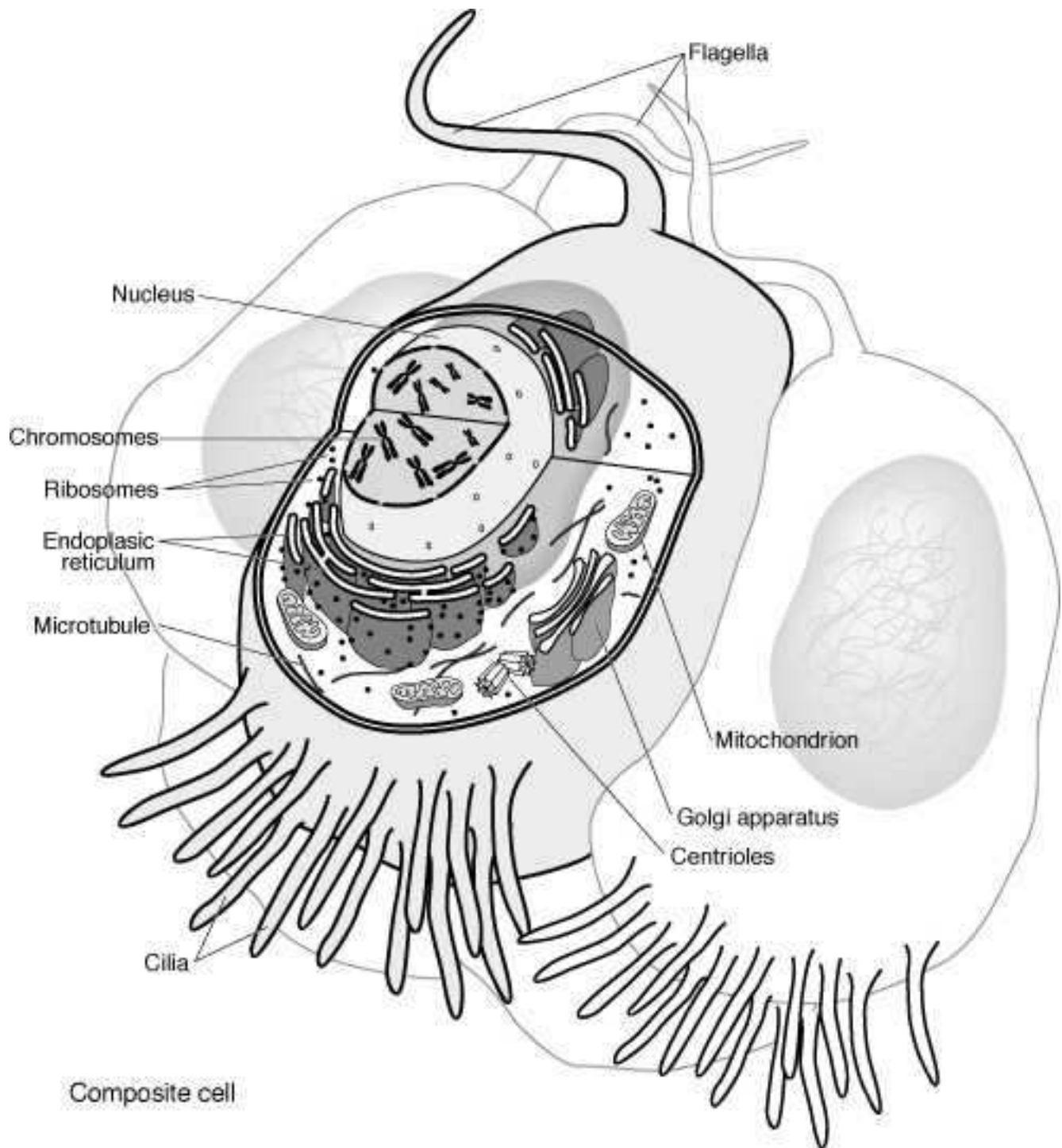


Cours 1

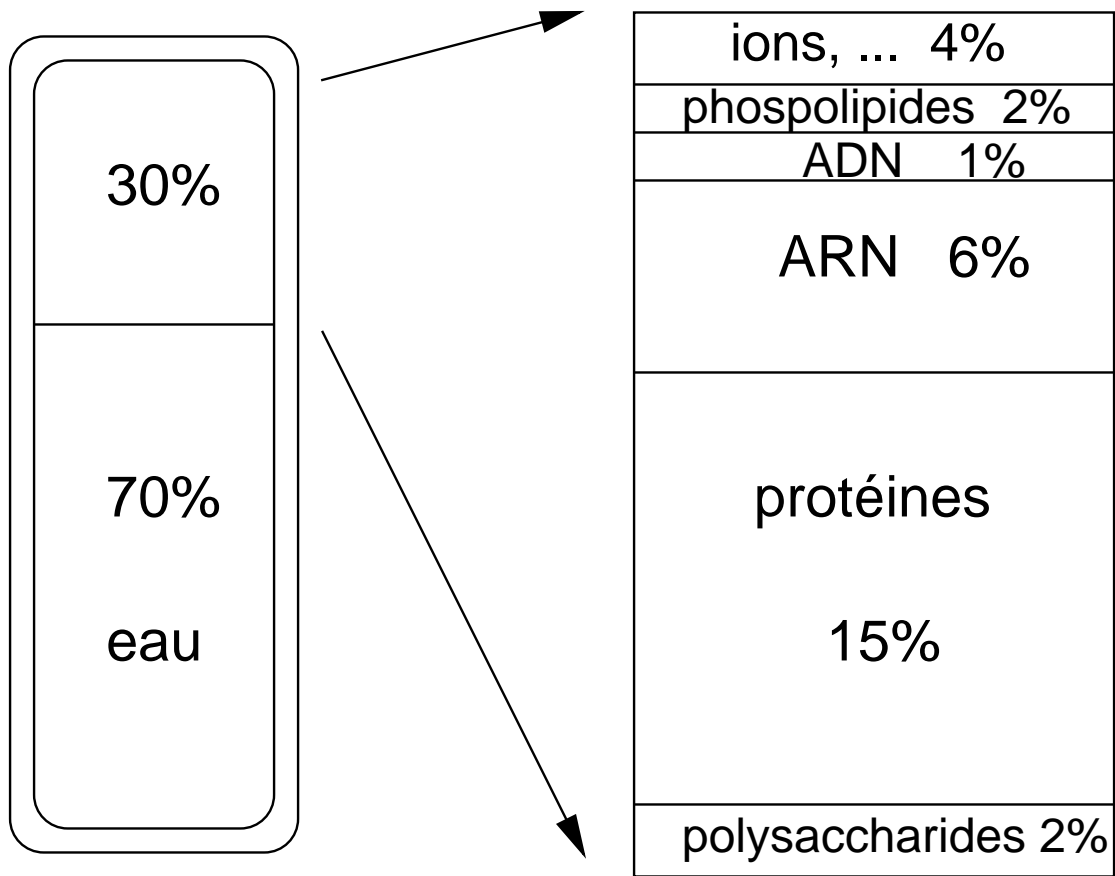
Comment tout savoir sur la biologie
moléculaire en moins de 90 minutes ?

... quitte à faire quelques simplifications

La cellule



Que trouve-t-on dans une cellule ?



Les protéines

- ▷ 3000 à 4000 protéines dans une cellule

- ▷ **les protéines sont partout :**
 - composant des muscles
 - transport: hémoglobine (oxygène), albumine (corps gras)...
 - régulation: insuline, ...
 - récepteur
 - anticorps
 - structure: collagène dans la peau, kératine dans les cheveux ...
 - ...

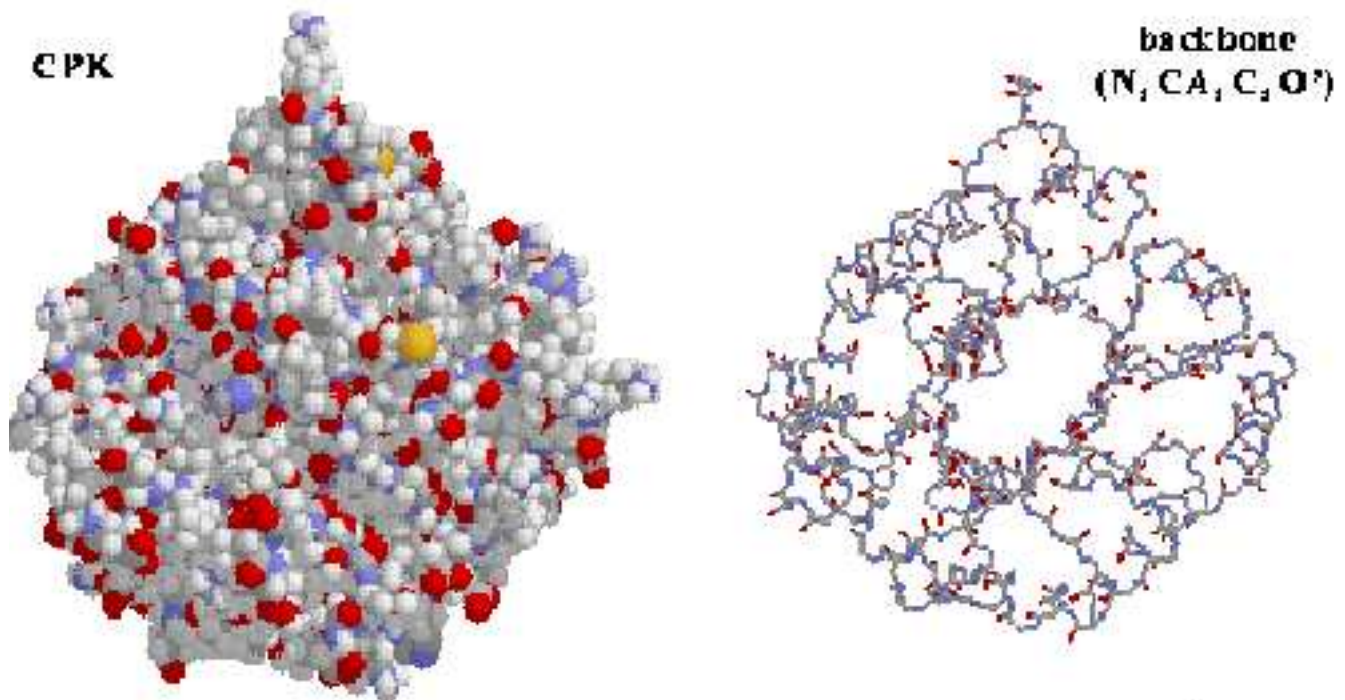
▷ 20 acides aminés distincts

A C D E F G H I K L M N P Q R S T V W Y

▷ longueur : 50 à 1000 acides aminés

▷ une protéine se replie "en pelote", adoptant une configuration spatiale caractéristique de sa fonction

triosc phosphate isomerase (TIM)



Qui fournit le *plan de construction*
des protéines ?

Acide Désoxyribonucléique

- ▷ Support matériel de l'hérédité
- ▷ Composé de quatre **bases** (ou **nucléotides**)

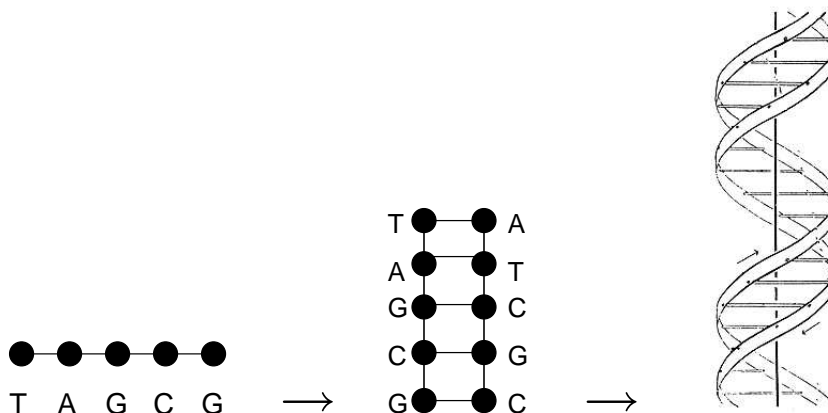
A : adénine
C : cytosine
G : guanine
T : thymine

- ▷ Un brin d'ADN est orienté

5' → ATACCGTATGCTA → **3'**

- ▷ Structure en double hélice

A ↔ **T**
C ↔ **G**



Ensemble de l'ADN (présent dans une cellule)
= ensemble de l'information génétique d'un organisme.
Chaque cellule contient un exemplaire complet du génome.

Chromosome

Le génome est réparti en petites molécules.

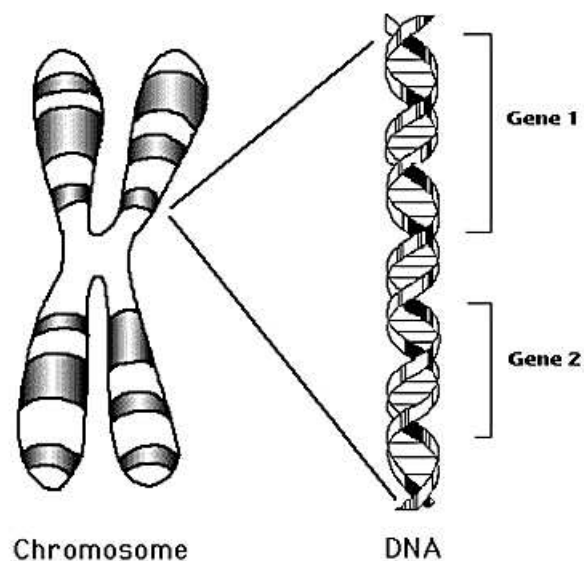
un gène = une fonction

Gène

Portion codante du génome : instruction pour former une protéine, ou un ARN

Tout le génome n'est pas réparti en gènes (environ 5% pour l'homme, par exemple).

Longueur d'un gène : de quelques centaines à un million de nucléotides.



Genes

- 23 paires de chromosomes
- génome : 3 milliards de bases
- gènes : de 30 000 à 35 000

Mus musculus - souris

- 21 paires de chromosomes
- génome : 3 milliards de bases
- gènes : de 30 000 à 35 000

Arabidopsis thaliana - plante des bords de chemins

- 5 paires de chromosomes
- génome : 1,2 milliard de bases
- gènes : environ 20 000

Saccharomyces cerevisiae - levure de bière

- 16 paires de chromosomes
- génome : 130 millions de bases
- gènes : environ 6 000

Escherichia coli - bactérie de l'intestin

- 1 paire de chromosomes
- génome : 46 millions de bases
- gènes : environ 4 000

De l'ADN à la protéine

1 - Transcription : ADN \rightarrow ARN (A, C, G, U)

Un gène est transcrit dans le sens 5' à 3' en une molécule d'**ARN messenger**, avec la complémentarité A \leftrightarrow U, C \leftrightarrow G.

2 - Maturation de l'ARN

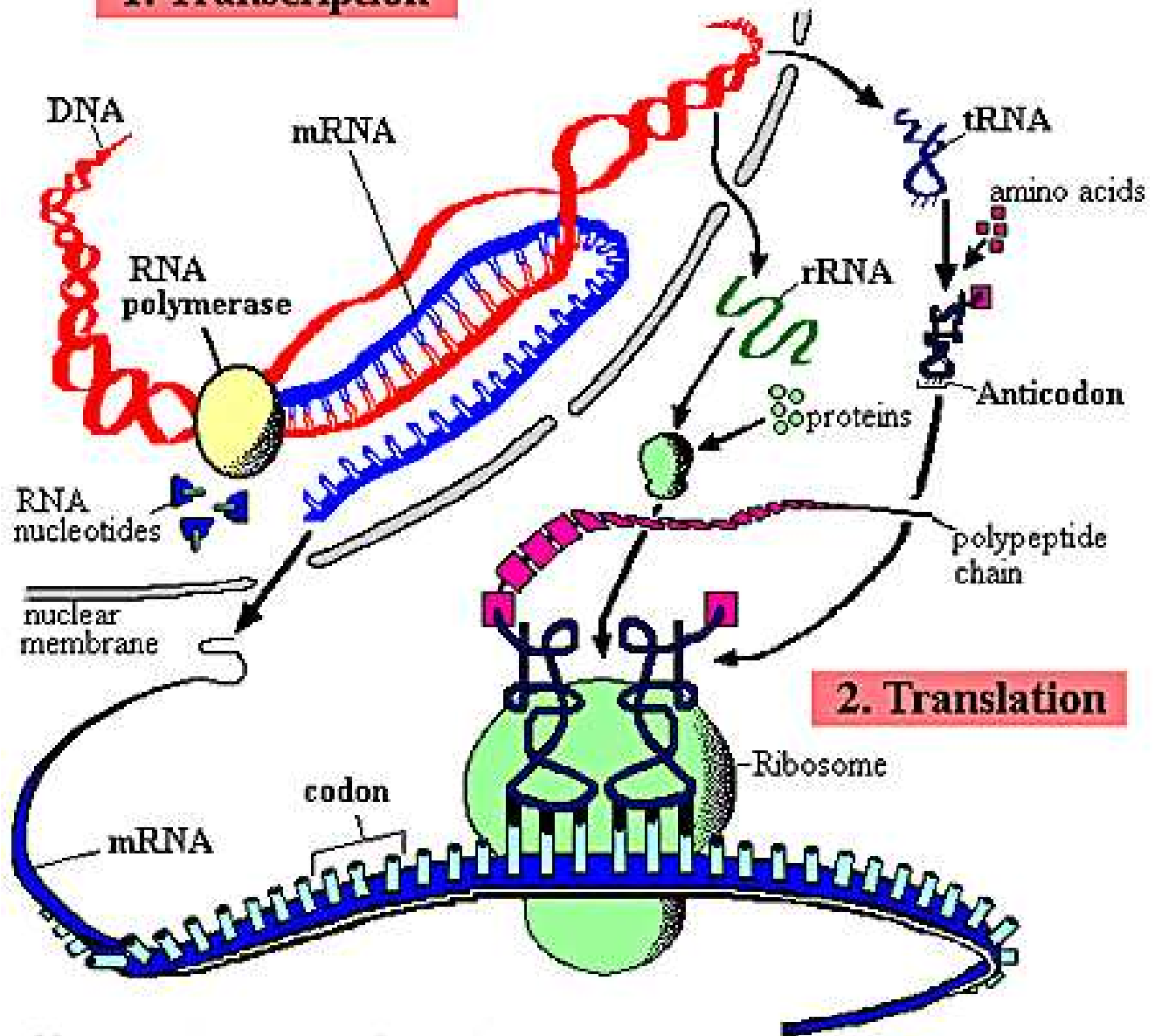
3 - Traduction : ARN \rightarrow protéine

Chaque triplet de nucléotides de l'ARN correspond à un acide aminé. Cette traduction est quasiment universelle. C'est le **code génétique**.



\rightarrow six phases de lectures

1. Transcription



Protein synthesis

Code génétique

		2ème						
		U	C	A	G			
1er						3ème		
U		F	S	Y	C	U		
		F	S	Y	C		C	
		L	S	!	!		A	
		L	S	!	W		G	
C		L	P	H	R	U		
		L	P	H	R		C	
		L	P	Q	R		A	
		L	P	Q	R		G	
A		I	T	N	S	U		
		I	T	N	S		C	
		I	T	K	R		A	
		M	T	K	R		G	
G		V	A	D	G	U		
		V	A	D	G		C	
		V	A	E	G		A	
		V	A	E	G		G	

! : codon stop

La bioinformatique

- ▷ Obtention de la séquence ADN : **séquençage**

Organismes séquencés, ou en cours de séquençage : plus de 20 espèces bactériennes, la levure, l'homme, la souris, le riz, le maïs, etc.

- ▷ Stockage des informations

Banques de données : Genbank (ADN), Swissprot (protéines), ...

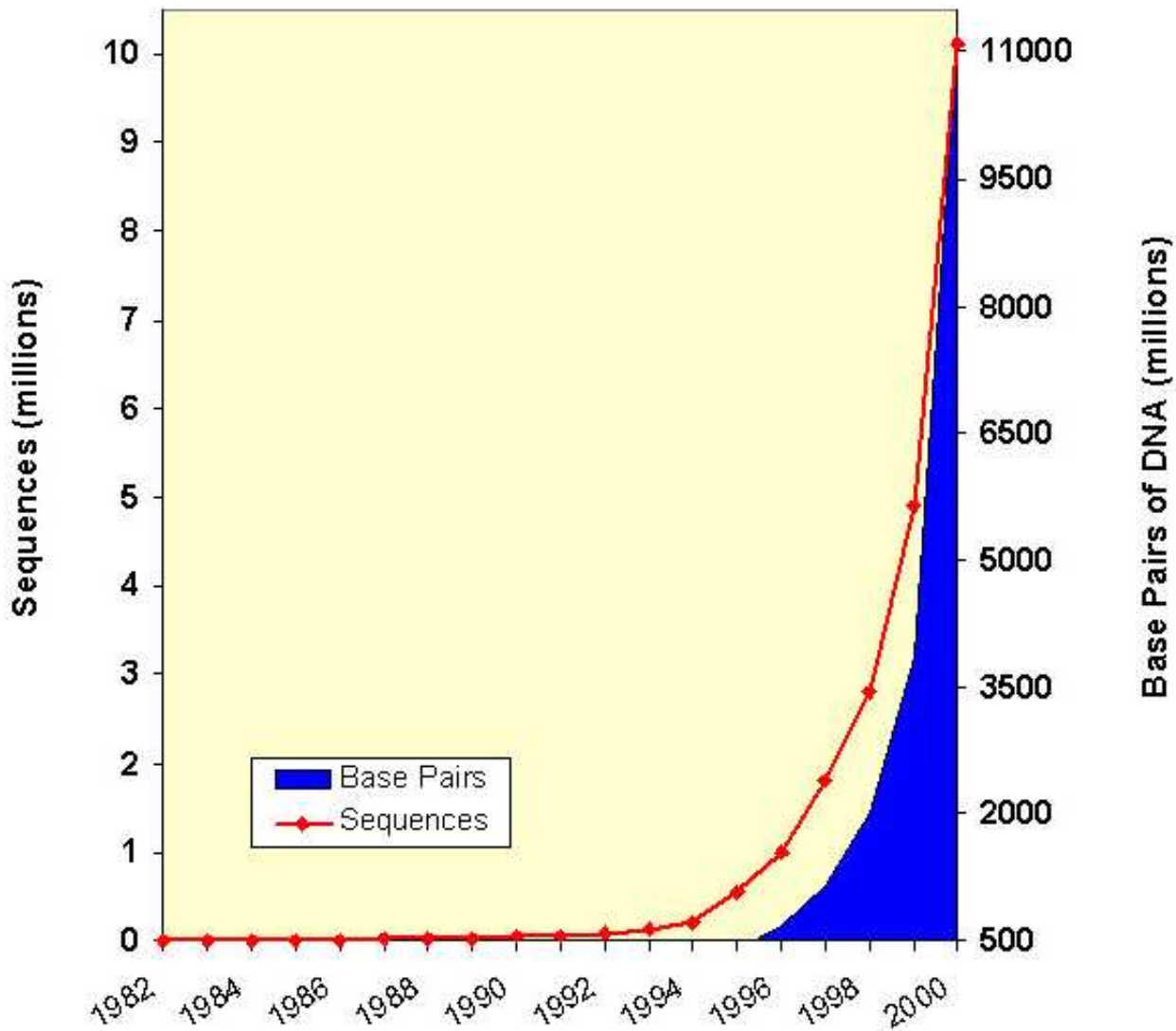
- ▷ Analyse de la séquence

de la séquence brute à la fonction

Repères historiques

- 1865** : Mendel et la théorie de l'hérédité
- 1944** : l'ADN est le support de l'information génétique, et non les protéines (Avery)
- 1951** : première protéine séquencée, insuline (Sanger)
- 1953** : structure en double hélice de l'ADN (Watson-Crick)
- 1967** : algorithme de reconstruction d'arbre phylogénétique
- 1970** : algorithme pour la comparaison de séquences protéiques (Needleman-Wunsh)
- 1974** : algorithme de prédiction de structure secondaire de protéines (Chou-Fasman)
- 1977** : séquençage d'ADN
- 1978** : constitution des premières banques de données
- 1990** : lancement du programme international de séquençage *Génome Humain*
- 1996** : séquence complète de la levure
- 2000** : premier brouillon du génome humain

Croissance des données de Genbank



16 milliards de bases (décembre 2001)

Exemple : recherche d'EST

- EST (Expressed Sequence Tag) :

Fragment d'ADN de quelques centaines de nucléotides correspondant à un extrait de gène.

- On peut localiser un gène dans le génome en retrouvant la séquence de l'EST.
- Généralisation : itérer le procédé en cherchant tous les EST connus, stockés dans une banque.

Algorithme de recherche de motifs exacts

texte : séquence plusieurs millions
motif : EST plusieurs centaines

Exemple : détermination de domaines conservés

Défensines :

Famille de protéines impliquées dans la défense contre les bactéries, les champignons et les virus à enveloppes.

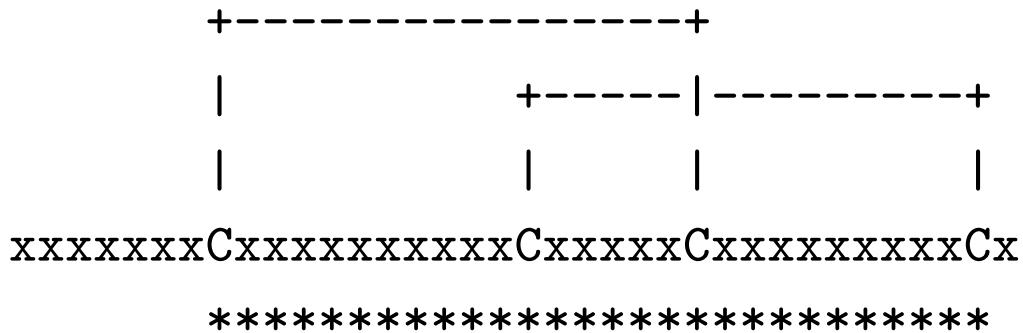
Exemples de mammifères

D1_MOUSE/64-92	CYCRSR.GCKGRERMNGTCRKGHLLYTLCC
D5_MOUSE/64-92	CYCRIR.GCKRRERVFGTCRNLFVTFVFC
D5_HUMAN/65-93	CYCRTG.RCATRESLSGVCEISGRLYRLCC
D1_RABIT/62-90	CACRRR.FCPNSERFSGYCRVNGARYVVRCC
D5_RABIT/65-93	CTCRRF.SCGFGERASGSCTVNGVRHTLCC
D6_HUMAN/72-99	CHCRR..SCYSTEYSYGTCTVMGINHRFCC
D2_RABIT/3-32	CVCRKQLLCSYRERRIGDCKIRGVRFPFCC
D3_RABIT/65-93	CACRRA.LCLPRERRAGFCRIRGRIHPLCC

Inférence de motifs

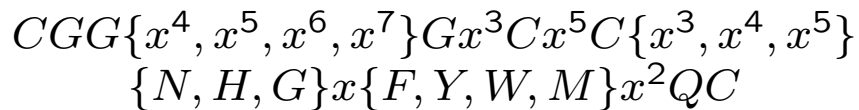
Exemple : *recherche de domaine dans une séquence protéique*

- Site de fixation de la cellulose



Les quatre cystéines sont impliquées dans des ponts di-sulfures.

- Description du motif

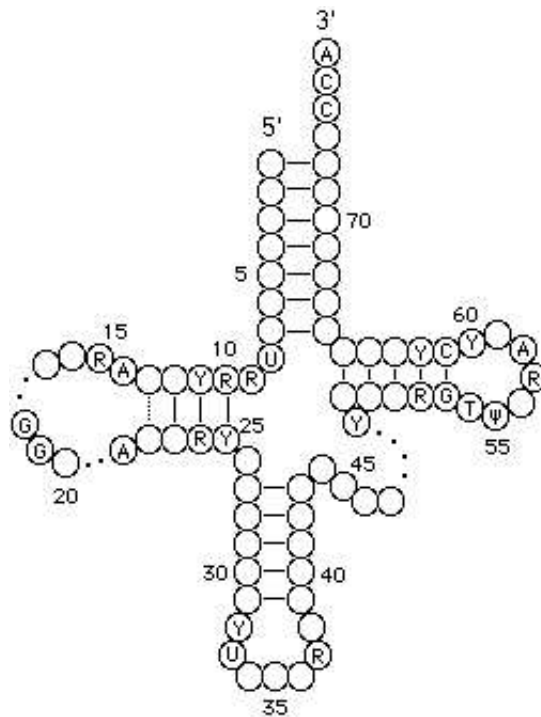


x : n'importe quel acide aminé

*Localisation
d'expressions régulières*

Exemple : *structure de l'ARN*

- L'ARN est une molécule simple brin qui se replie en formant des tiges.



- Le séquençage ne donne pas d'information sur la localisation de ces tiges.

*Algorithme de recherche de palindromes,
d'optimisation du nombre de palindromes.*

Champs d'application

Que peut-on faire avec des séquences ?

