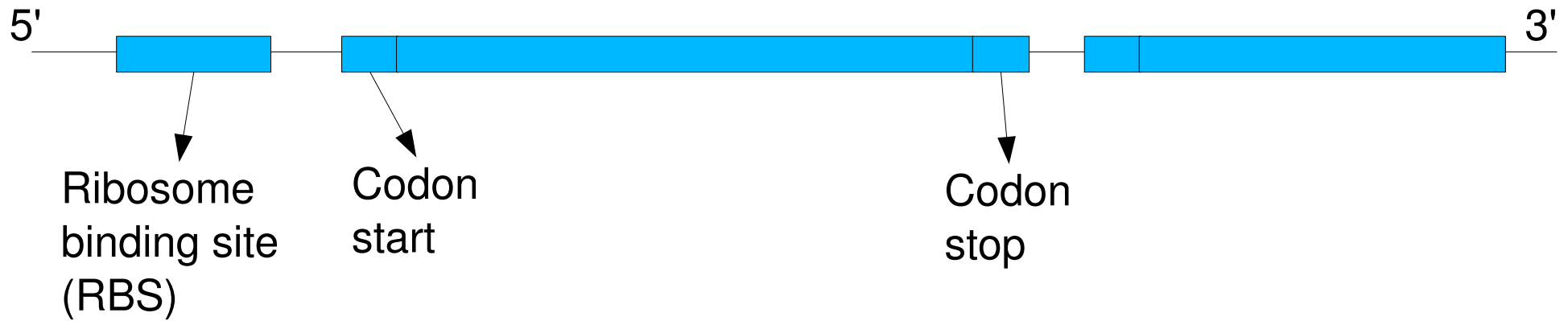
★ Terminaison de la transcription

★ Région à symétrie imparfaite

★ Région riche en AT

Transcription chez les Procaryotes

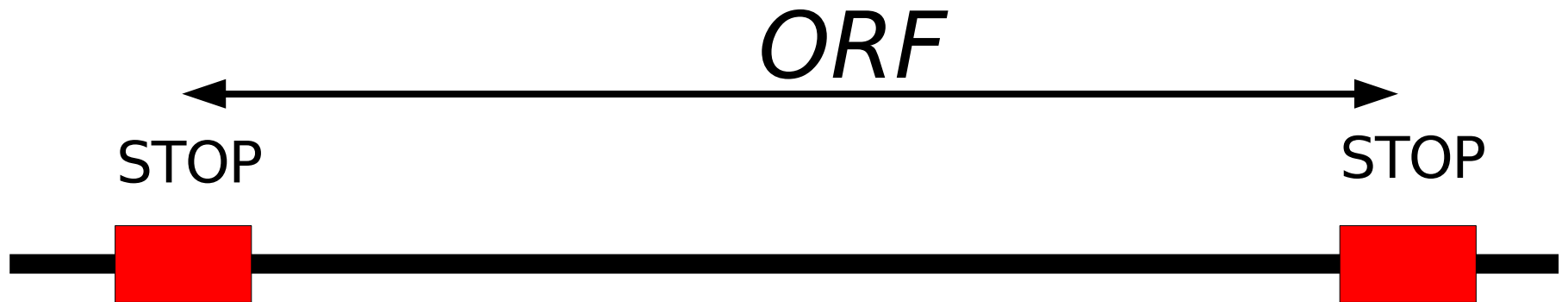
ARN polycistronique



ORF : Open Reading Frame

(= Phase ouverte de lecture)

La phase ouverte (ORF, Open Reading Frame) est la région de l'ADN qui sépare deux codons STOP (donc potentiellement codante).



		1															2	
1							0										0	
-----+-----+																		
t	g	c	t	g	a	g	c	a	g	c	t	c	c	a	g	a	c	t

.	T			C			A			G			.
T	TTT	Phe	F	TCT	Ser	S	TAT	Tyr	Y	TGT	Cys	C	T
	TTC	Phe	F	TCC	Ser	S	TAC	Tyr	Y	TGC	Cys	C	C
	TTA	Leu	L	TCA	Ser	S	TAA	Stop		TGA	Stop		A
	TTG	Leu	L	TCG	Ser	S	TAG	Stop		TGG	Trp	W	G
C	CTT	Leu	L	CCT	Pro	P	CAT	His	H	CGT	Arg	R	T
	CTC	Leu	L	CCC	Pro	P	CAC	His	H	CGC	Arg	R	C
	CTA	Leu	L	CCA	Pro	P	CAA	Gln	Q	CGA	Arg	R	A
	CTG	Leu	L	CCG	Pro	P	CAG	Gln	Q	CGG	Arg	R	G
A	ATT	Ile	I	ACT	Thr	T	AAT	Asn	N	AGT	Ser	S	T
	ATC	Ile	I	ACC	Thr	T	AAC	Asn	N	AGC	Ser	S	C
	ATA	Ile	I	ACA	Thr	T	AAA	Lys	K	AGA	Arg	R	A
	AUG	Met	M	ACG	Thr	T	AAG	Lys	K	AGG	Arg	R	G
G	GTT	Val	V	GCT	Ala	A	GAT	Asp	D	GGT	Gly	G	T
	GTC	Val	V	GCC	Ala	A	GAC	Asp	D	GGC	Gly	G	C
	GTA	Val	V	GCA	Ala	A	GAA	Glu	E	GGA	Gly	G	A
	GTG	Val	V	GCG	Ala	A	GAG	Glu	E	GGG	Gly	G	G

			1				2
1			0				0
- - - - - + - - - - - +							
tgc	tga	gca	gct	cca	gag	aact	
C	*	A	A	P	E	T	

							1						2					
1							0						0					
-----+-----+-----+-----+-----+																		
t	g	c	t	g	a	g	c	a	g	c	t	c	c	a	g	a	c	t
C			*		A		A		P		E		T					
	A		E		Q		L		Q		R		.	.	.			
	L		S		S		S		R		D		.	.				

5' -tgctgagcagctccagagact-3'

5' -tgctgagcagctccagagact-3'

: : : : : : : : : : : : : : : : : :

3' -acgactcgtcgaaggctctctga-5'



5' -agtctctctggagctgctcagca-3'

S L W S C S A

Prédiction des ORFs

- ★ORF (Open Reading Frame) : Phase Ouverte

- ★Région qui sépare deux codons STOP

- ★A l'intérieur, région CDS qui commence par codon START et finit par codon STOP

Il y a **six** cadres de lecture possibles dans la séquence :
trois par brin, avec un décalage de un nucléotide pour passer d'une phase à une autre, et sur les deux brins.

6 séquences virtuelles différentes

Prédiction des ORFs: méthodes (1)

Recherche des phases ouvertes longues

- ★ Si distribution aléatoire, un codon stop tous les 21 codons ($64/3=21$)
- ★ Protéine moyenne = 1000pb
- ★ Recherche de longues séquences sans codon stop pour les 3 cadres de lecture.
- ★ Quand détection, programme balaye dans le sens opposé, un codon START
- ★ Algorithme incapable de détecter les gènes courts (< 300pb)
- ★ + de ORFs que de gènes. Ex: 6500 ORFs dans ADN de *E. Coli* alors que 4400 gènes

Prédiction des ORFs: méthodes (2)

Fréquences des codons dans les régions codantes

- ★Analyse de l'usage des codons pour le comparer à celui d'un gène typique de l'organisme
- ★Biais de l'usage du code génétique = nature fonctionnelle des gènes
- ★Méthodes statistiques pour déduire une carte de distribution des gènes selon l'usage des codons
- ★Corrélation entre le biais d'usage et les fonctions biologiques des gènes

Prédiction des ORFs: méthodes (3)

Fréquences des codons dans les régions codantes: méthodes GC3

- ★le contenu en G+C du génome de certaines bactéries (*Streptomyces*) est très élevé -> biais dans l'usage des codons.
- ★Les gènes de *Streptomyces* ont une 3ème lettre dont le contenu en G+C est de 92%
- ★Distribution très élevée en G+C sur la 3ème lettre du codon (GC3)
- ★Algorithme recherche donc les zones riches en G+C qui sont potentiellement des phases ouvertes
- ★FramePlot (www.nih.go.jp/~jun/frameplot.pl)

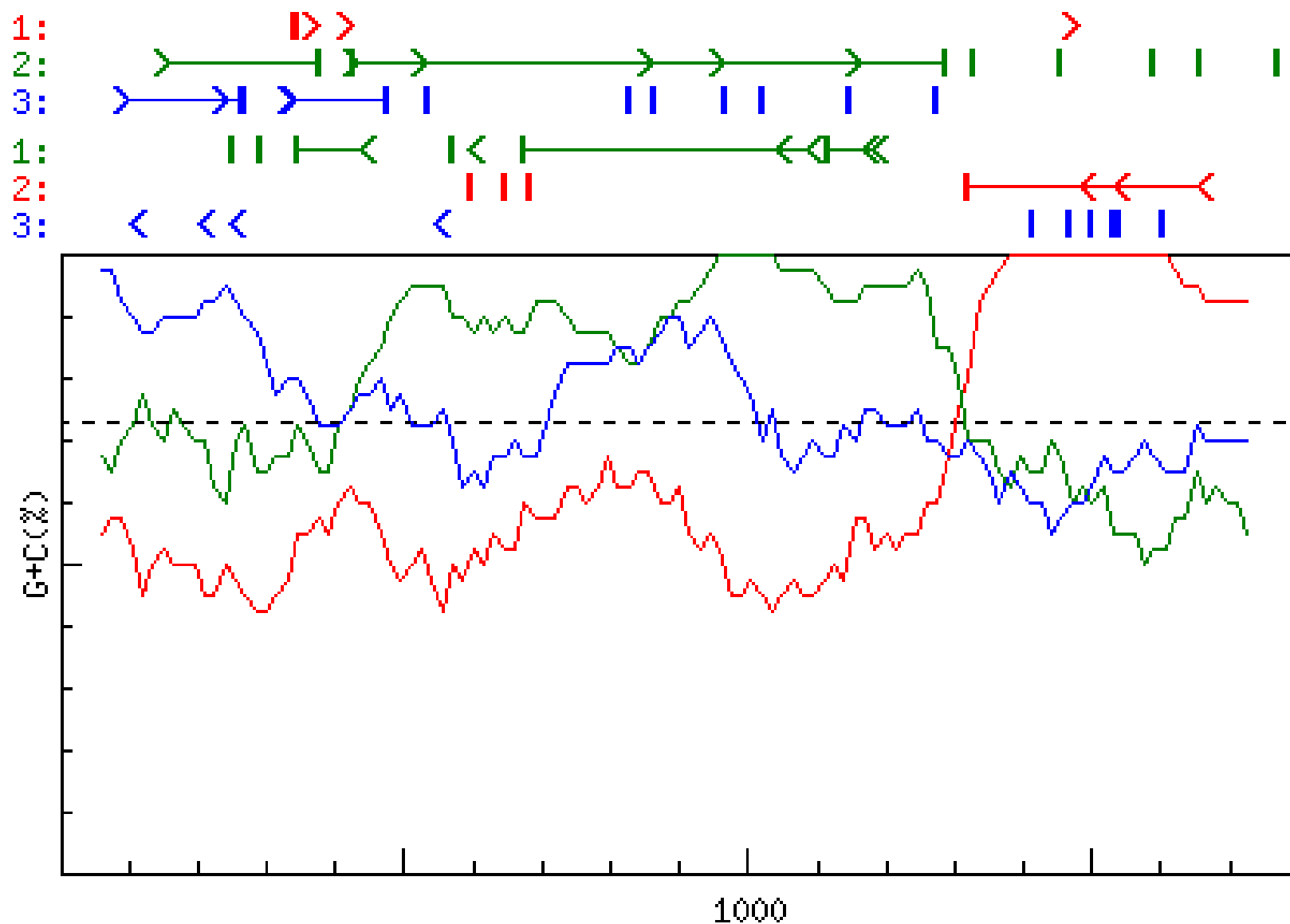
ORGANISM Streptomyces griseus

FramePlot 2.1 - (c) 1996-1998, ISHIKAWA Jun

Target: STHKAN 1794 bp; 72.7% G+C (dashed line)

Window: 40, Step: 5, Start codon [>]: ATG

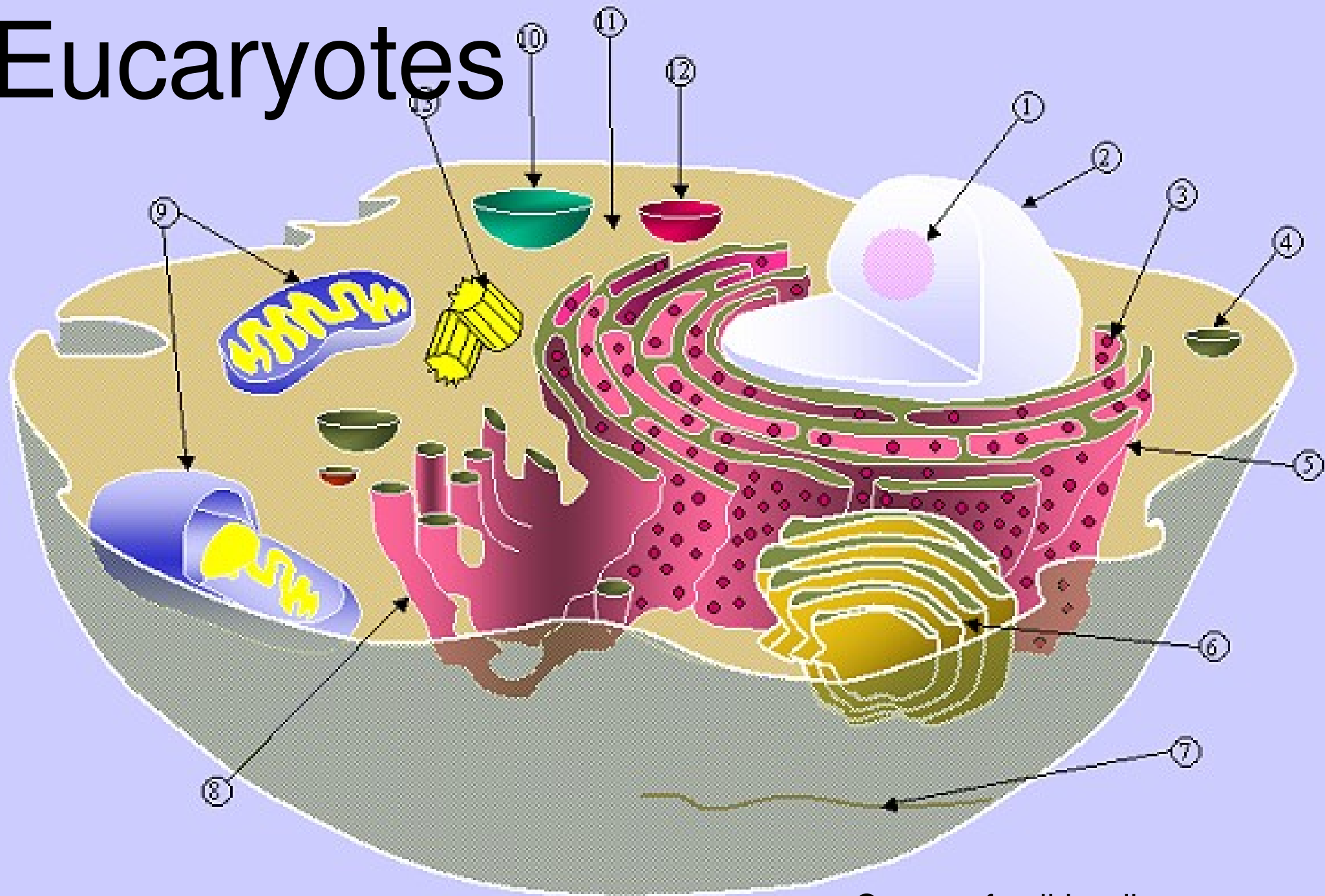
Minimum ORF: 20, Date: Jul 13 11:44:22 1998



LOCUS STMKAN 1794 bp DNA BCT 17-
APR-1992
DEFINITION S.griseus kan gene encoding aminoglycoside 3-N-
acetyltransferase.
ACCESSION D00681
NID g217013
KEYWORDS aminoglycoside 3-N-acetyltransferase; kan gene.
SOURCE S.griseus (strain SS-1198PR) genomic DNA.
ORGANISM Streptomyces griseus

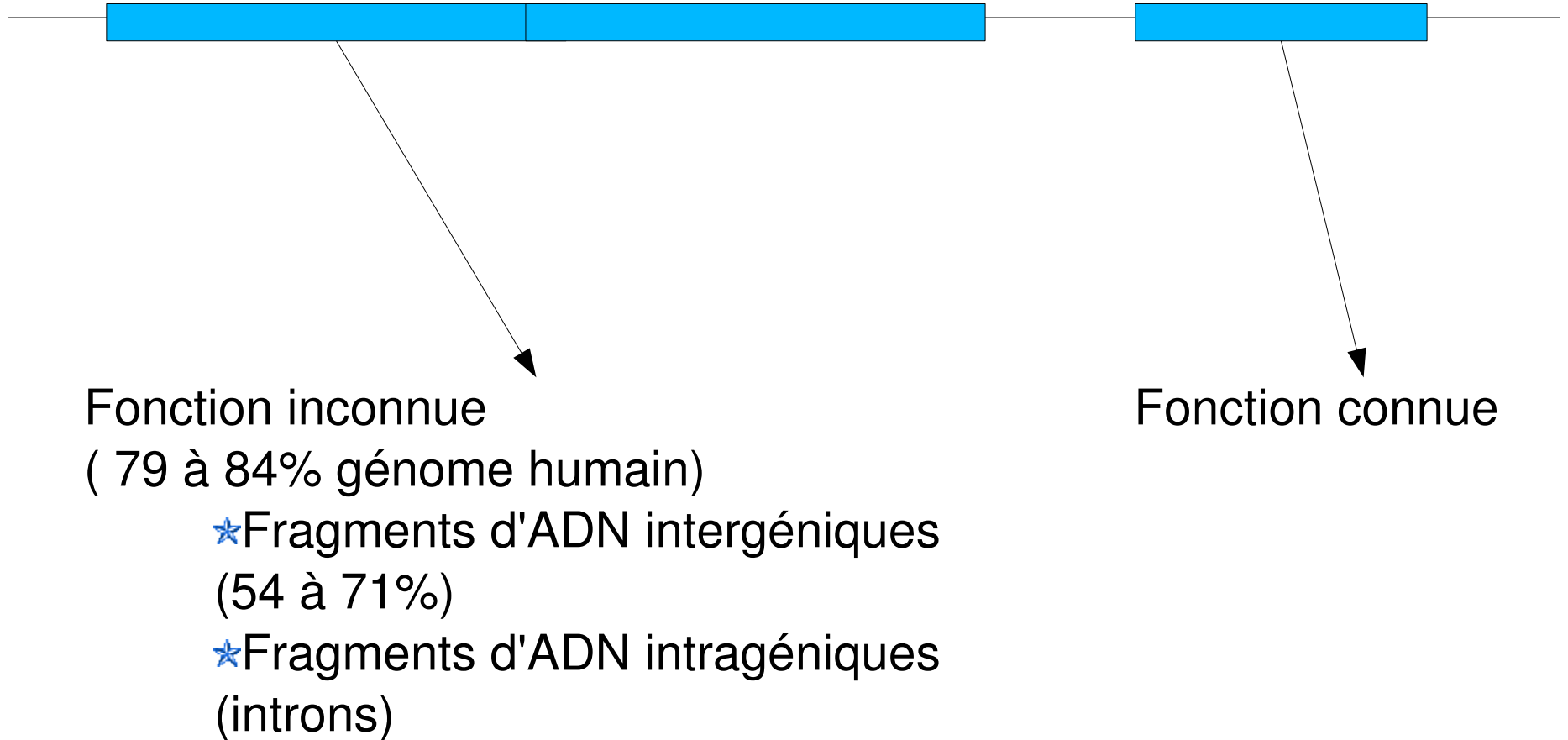
FEATURES Location/Qualifiers
 source 1..1794
 /organism="Streptomyces griseus"
 -10_signal 338..343
 mRNA 350..>1288
 /note="mRNA"
 CDS 434..1288
 /gene="kan"
 /codon_start=1
 /product="aminoglycoside 3-N-acetyltransferase"
 /db_xref="PID:g217014"
 /transl_table=11

Eucaryotes

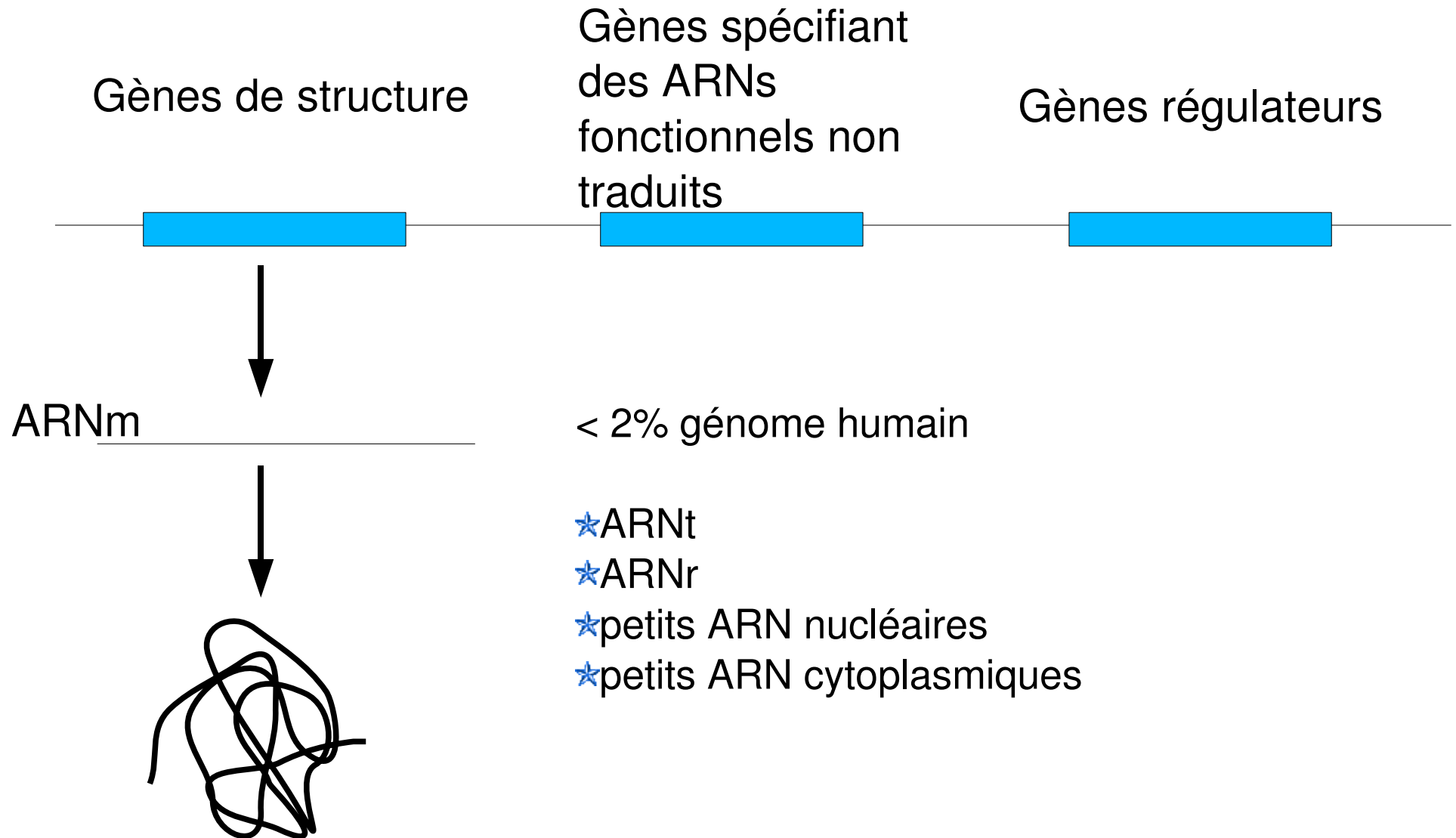


Source: fr.wikipedia.org

Structure du génome



Compartimentation fonctionnelle du génome (2)



Gènes régulateurs

Gènes réplicateurs

- ★ Fonction: spécifier les sites d'initiation et terminaison de la réplication
- ★ Origines de la réplication (1 à 3% du génome humain)

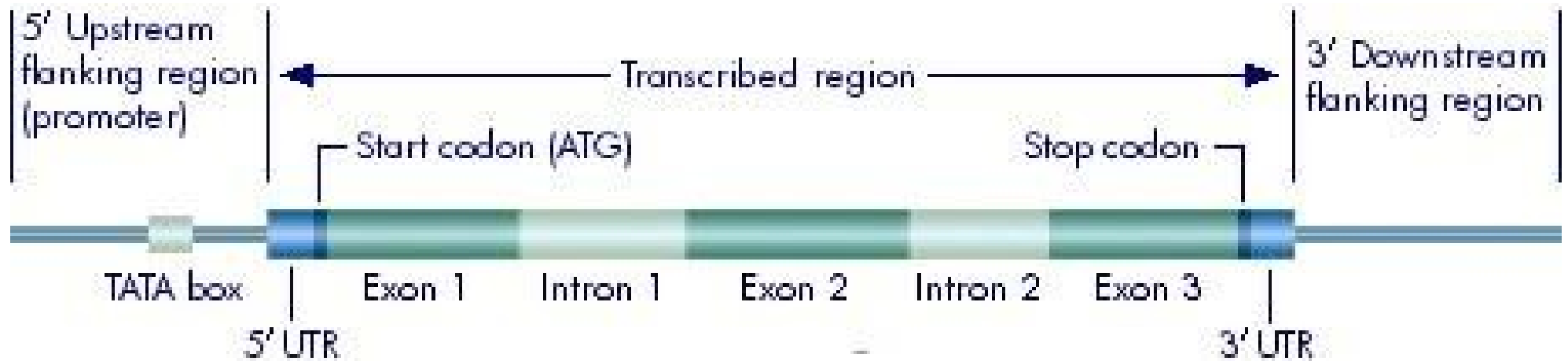
★ Gènes recombinateurs

- ★ Fonction: fournir des sites de reconnaissance pour les enzymes de recombinaison

★ Gènes ségrégateurs

- ★ Fonction: fournir les sites d'attachement pour la ségrégation des chromosomes.
- ★ Les centromères (< 10% du génome humain ?)
- ★ Les télomères (0,01 à 0,1% du génome humain)
- ★ Les SAR (Scaffold Attachment Region) ou MAR (Matrix Attachment Region) pour attachement de l'ADN au squelette chromosomal (~1% du génome humain)

Gènes protéiques : structure (1)



- ★ Régions de régulation

 - ★ promoteurs

 - ★ enhancers et silencers

- ★ 1er exon : codon start

- ★ dernier exon : codon stop

- ★ NCR : Non Coding Region

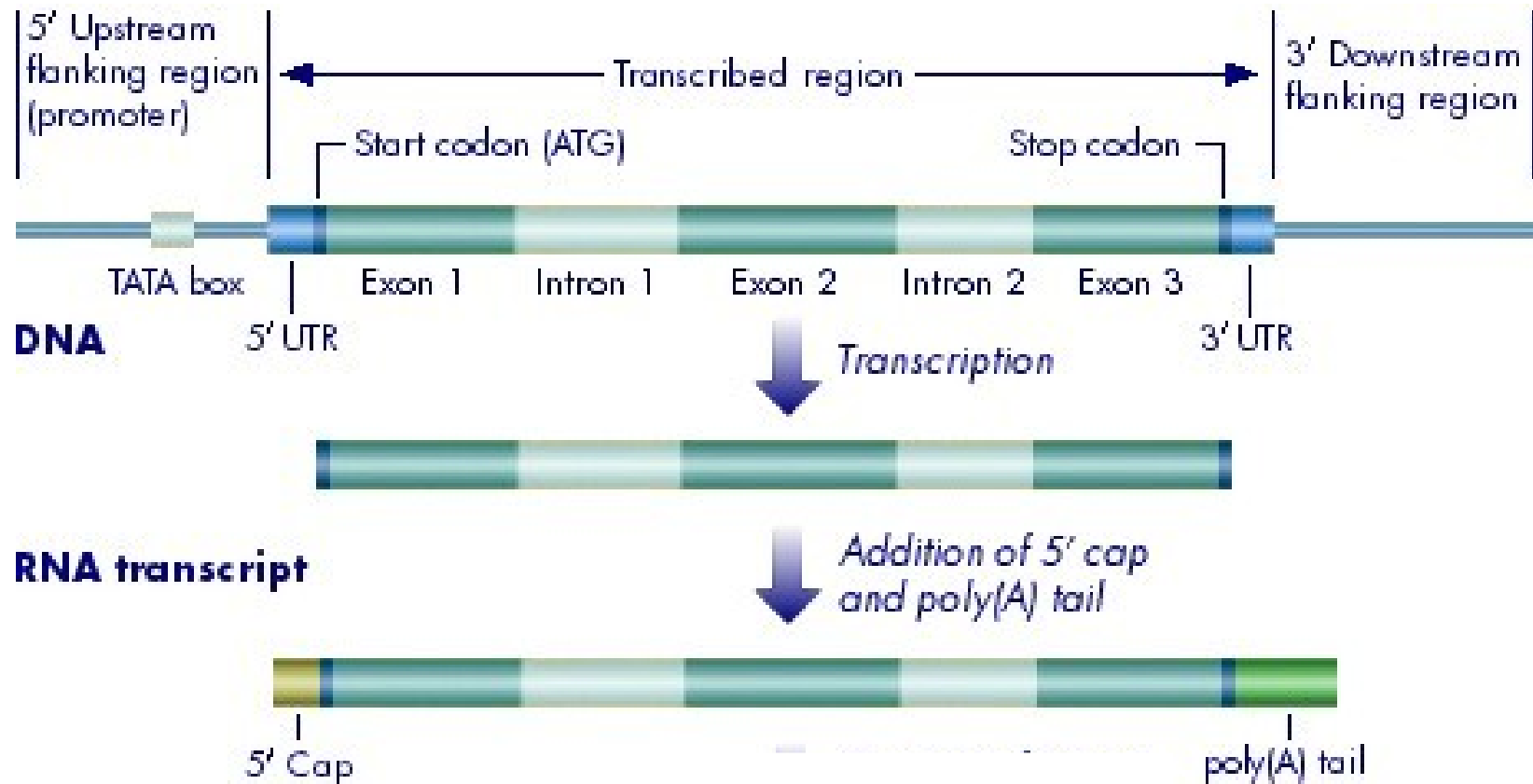
- ★ FLR : Flanking Region

- ★ UTR: Untranslated Region

- ★ CDS : Coding Sequence

Chez les
Eucaryotes

Gènes protéiques: structure de l'ARNm mûre (2)



- ★ Lecture du signal de fin de gène sur ADN
 - ★ Signal de polyadénylation AATAAA ---> queue polyA
 - ★ Région riche en AT
- ★ Transcription intégrale des exons et introns : transcrit primaire (pré-mRNA)
 - ★ Addition d'un 5'cap
 - Addition d'un 3'polyA

Gènes protéiques: structure de l'ARNm mûre (2)

- ★ Maturation du précurseur par excision-épissage
 - ★ Jonction exon-intron:
 - ★ Les introns commencent par GU (site donneur d'épissage)
 - ★ Les introns se terminent par AG (site accepteur d'épissage)
 - ★ A -30 nucléotides de AG, AMP dit A du branchement

RNA transcript



mRNA



Gènes protéiques : structure (2)

Régulation de la transcription chez les Eucaryotes

★ Autres séquences d'ADN situées en 5' du gène : éléments cis-régulateurs

★ TATAAA

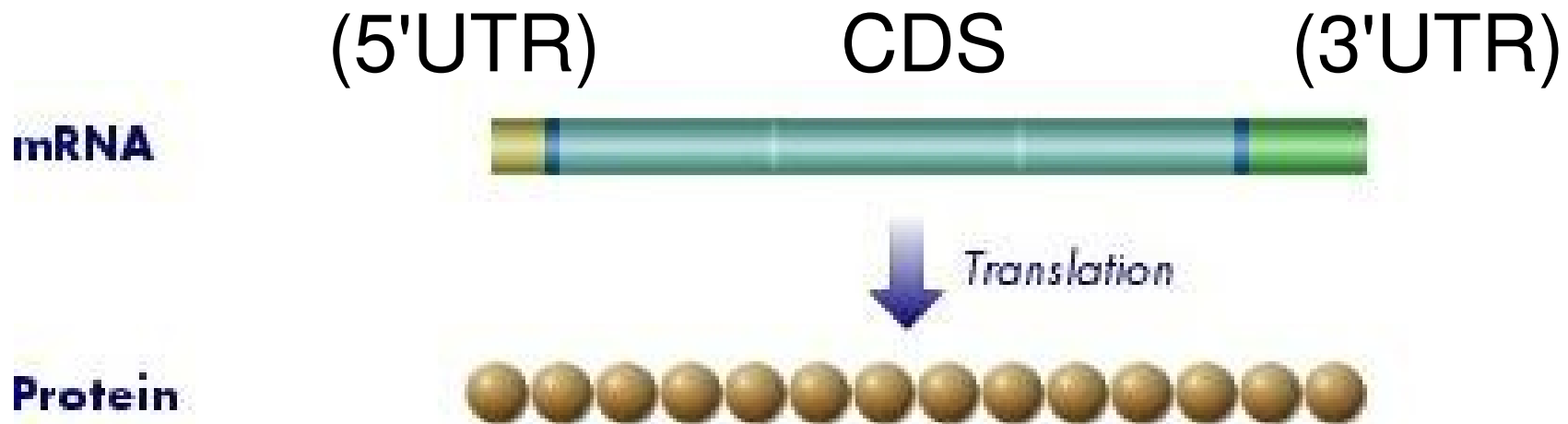
★ GGGCGG

★ CCAAT

★ ATGCAAAT des immunoglobulines

★

Gènes protéiques: structure de l'ARNm mûre (3)



Prédiction des ORFs: méthodes (4)

Recherche des signaux

- ★ Complémentaire de la recherche des ORFs longues
- ★ Recherche des signaux nécessaires à la traduction. En particulier, le site de fixation du ribosome (RBS)
- ★ Chez les eucaryotes, les séquences particulières séparant intron/exon (jonctions intron-exon et exon-intron)
- ★ Motifs ni universels, ni spécifiques
- ★ Systèmes de prédictions sont des programmes à apprentissage développés pour un organisme donné
 - ★ Modèles de Markov (approche probabiliste)
 - ★ Réseaux neuronaux
 - ★ GENSCAN, GAIL (pour le génome humain)

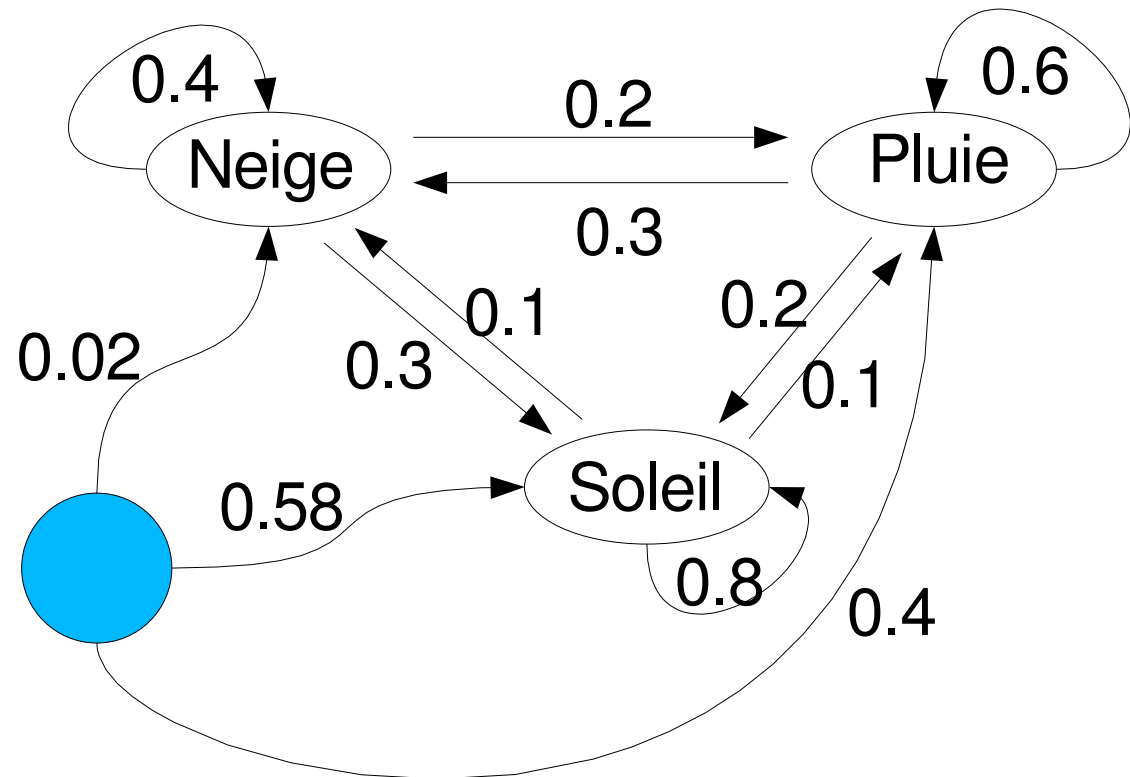
Prédiction des ORFs: chaînes de Markov

	Neige	Pluie	Soleil
Neige	0.4	0.3	0.3
Pluie	0.2	0.6	0.2
Soleil	0.1	0.1	0.8

Probabilité qu'il fasse 'Soleil' les 5 prochains jours:

$$P = 0.58 * 0.8 * 0.8 * 0.8 * 0.8 * 0.8$$

$$P = 0.2375$$



Prédiction des ORFs: chaînes de Markov

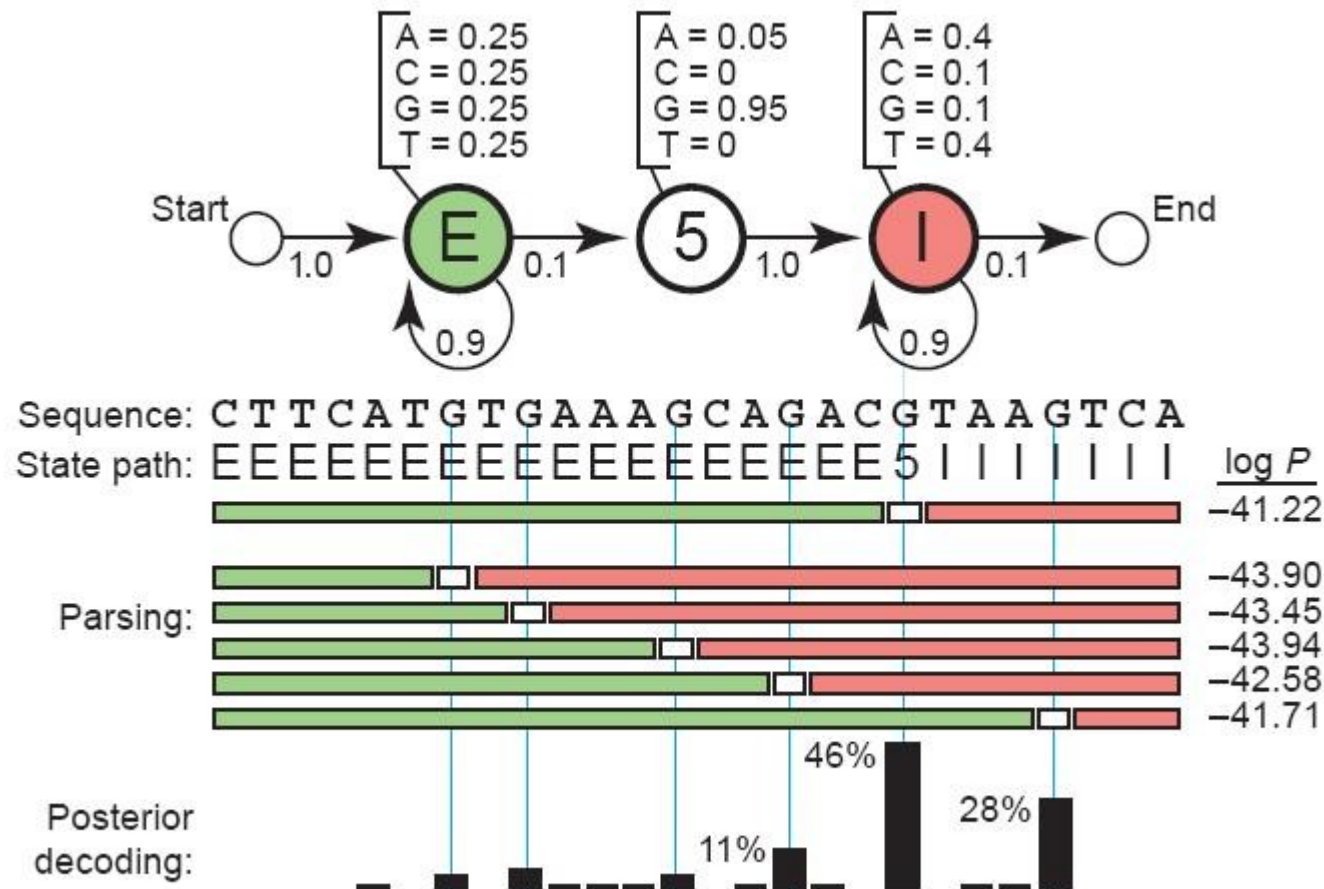


Figure 1 A toy HMM for 5' splice site recognition. See text for explanation.

Prédiction des ORFs: méthodes (5)

Confirmation expérimentale

- ★ Isoler et séquencer les ARNm puis les comparer aux banques de données de séquences codantes
- ★ Identification des gènes si séquences homologues



ORF Finder (Open Reading Frame Finder)

PubMed

Entrez

BLAST

OMIM

Taxonomy

Structure

NCBI

Tools
for data mining

GenBank
sequence
submission support
and software

FTP site
download data and
software

The ORF Finder (Open Reading Frame Finder) is a graphical analysis tool which finds all open reading frames of a selectable minimum size in a user's sequence or in a sequence already in the database. This tool identifies all open reading frames using the standard or alternative genetic codes. The deduced amino acid sequence can be saved in various formats and searched against the sequence database using the WWW BLAST server. The ORF Finder should be helpful in preparing complete and accurate sequence submissions. It is also packaged with the Sequin sequence submission software.

Enter GI or ACCESSION
or sequence in FASTA format

OrfFind

Clear

FROM: **TO:**

[Genetic codes](#)

1 Standard



New GENSCAN Web Server at MIT - Mozilla Firefox

FichierEditionAffichageAller àMarque-pagesOutils?

←

→

↻

✖

🏠

http://genes.mit.edu/GENSCAN.html

OK

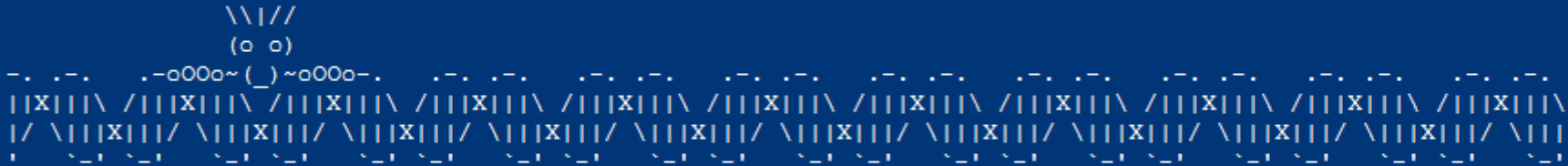
🔍pubmed

HotmailPersonnaliser les liensWindows MediaWindowshttp://www.vvf-vac...

New GENSCAN Web Server at MITNCBI Sequence Viewer v2.0cours 1.PDF (objet application/pdf)gene recognition papers (1997)

The New GENSCAN Web Server at MIT

Identification of complete gene structures in genomic DNA



[For information about Genscan, click here](#)

This server provides access to the program Genscan for predicting the locations and exon-intron structures of genes in genomic sequences from a variety of organisms.

This server can accept sequences up to 1 million base pairs (1 Mbp) in length. If you have trouble with the web server or if you have a large number of sequences

✖ Rechercher : genscan

🔍 Occurrence suivante

🔍 Occurrence précédente

🔍 Surligner

☐ Respecter la casse

⚠ Phrase non trouvée

Terminé

🟢

☁

☀

☀

GeneMark™ - Gene prediction software programs - Mozilla Firefox

Fichier Edition Affichage Aller à Marque-pages Outils ?

← → ↺ × 🏠 🐝

http://exon.gatech.edu/GeneMark/

OK genemark

Hotmail Personnaliser les liens Windows Media Windows http://www.vvf-vac...

GeneMark™ - Gene prediction softw... NCBI Sequence Viewer v2.0 cours 1.PDF (objet application/pdf) gene recognition papers (1997)

GeneMark™

A family of gene prediction programs developed by Mark Borodovsky's [Bioinformatics Group](#) at the [Georgia Institute of Technology](#), Atlanta, Georgia.

What's New:

Gene identification in novel eukaryotic genomes by self-training algorithm: [GeneMark.hmm-ES](#)

Supported by NIH



Gene Prediction in Bacteria and Archaea



For bacterial and archaeal gene prediction you can use the parallel combination of [GeneMark-P](#) and [GeneMark.hmm-P](#). For a novel genome you can use either the [Heuristic models](#) option (if the sequence is shorter than 1 Mb) or the self-training program [GeneMarkS](#) (aka GeneMark.hmm-PS).

Gene Prediction in Eukaryotes



For eukaryotic gene prediction you can use the parallel combination of [GeneMark-E](#) and [GeneMark.hmm-E](#). For a novel genome (the one whose name is not in the list of available models) you can run [GeneMark.hmm-ES](#), the self-training program (just 10MB sequence is needed for training).

Gene Prediction in Viruses

Powered by IBM



Borodovsky Group

Gene Prediction Programs

- ♦ [GeneMark](#)
- ♦ [GeneMark.hmm](#)
- ♦ [GeneMarkS](#)
- ♦ [Heuristic models](#)
- ♦ [Frame-by-Frame](#)

Information

- ♦ [Background](#)
- ♦ [References](#)
- ♦ [In GenBank](#)
- ♦ [FAQ](#)
- ♦ [Contact](#)

Databases of predicted genes

Rechercher : genscan

☒ Occurrence suivante ☐ Occurrence précédente

☐ Surligner ☐ Respecter la casse

Phrase non trouvée

Terminé



Applications

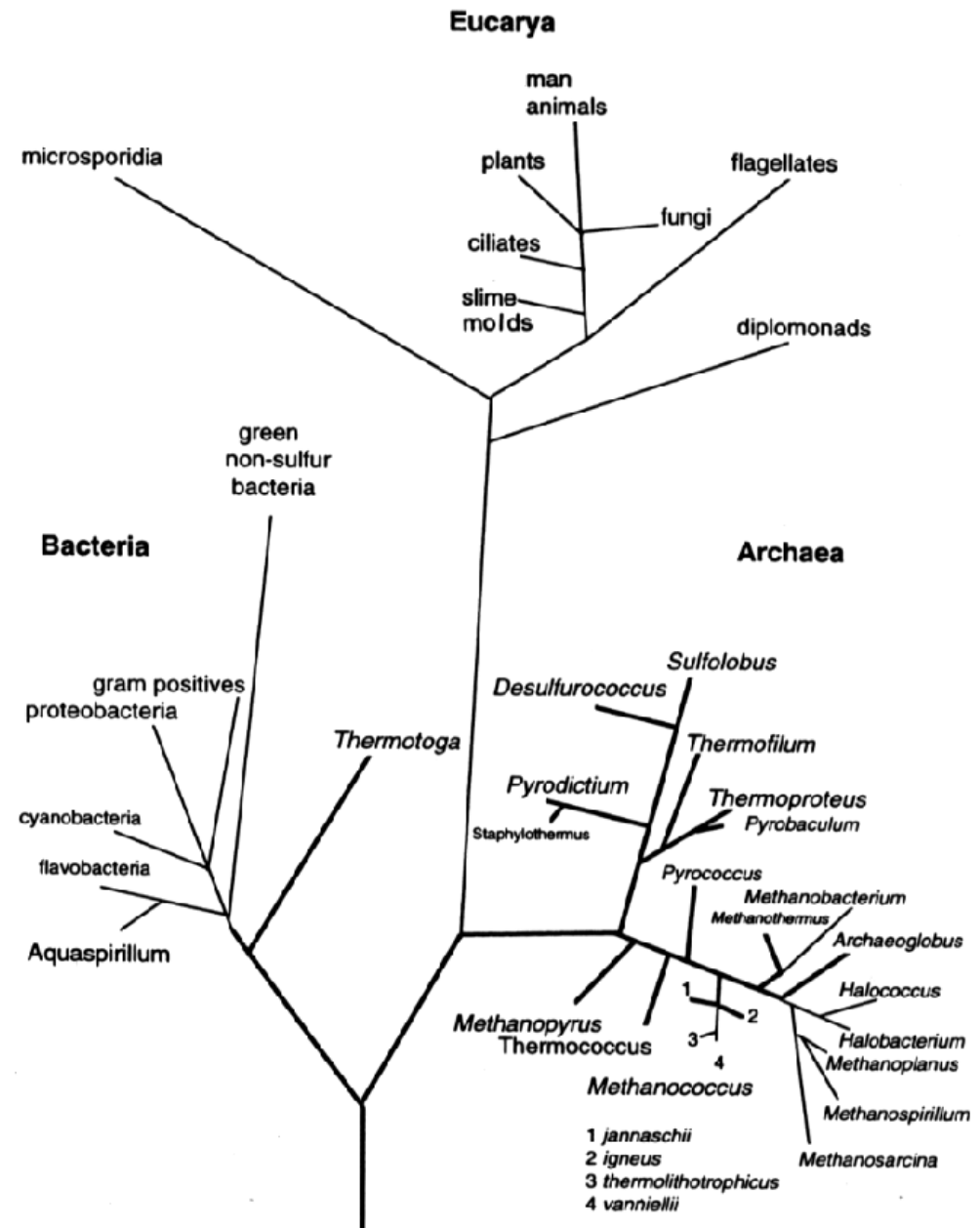


Phylogénie

Evolution

- Mutation
- Fixation

au niveau ADN



Fonction d'un gène

- ★Fonction moléculaire

- ★Fonction cellulaire

- ★Fonction dans le développement

- ★Fonction d'adaptation

- ★Produit d'un même gène <-> plusieurs fonctions moléculaires

- ★Un gène peut coder +sieurs produits différents à fonctions différentes:

 - ★Modifications des produits de transcription (épissage alternatif)

 - ★Modifications post-traductionnelles

Détermination de la fonction d'un gène

Approche bioinformatique

- ★ Par similarité de séquence
- ★ Par similarité de structure du produit du gène
- ★ Connaissance de la structure (et identification) du gène ne donne pas généralement accès à la fonction.

Approche expérimentale

- ★ Par l'étude de l'expression du gène dans l'espace et dans le temps
- ★ Par la localisation subcellulaire et les modifications post-traductionnelles de la protéine
- ★ Par observation des effets de l'altération ou de la délétion du gène dans l'organisme (technique de *knock-out* ou surexpression d'un gène)

