

Introduction à la biologie moléculaire et à la bio-informatique

Cours de Master Recherche M2, 2004/2005

Jean-Philippe Vert

Jean-Philippe.Vert@mines.org

Plan

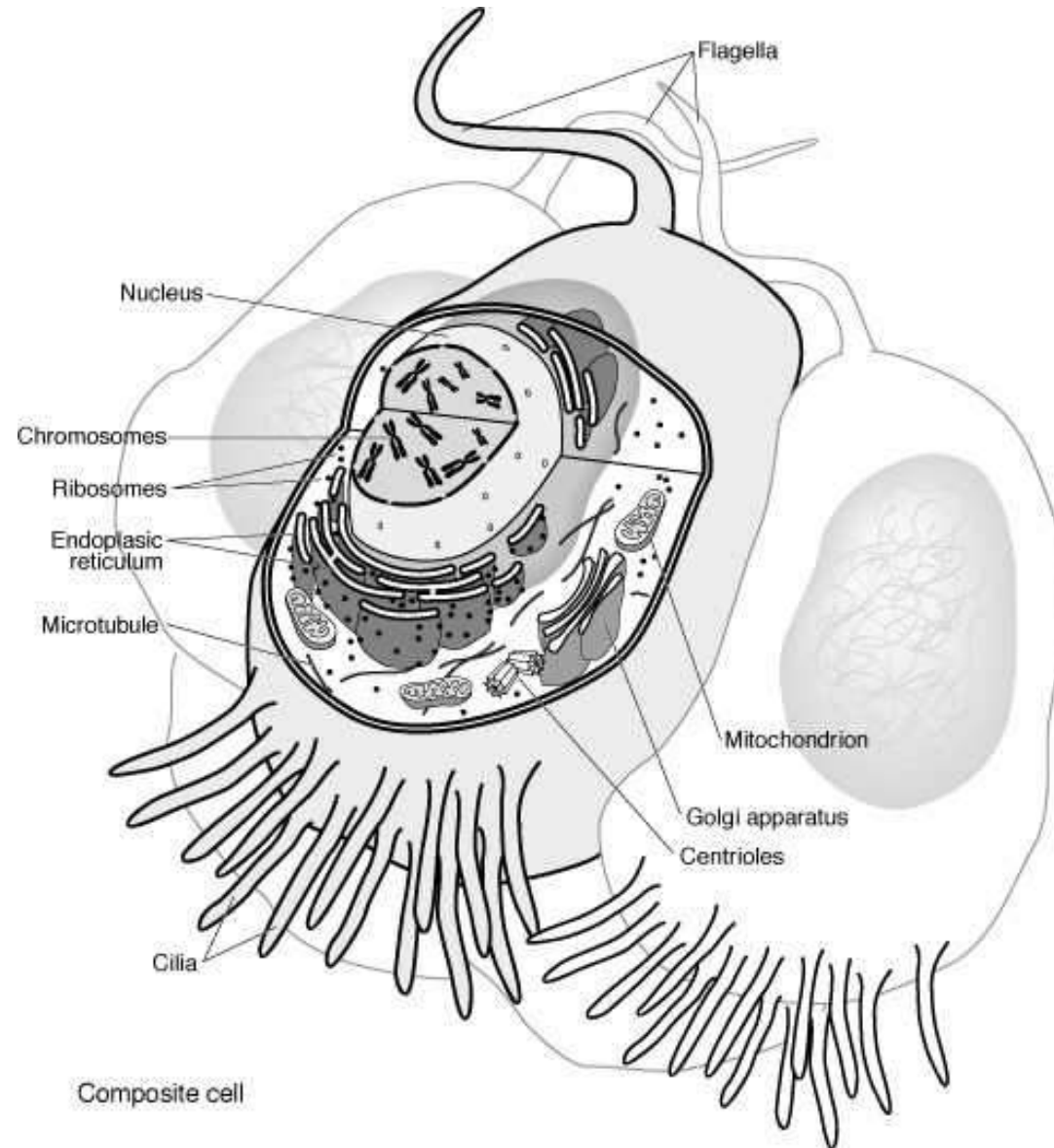
- Organismes et cellules
- Molécules de la vie
- Gènes et génomes
- Technologies et données
- Challenges en bio-informatique

Organismes et cellules

Cellules

- Tout organisme *vivant* est composé de *cellules*
- Une cellule est une solution contenant différentes molécules entourée d'une *membrane*
- Il y a des organismes *unicellulaires* (bactéries, levure...) ou *multicellulaires*.
- Exemple: il y a environ 6×10^{23} cellules dans un humain, de 320 types différents (peau, muscles, neurones...)

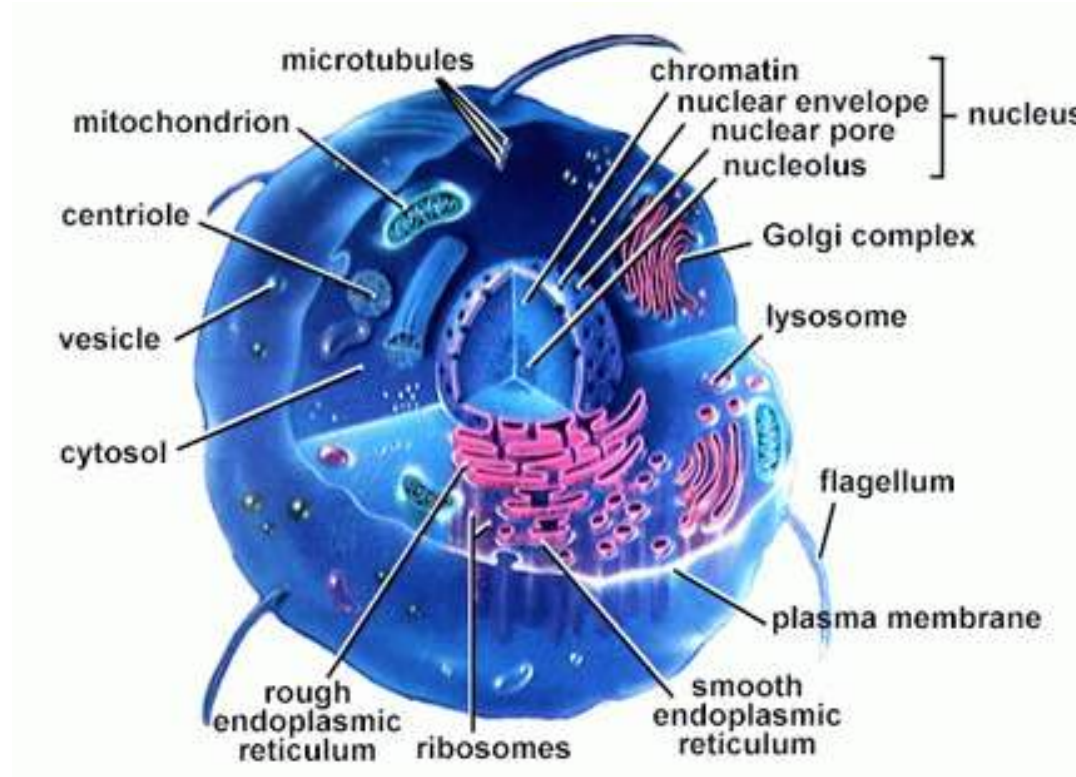
Cellules



Classification des organismes

- On distingue généralement les *eukaryotes* des *prokaryotes*
- Les prokaryotes (eux-mêmes subdivisés en *bactéries* et *archéens*) sont unicellulaires, de petite taille (typiquement $1\mu m$), et ont une structure simple
- Les eukaryotes sont uni- ou multicellulaires, plus grands, et ont une structure plus complexes
- La vie est apparue il y a 3,8 milliards d'années, tous les organismes proviennent d'un ancêtre commun

La cellule eukaryote



Différents organelles. Un noyau qui contient l'ADN (chromosomes).

Caractéristiques de la cellule

- La plupart des cellules sont capables de *grossir* et de se *diviser* (exception: neurones)
- Elles ont un *métabolisme*, i.e., importent des nutriments et les convertissent en molécules utiles et énergie
- Elles peuvent réagir à leur environnement

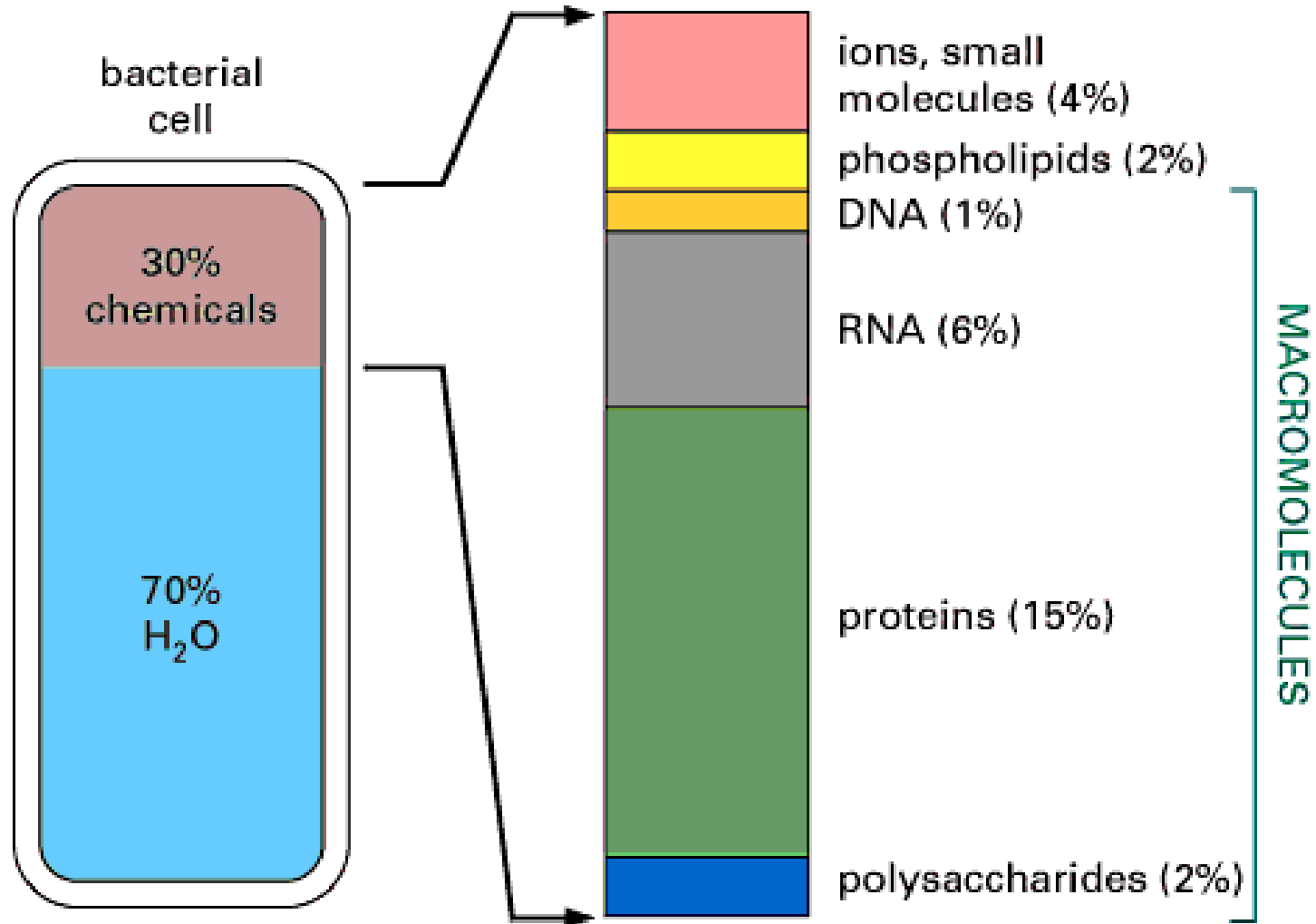
Les molécules de la vie

Types de molécules

On les regroupe en 4 grandes familles:

- les petite molécules
- les protéines
- l'ADN
- l'ARN

Dans la cellule



Petites molécules

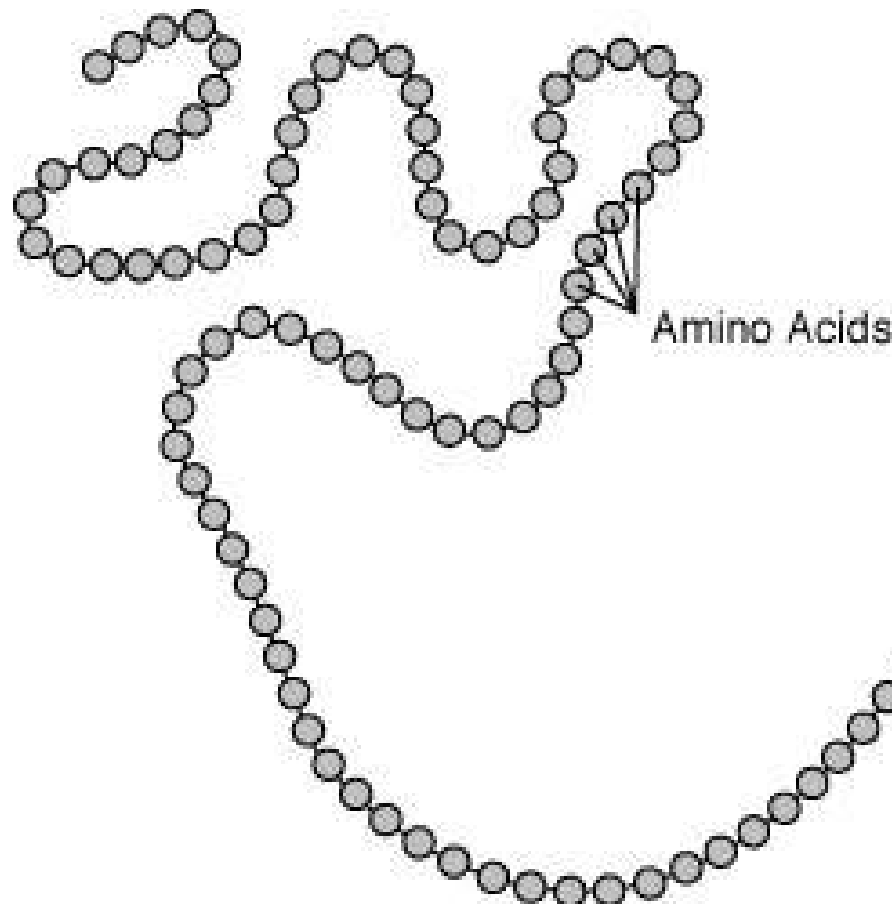
- Petites molécules ayant un rôle: *ATP*, *NADPH* stockent l'énergie
- Sucres, lipides (sources d'énergie, structure des membranes)
- *Acides aminés* et *nucléotides*, qui sont les blocs de base pour former les *protéines* et l'*ADN/ARN*.

Protéines

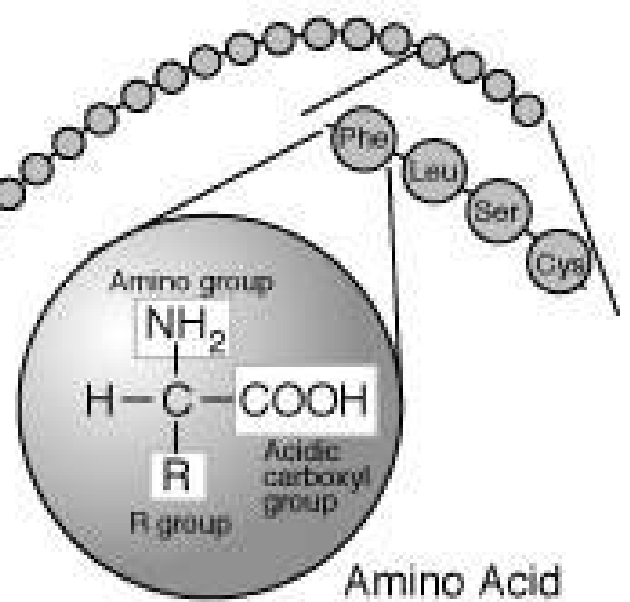
Les protéines représentent 20% du poids de la cellule (eau=70%). Elles ont de *multiples fonctions*:

- *Structurale* : ex: le collagène relie les os et les tissus
- *Catalytique* : les *enzymes* catalysent une multitude de réactions biochimique (formant le métabolisme). Ex: la hexokinase permet la conversion du glucose au glucose-6-phosphate
- Les *protéines membranaires* maintiennent l'environnement cellulaire, régulent le volume de la cellule, créent des gradients ioniques pour les muscles et le système nerveux...

Protéine = polymère d'acides aminés



Primary protein structure is sequence of a chain of amino acids



Structure primaire

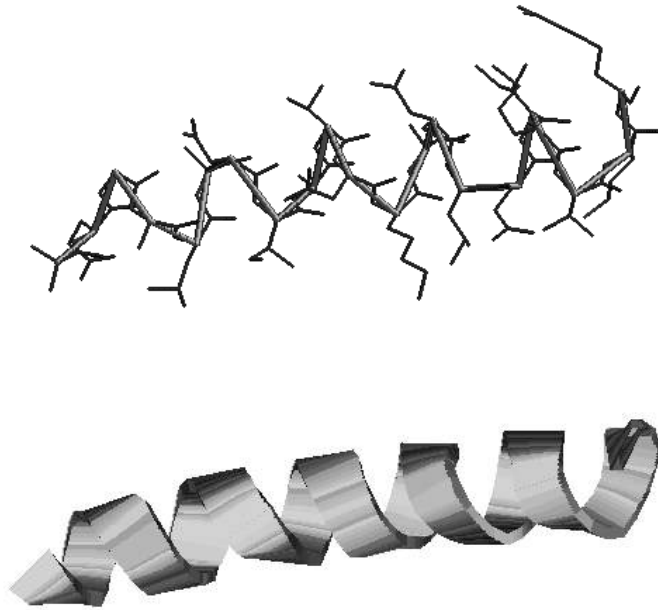
Il y a *20 acides aminés*. On peut donc représenter la structure chimique d'une protéine comme un texte sur un alphabet de 20 lettres.

Exemple: l'insuline:

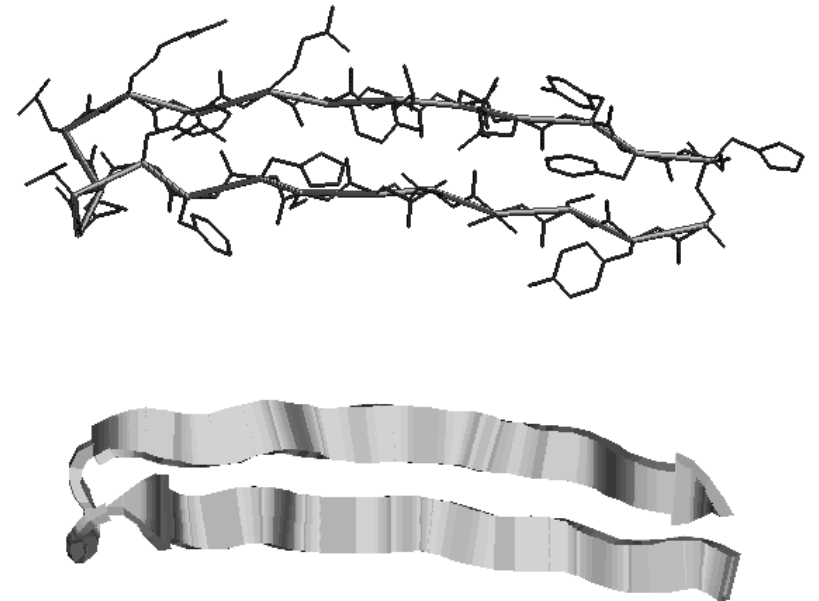
FVNQHLCGSHLVEALYLVCGERGFFYTPKA

Structure secondaire

Hélice α



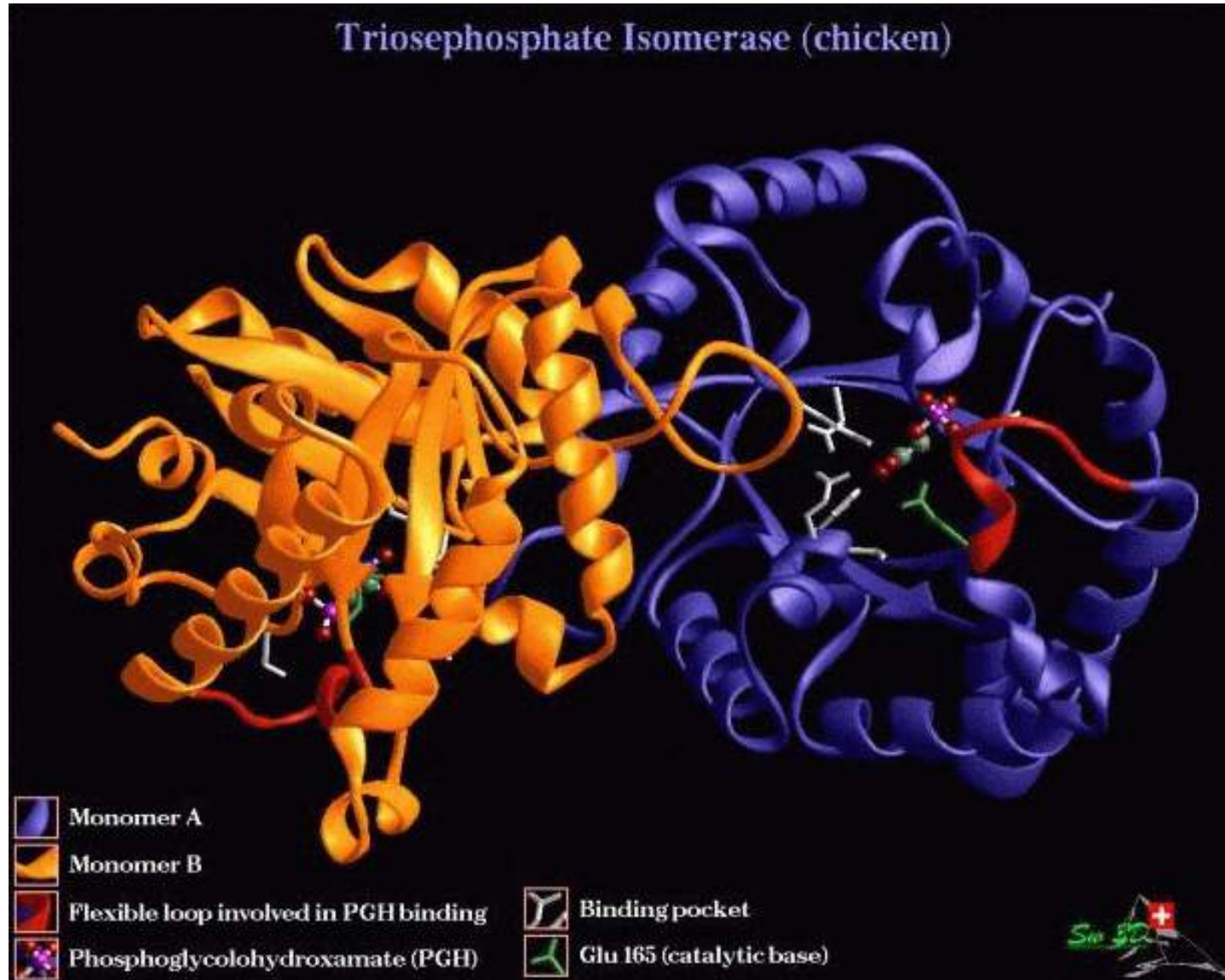
Feuillet β



Structure tertiaire



Structure quaternaire

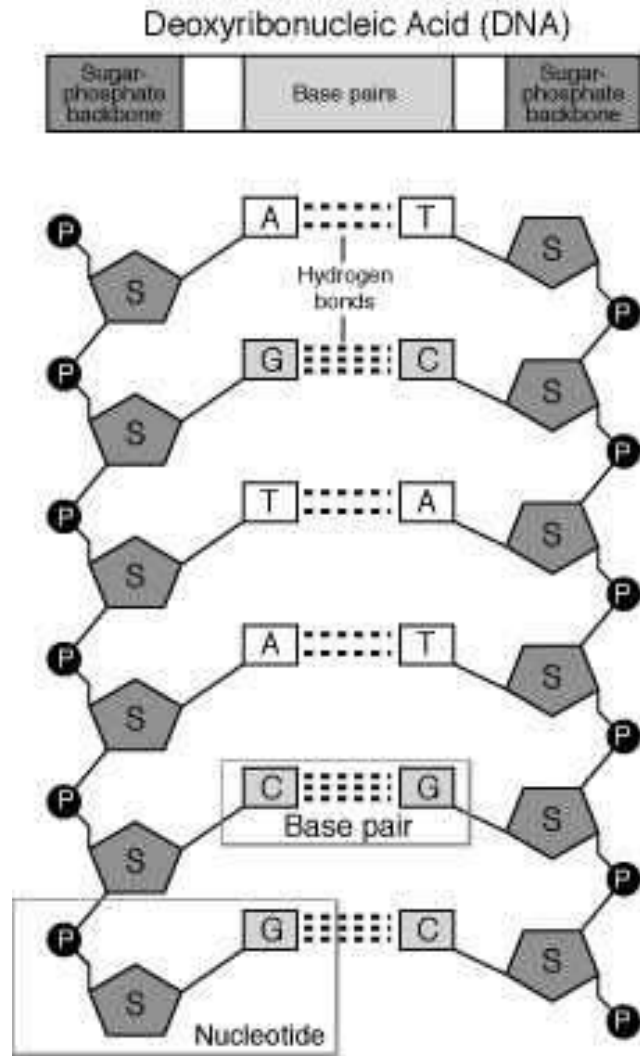


ADN

- L'*acide desoxyribonucléique* (ADN) est la molécule, présente dans toutes les cellules, qui contient l'information génétique transmise entre générations.
- L'ADN peut être en *simple brin* ou *double brin*.
- Un brin simple (aussi appelé polynucléotide) est un polymère linéaire composé de 4 *nucléotides*: adénosine (A), cytosine (C), guanine (G) et thymine (T)
- On représente un polynucléotide par une séquence orientée de lettres:

5' -A-T-T-C-A-G-G-C-A-T-T-A-G-C- 3'

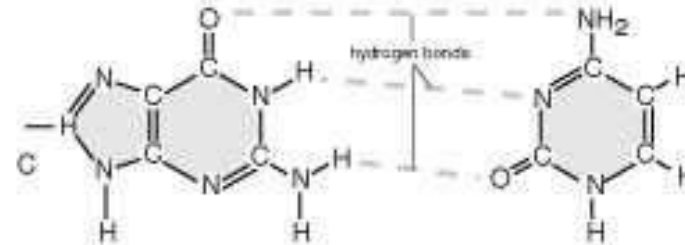
ADN double brin



Nitrogenous Bases

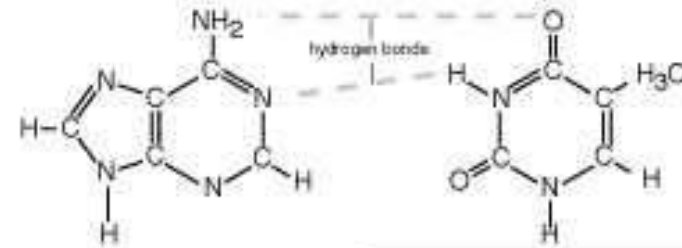
G Guanine

C Cytosine

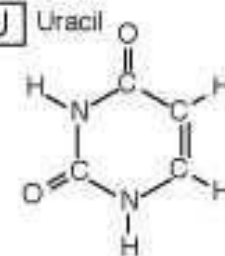


A Adenine

T Thymine

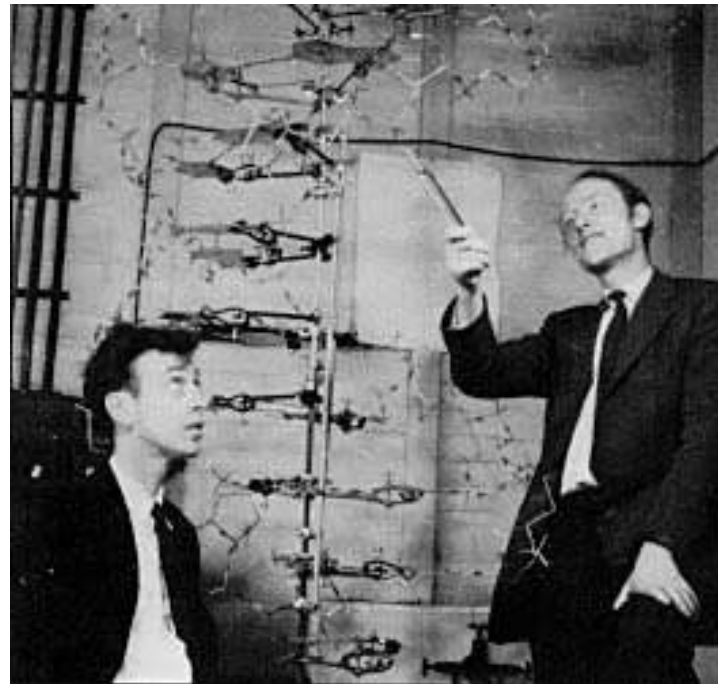
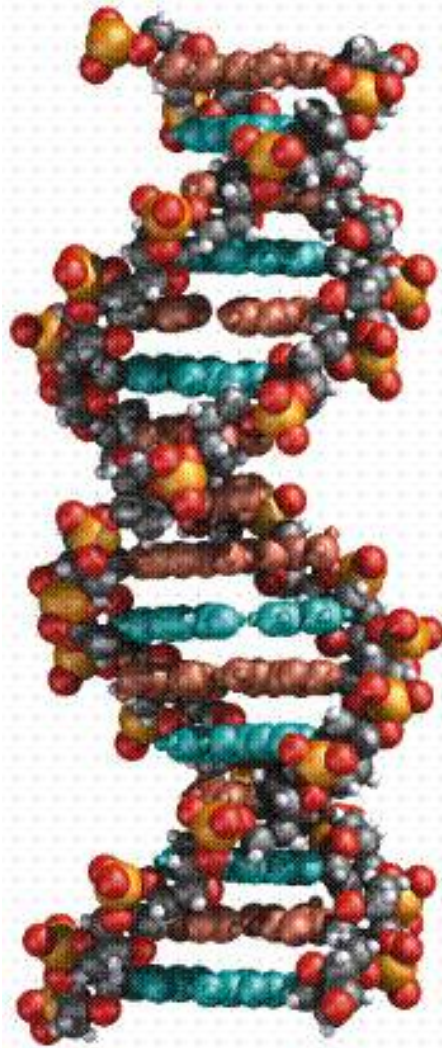


U Uracil



replaces Thymine in RNA

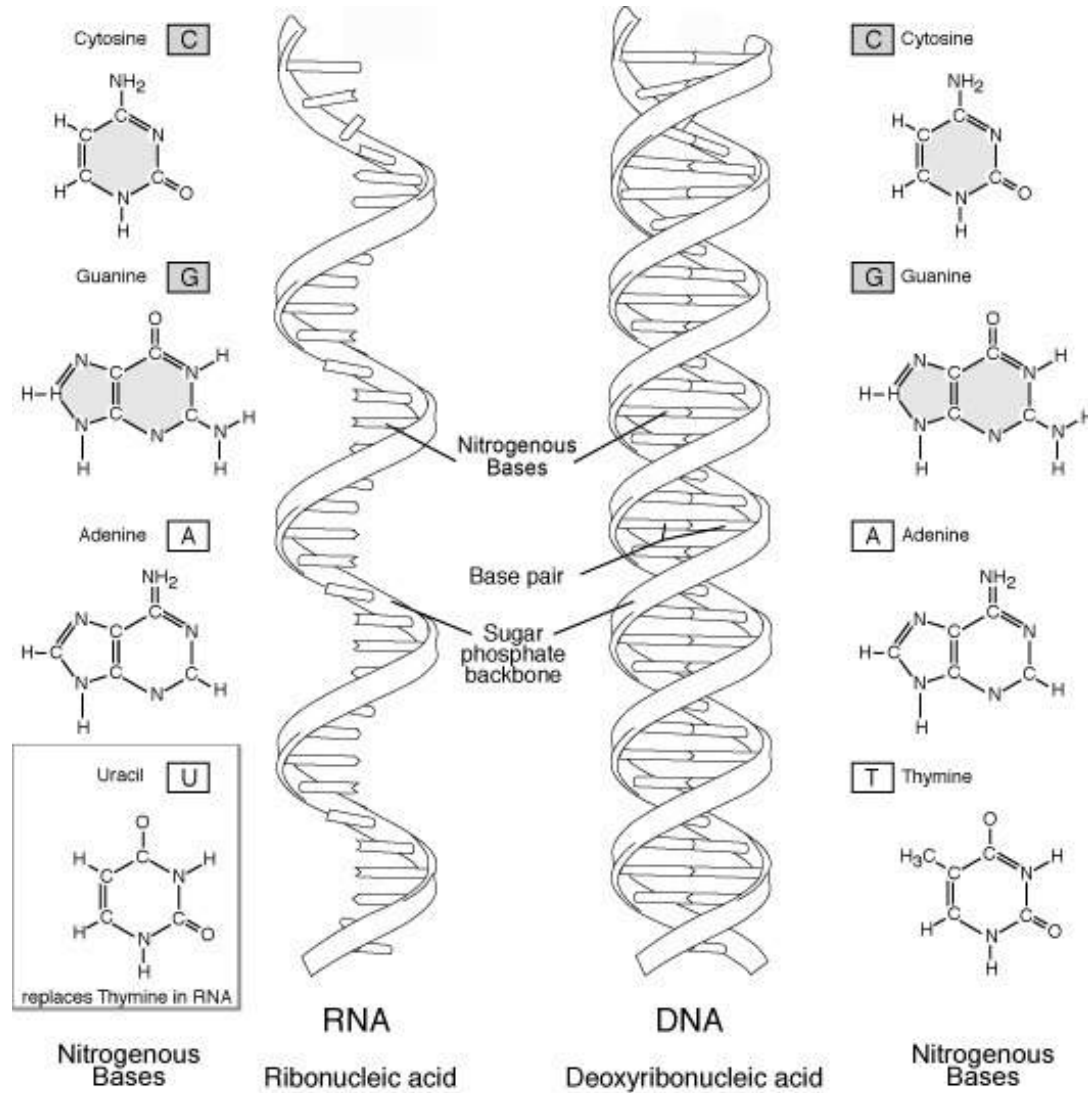
structure de l'ADN (Watson et Crick, 1953)



ADN et information

- La double hélice est stable, quelle que soit la séquence de nucléotides
- Parfait pour stocker 2 bits/base
- Distance entre 2 bases = 0.34nm, donc $6 \cdot 10^7$ bits/cm = *75ko/cm*
- Par repliement de l'ADN en 3D, on peut théoriquement monter à $2 \cdot 10^{21}$ *bits/cm³*

ARN



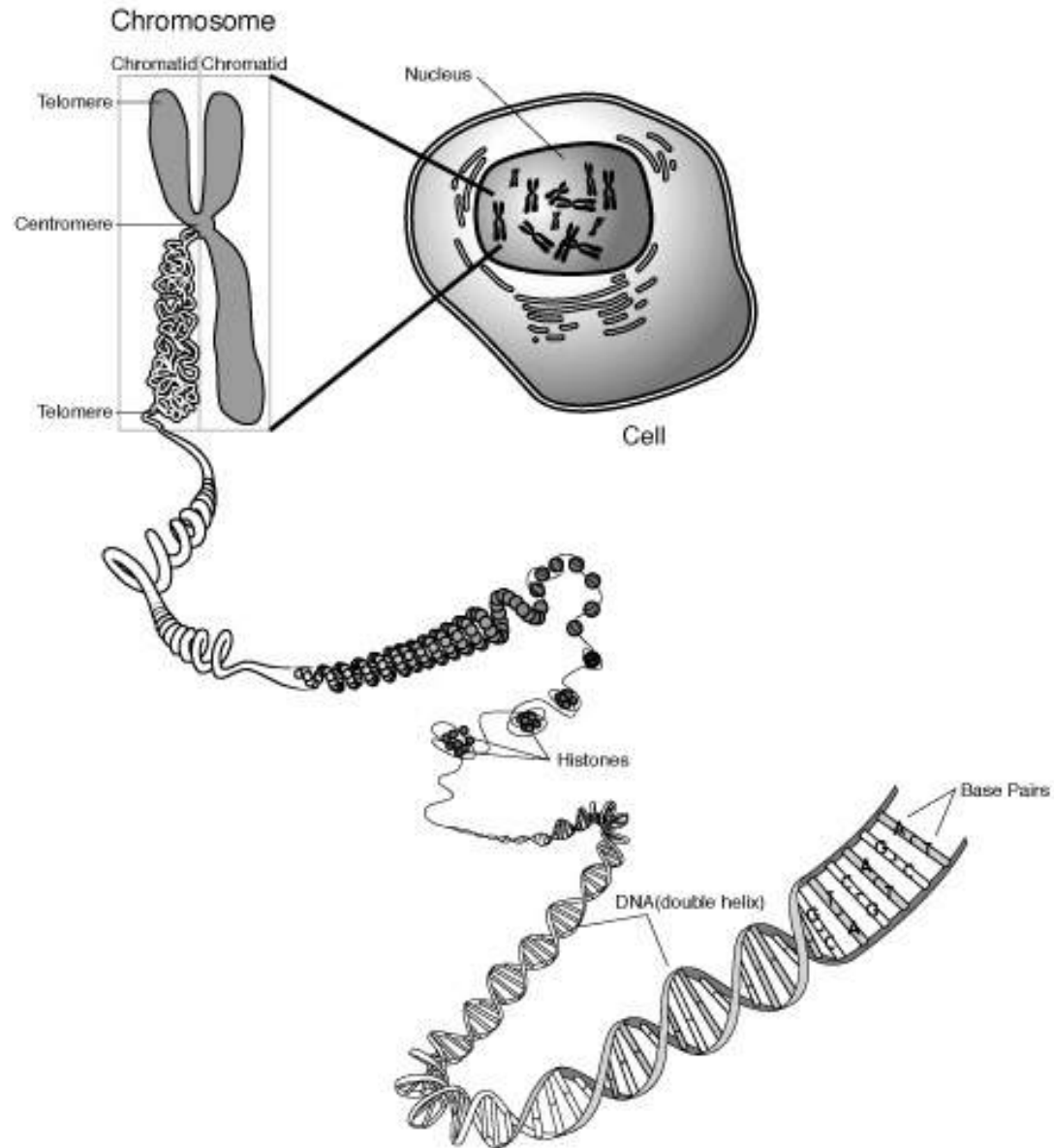
ARN

L'ARN (*acide ribonucléique*) ressemble beaucoup à l'ADN mais:

- Le sucre de l'ADN (désoxyribose) est remplacé par une autre sucre dans l'ARN (ribose)
- La thymine (T) de l'ADN est remplacée par *l'uracile (U)* dans l'ARN.
- L'ARN peut s'apparier avec un ARN complémentaire, mais les ARN sont généralement simple brin et sont donc le siège d'*appariements intramoléculaires*.
- On connaît depuis longtemps *3 types d'ARN*: ARN messagers (ARNm), ARN ribosomiques (ARNr), ARN de transfert (ARNt). Mais on en découvre de nouveaux depuis quelques années...

Gènes et génomes

ADN et chromosomes

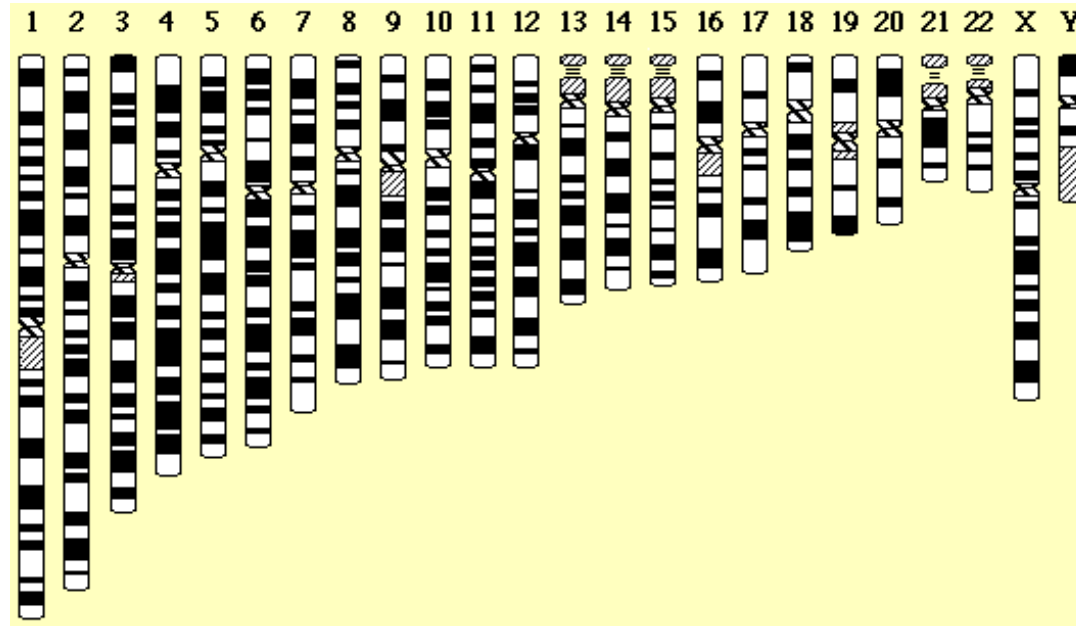


Génome

- Toutes les cellules d'un organisme ont (à peu près) le même ADN, appelé *génom*

Organisme	Chromosomes	Taille du génome (bp)
Bactéries	1	400,000 a 10,000,000
Levure	12	14,000,000
Mouche	4	300,000,000
Homme	46	6,000,000,000

Génomes humains



22 paires de chromosomes + chromosomes X/X ou X/Y =
46 chromosomes.

Séquencage

- *Séquencer* = déterminer la séquence des lettres d'un ADN
- 1995: premier génome bactérien séquencé
- levure (1997), mouche (2000), homme (2003)...
- Approche “shotgun”: les plus grands problèmes pour le séquencage des eukaryotes supérieurs sont *informatique* (assemblage)!

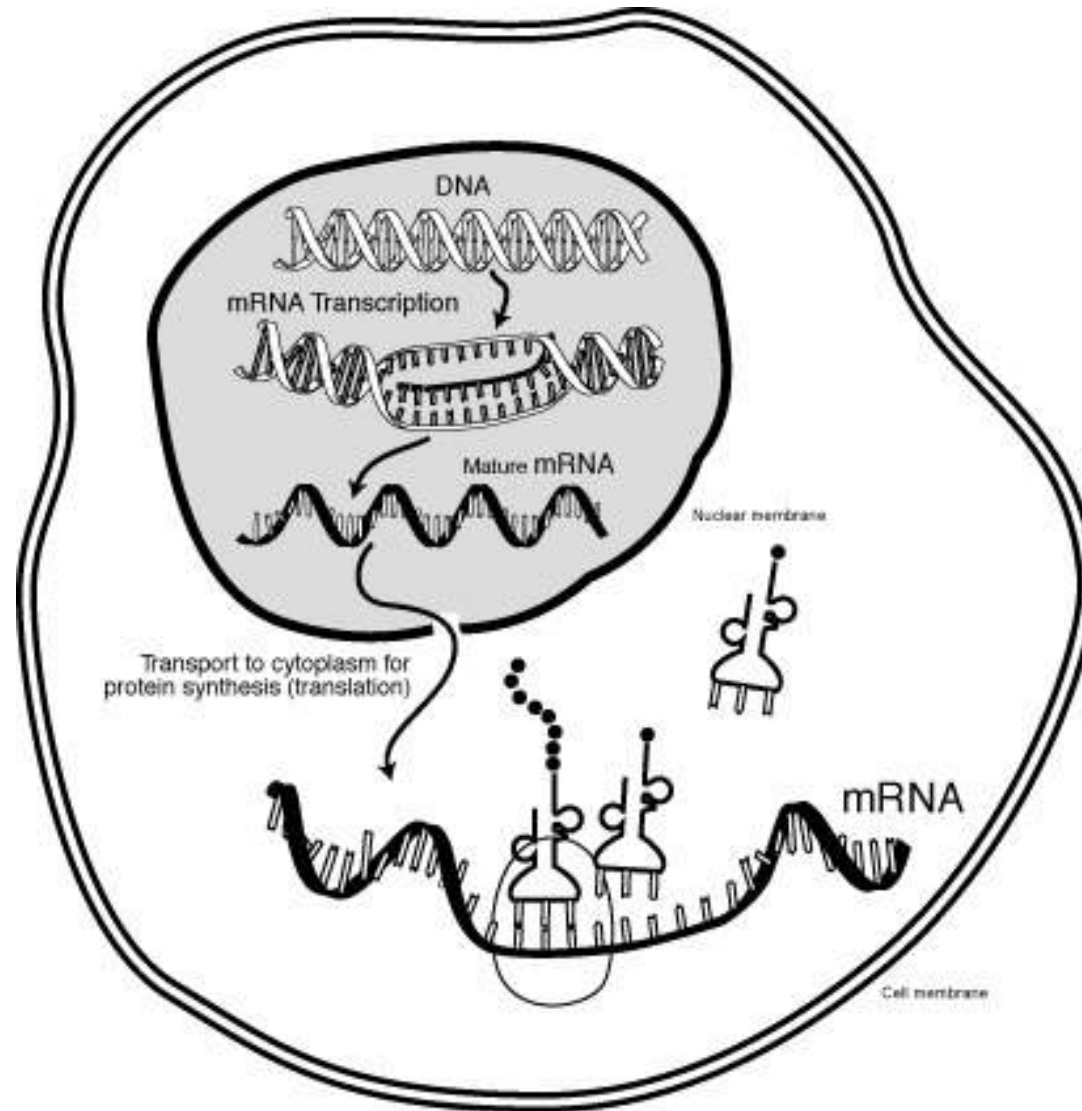
Gène

Une partie continue d'un brin d'ADN, à partir de laquelle une machinerie moléculaire complexe peut lire de l'information (encodée dans les lettres A, C, G, T) et créer une protéine particulière

Dogme central

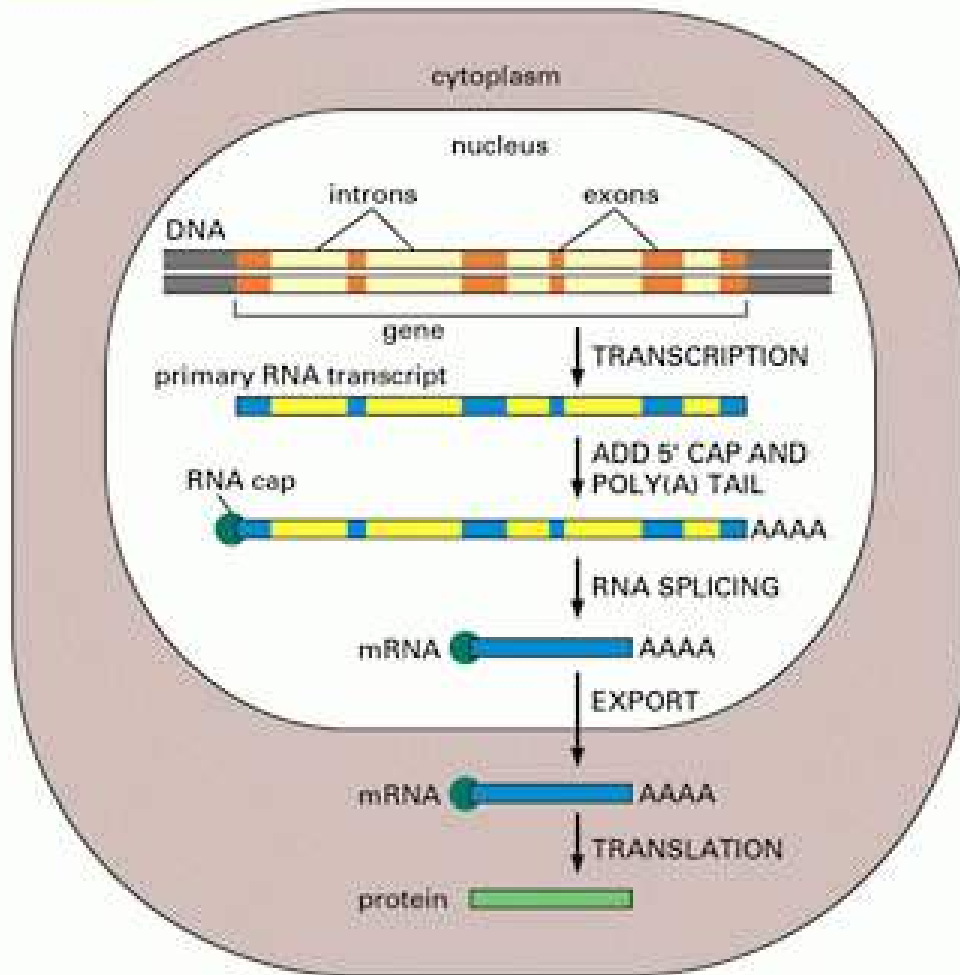


ARN messenger

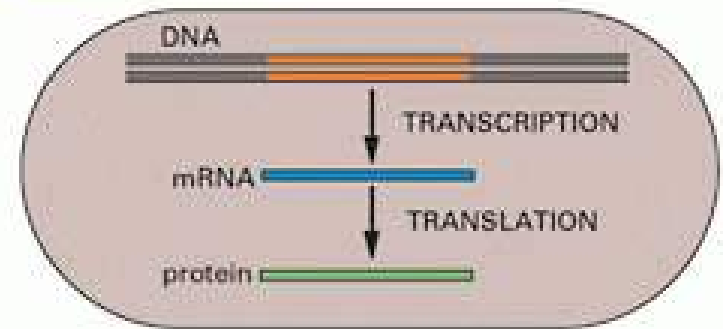


De l'ADN aux protéines

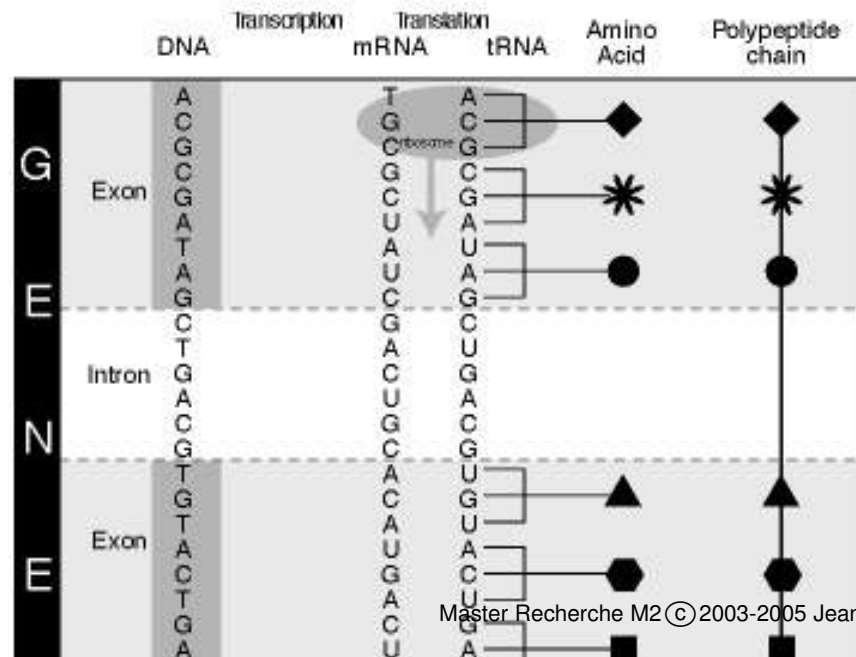
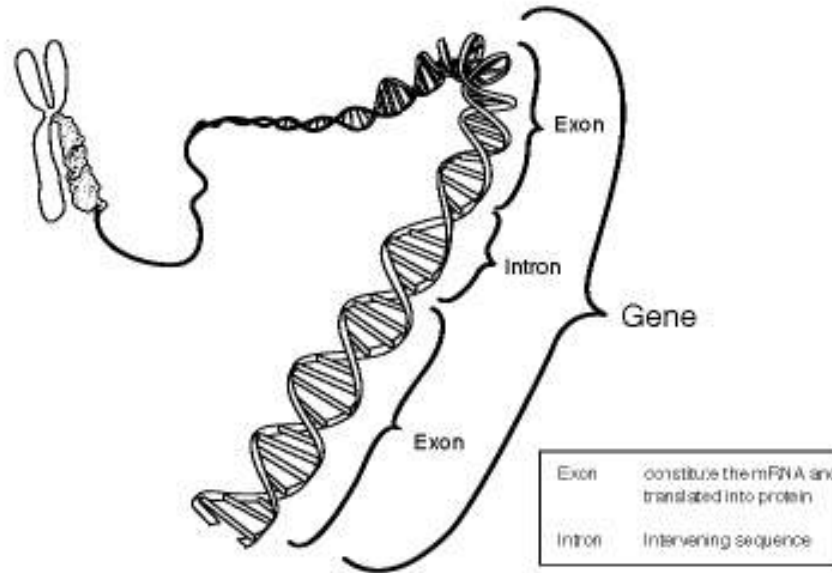
(A) EUCARYOTES



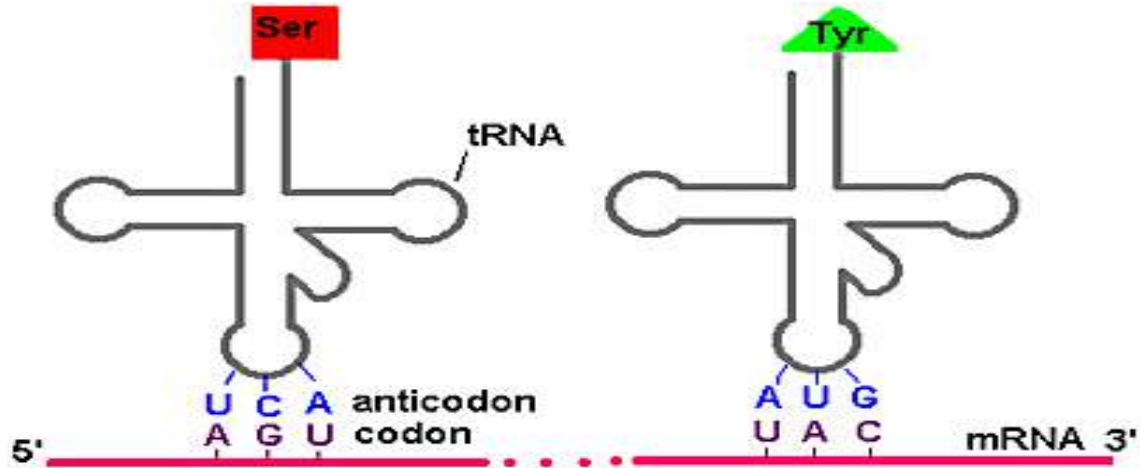
(B) PROCARYOTES



Gènes



Code génétique



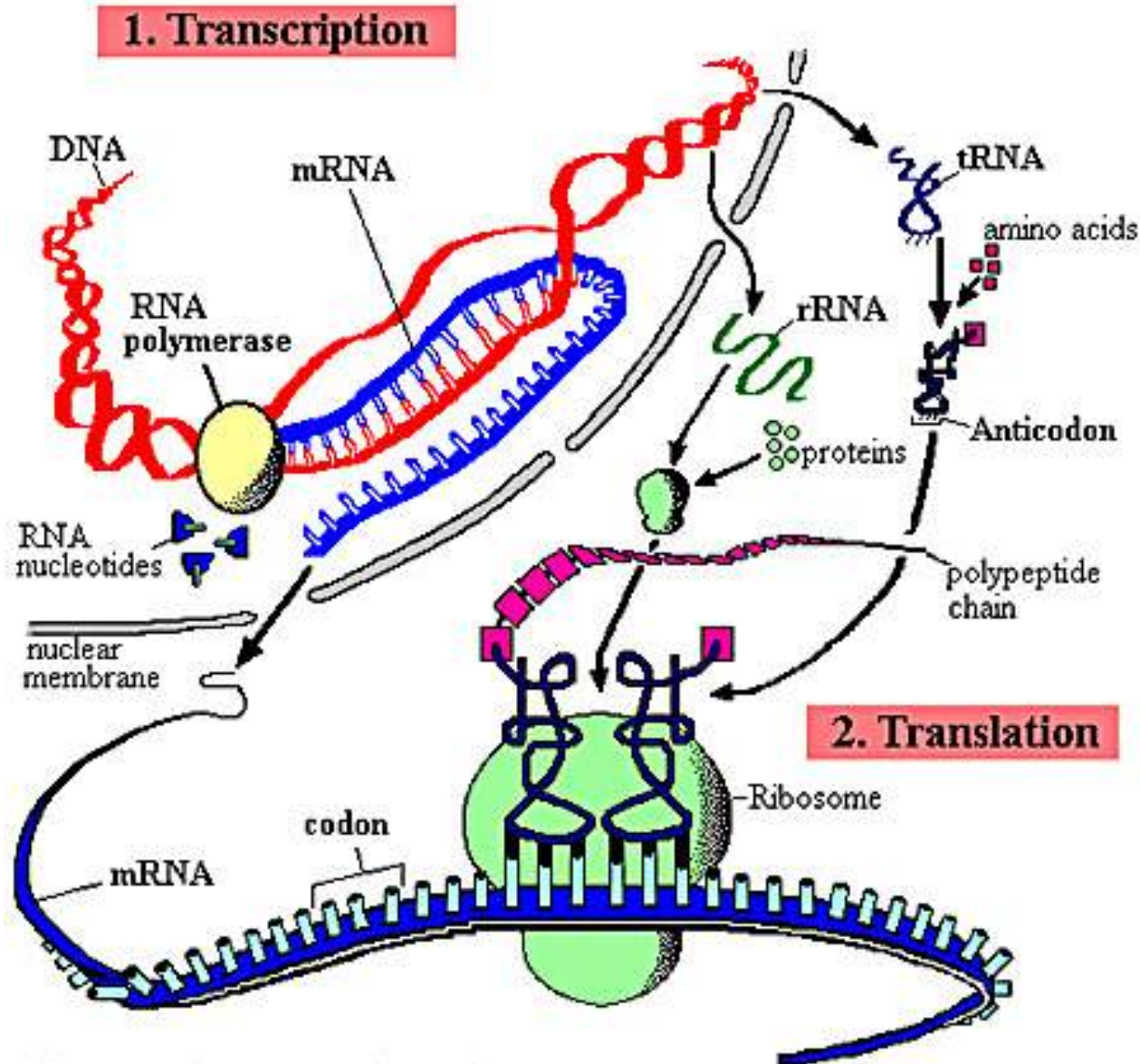
2nd base in codon

		U	C	A	G		
1st base in codon	U	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr STOP STOP	Cys Cys STOP Trp	UCA CAG	
	C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	UCA CAG	
	A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	UCA CAG	
	G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	UCA CAG	

3rd base in codon

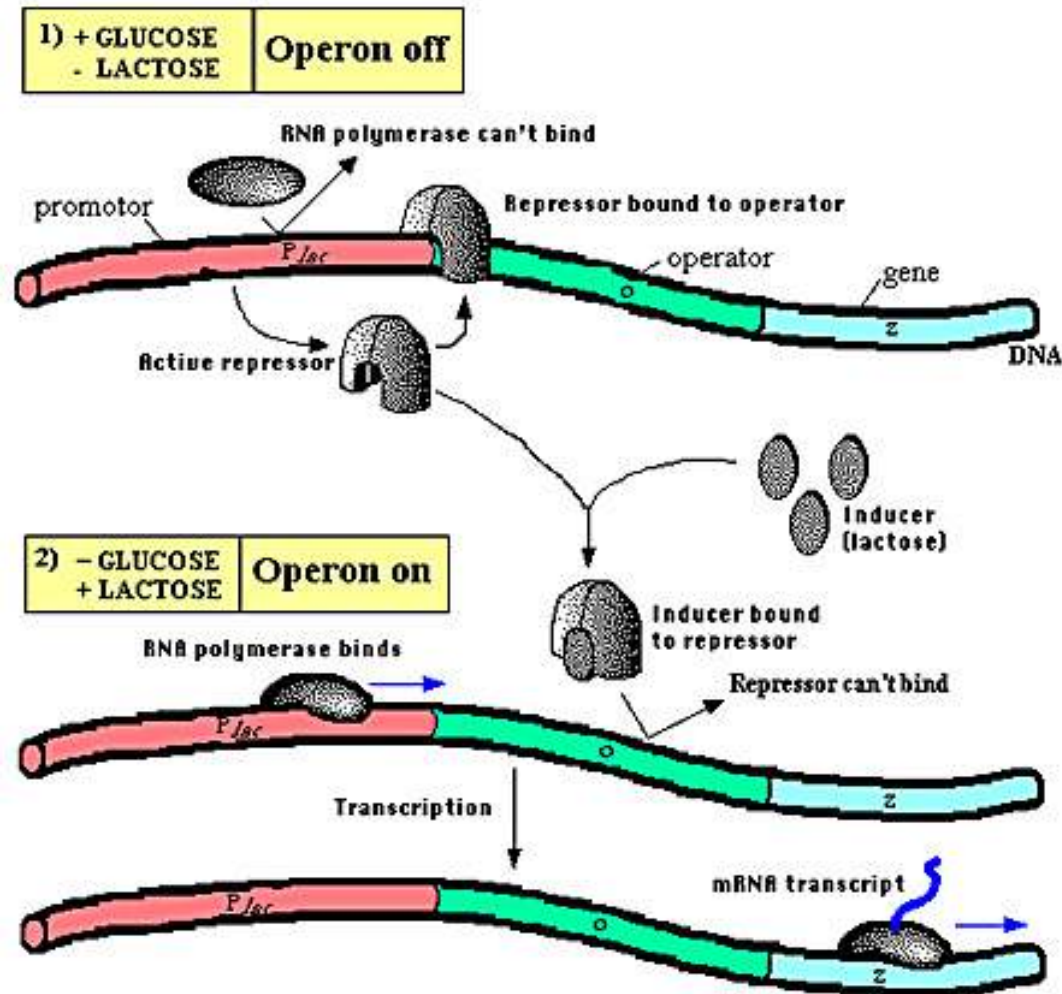
The Genetic Code

De l'ADN à la protéine



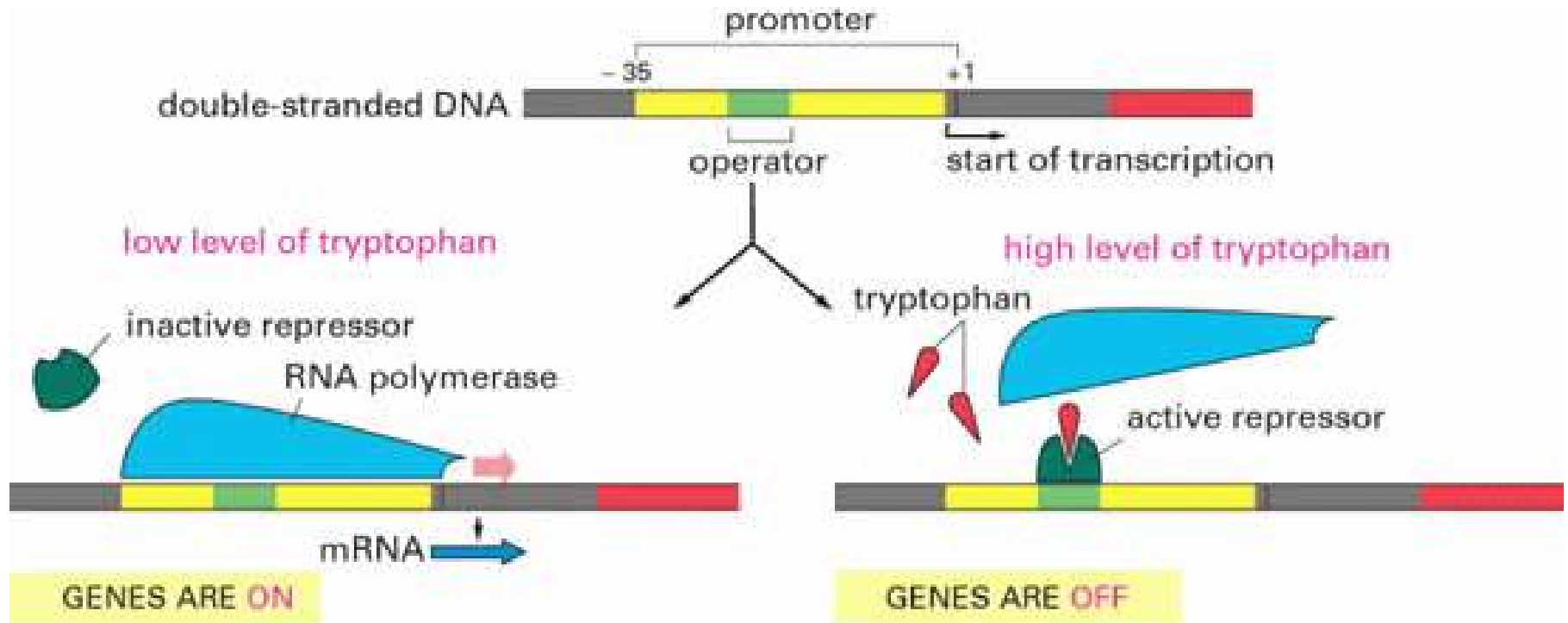
Protein synthesis

Contrôle de l'expression (induction)

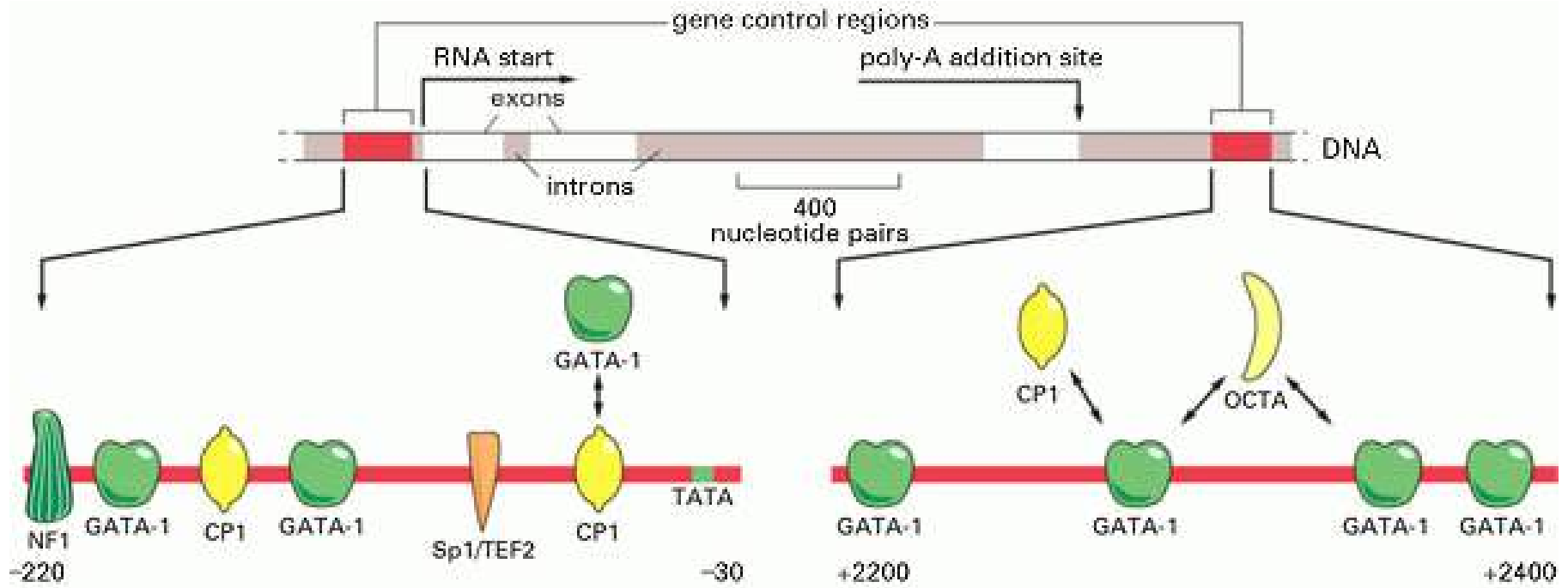


Induction of the *lac* Operon

Contrôle de l'expression (répression)

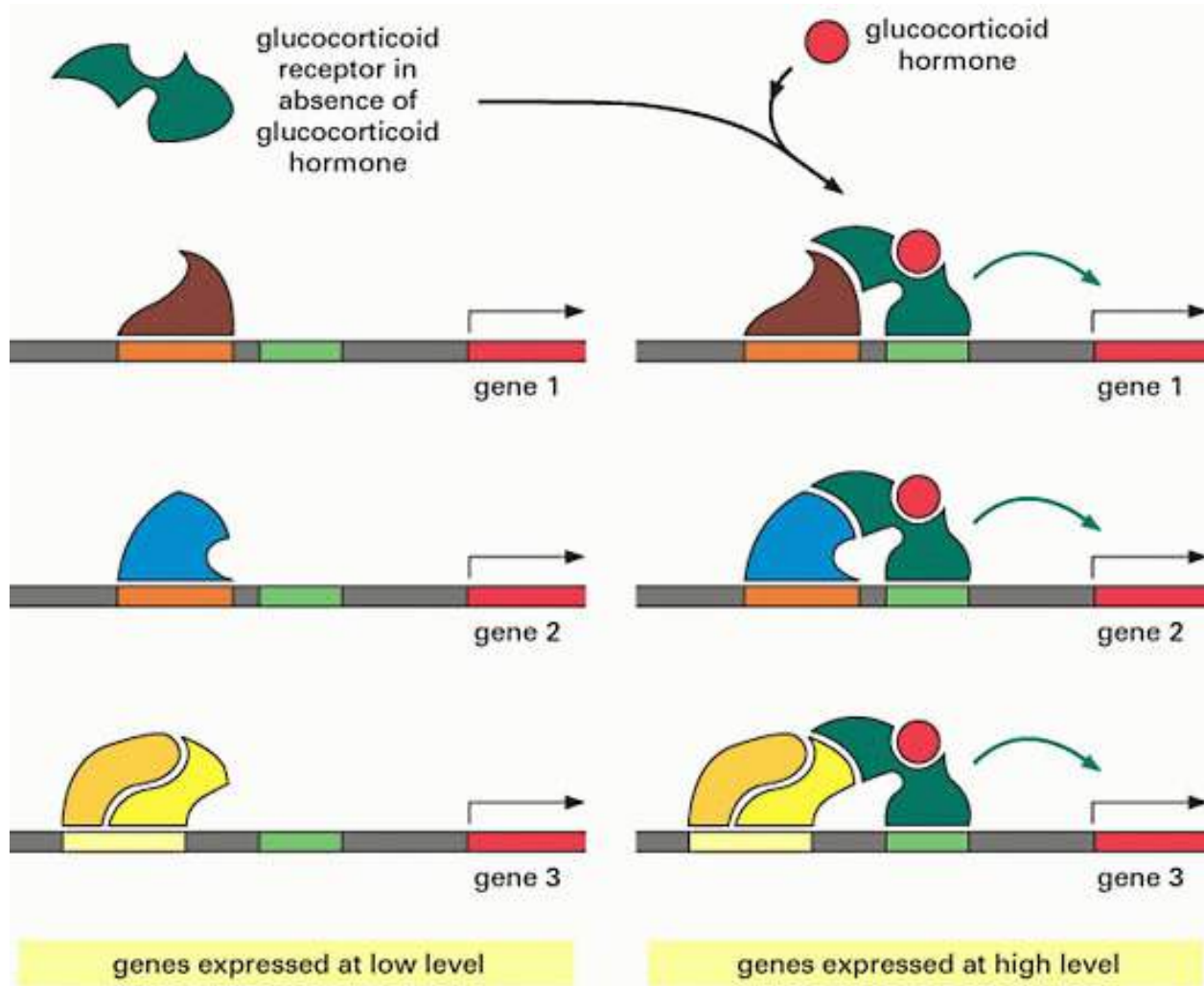


Exemple: B-globine humaine

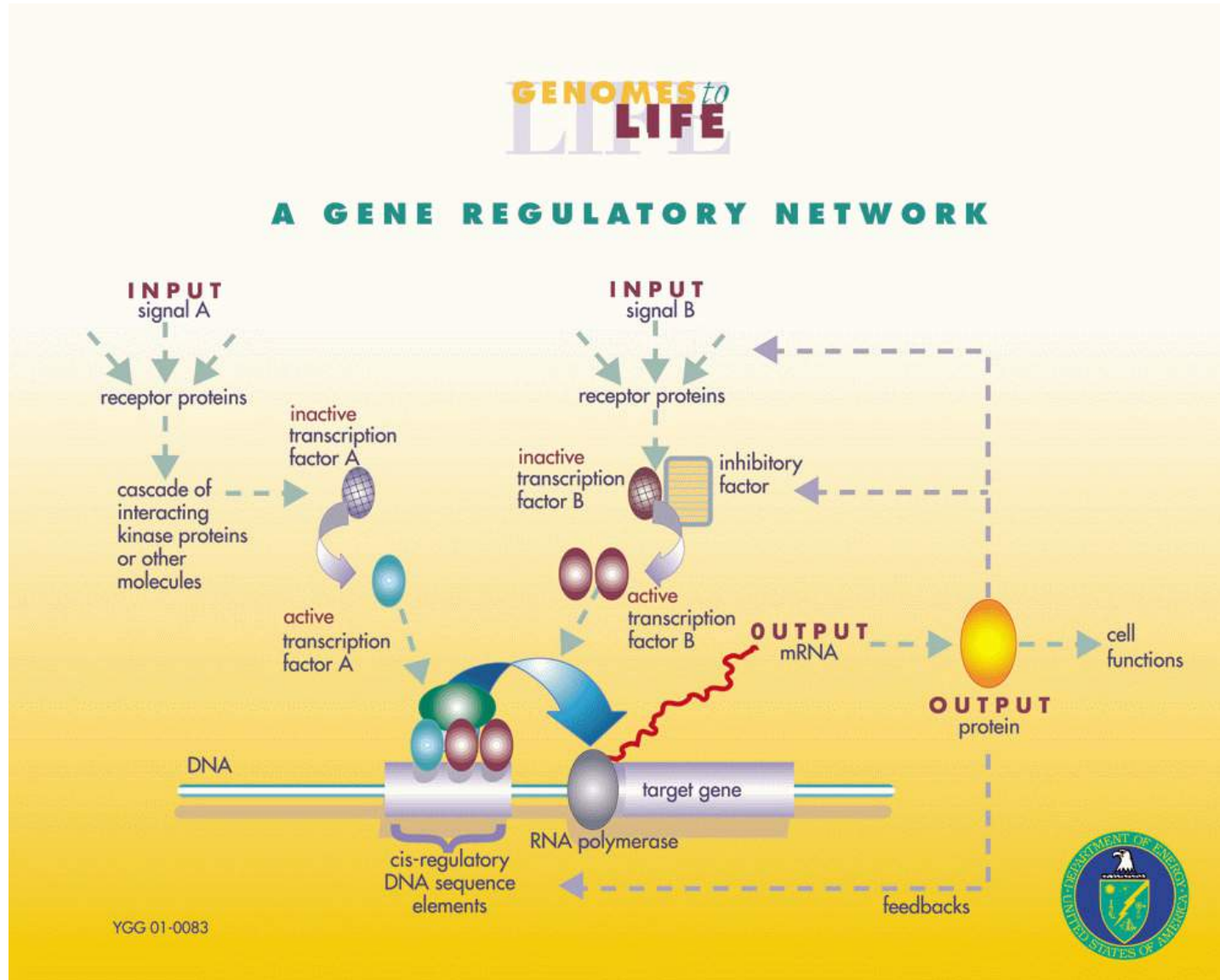


La B-globin joue un rôle important dans le développement des cellules rouges du sang. Certaines protéines régulatrices, comme CP1, sont présentes dans de nombreuses cellules, mais d'autres, comme GATA-1, ne se trouvent que dans quelques types de cellules, dont les précurseurs des cellules rouges.

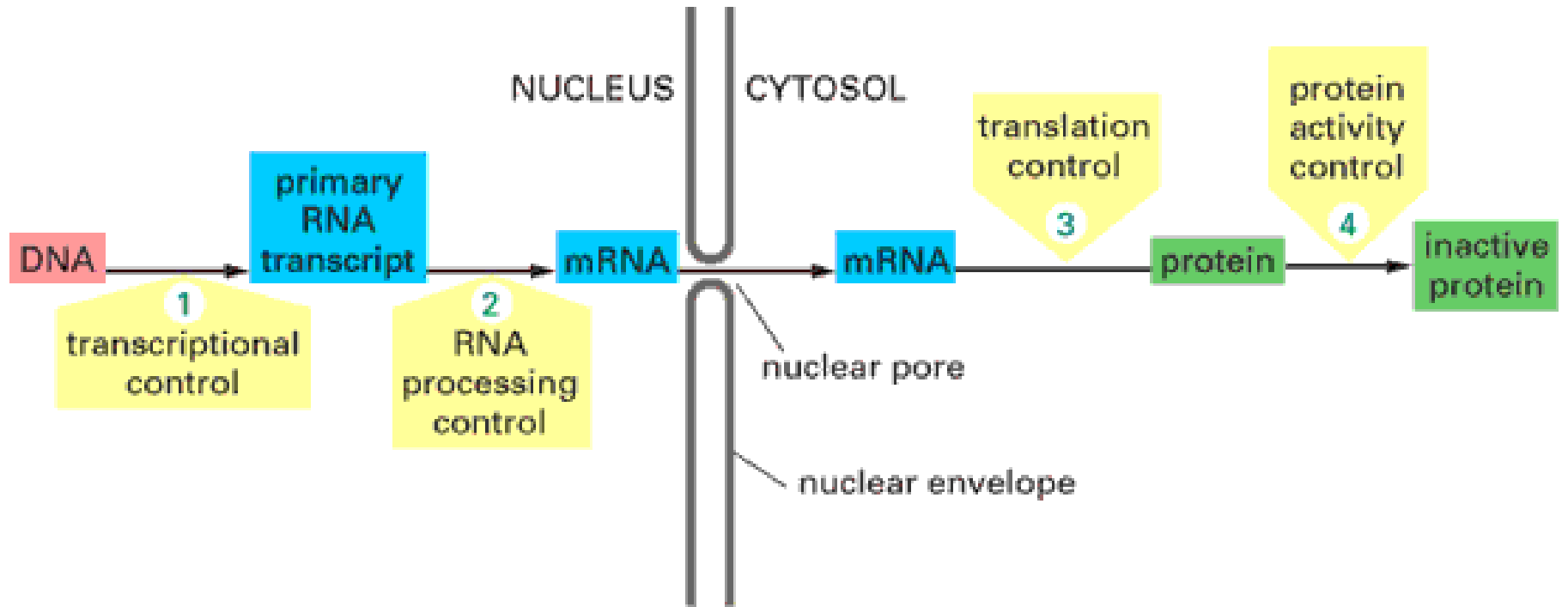
Coordination du contrôle



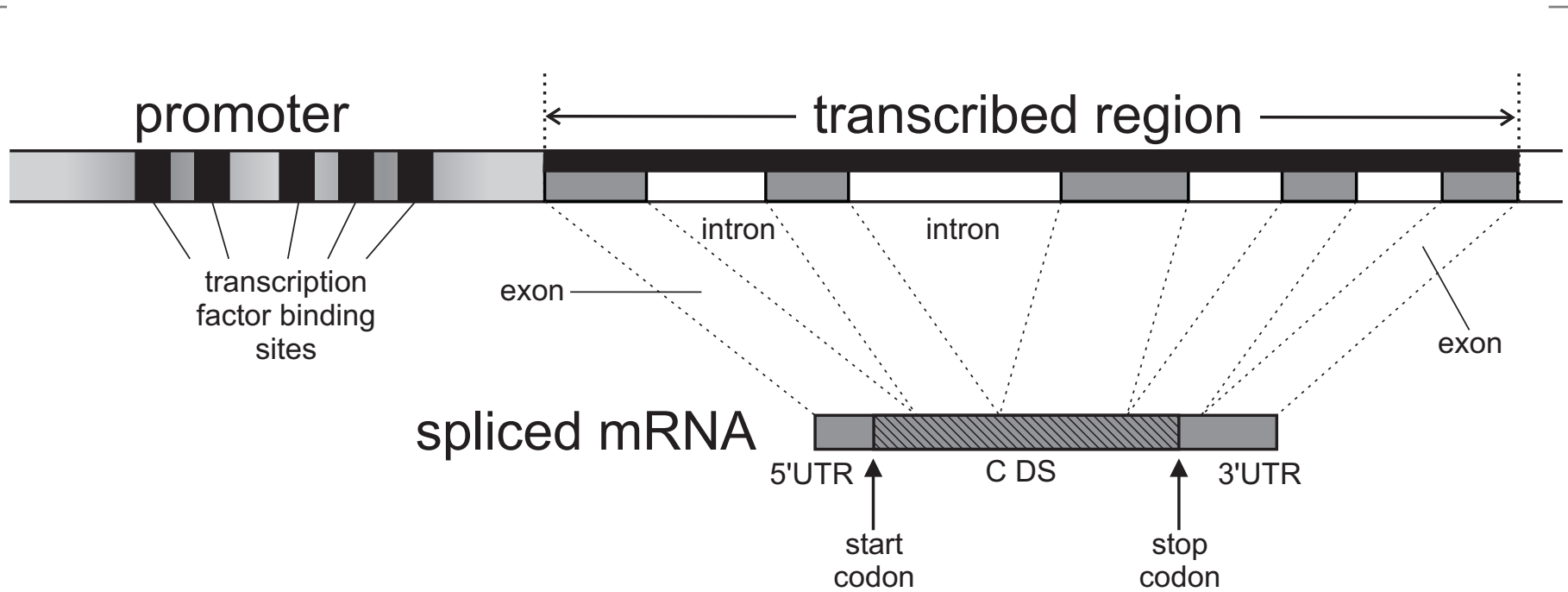
Réseau de régulation



autres contrôles: du gène à la protéine active

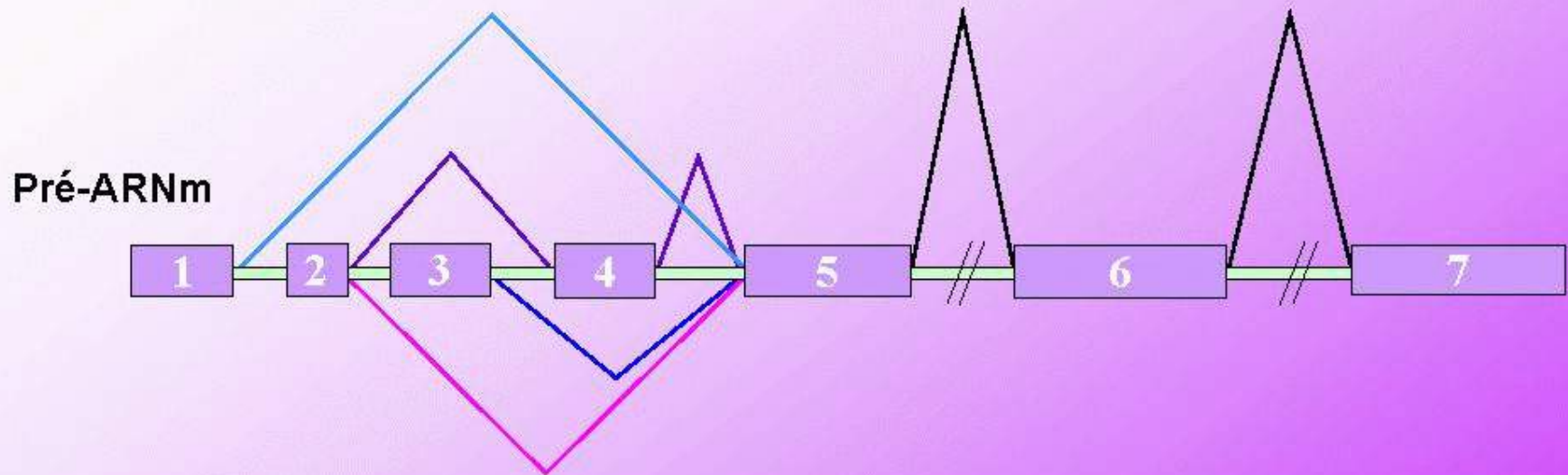


Epissage

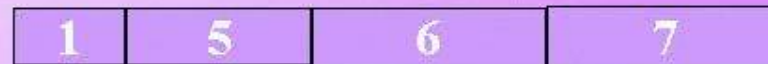


Épissage alternatif

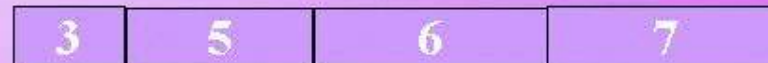
Schéma récapitulatif de l'épissage alternatif



• ARNm 1 :



• ARNm 2 :



• ARNm 3 :



• ARNm 4 :



Nouveau dogme

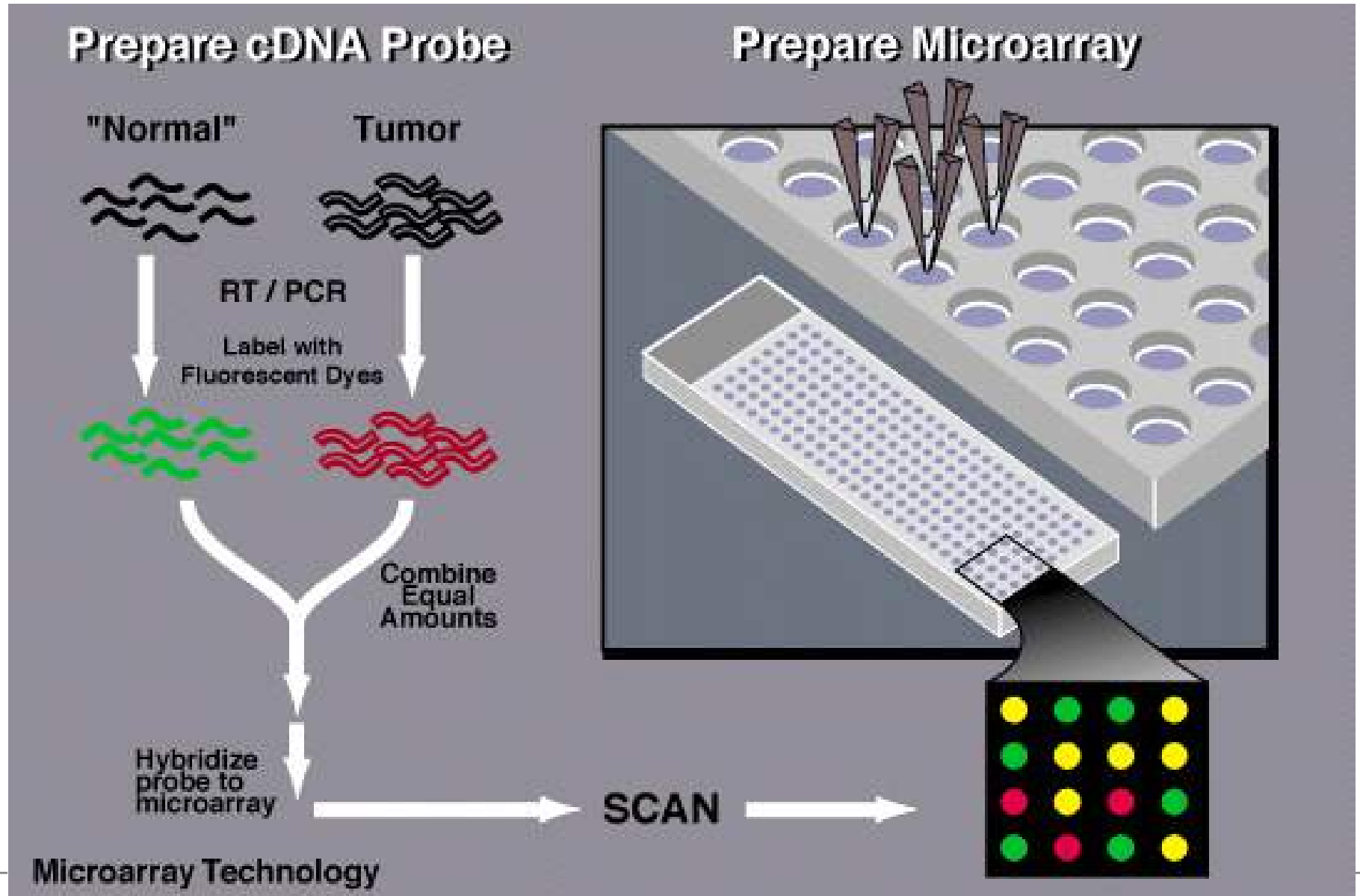
- Avant: 1 gene = 1 ARNm = 1 protéine
- Maintenant: 1 gene = x ARNm = xy protéines
- Rappel: 30,000 genes (?) chez l'homme

Technologies et données

Séquenceur



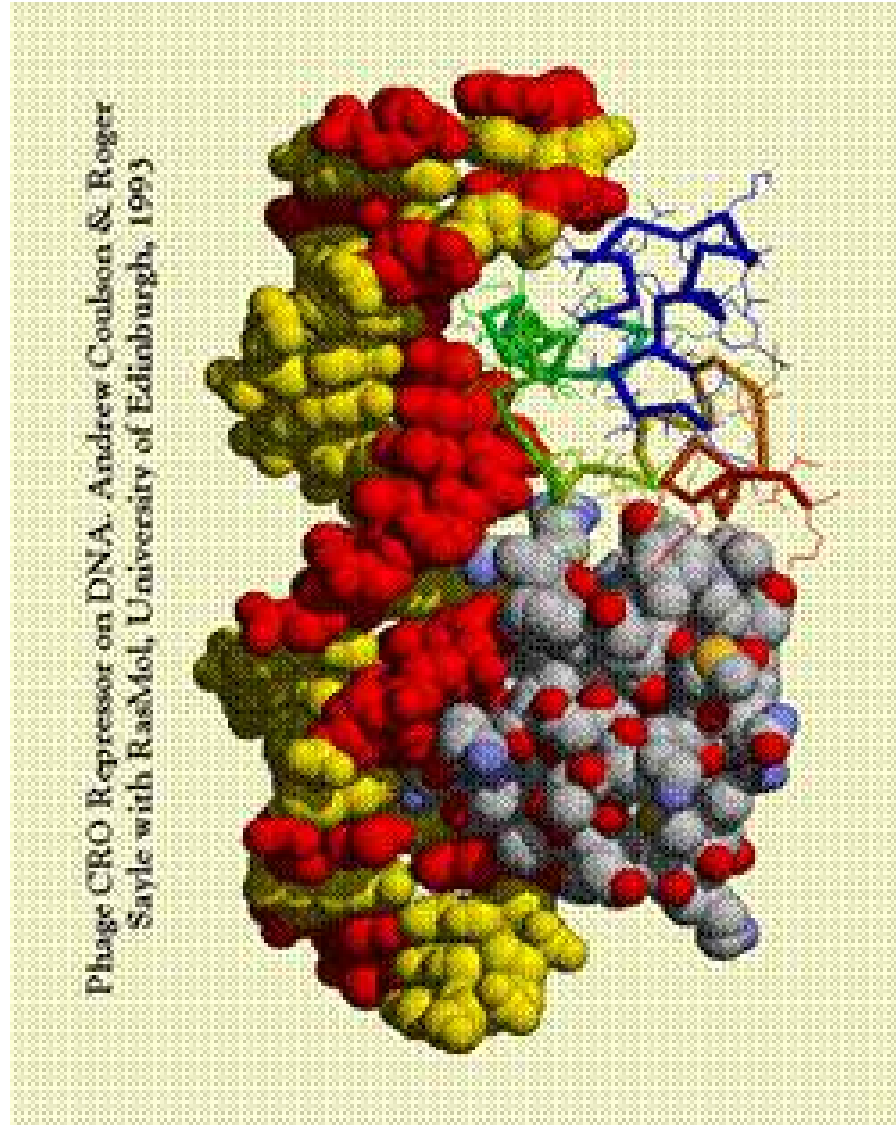
Microarrays



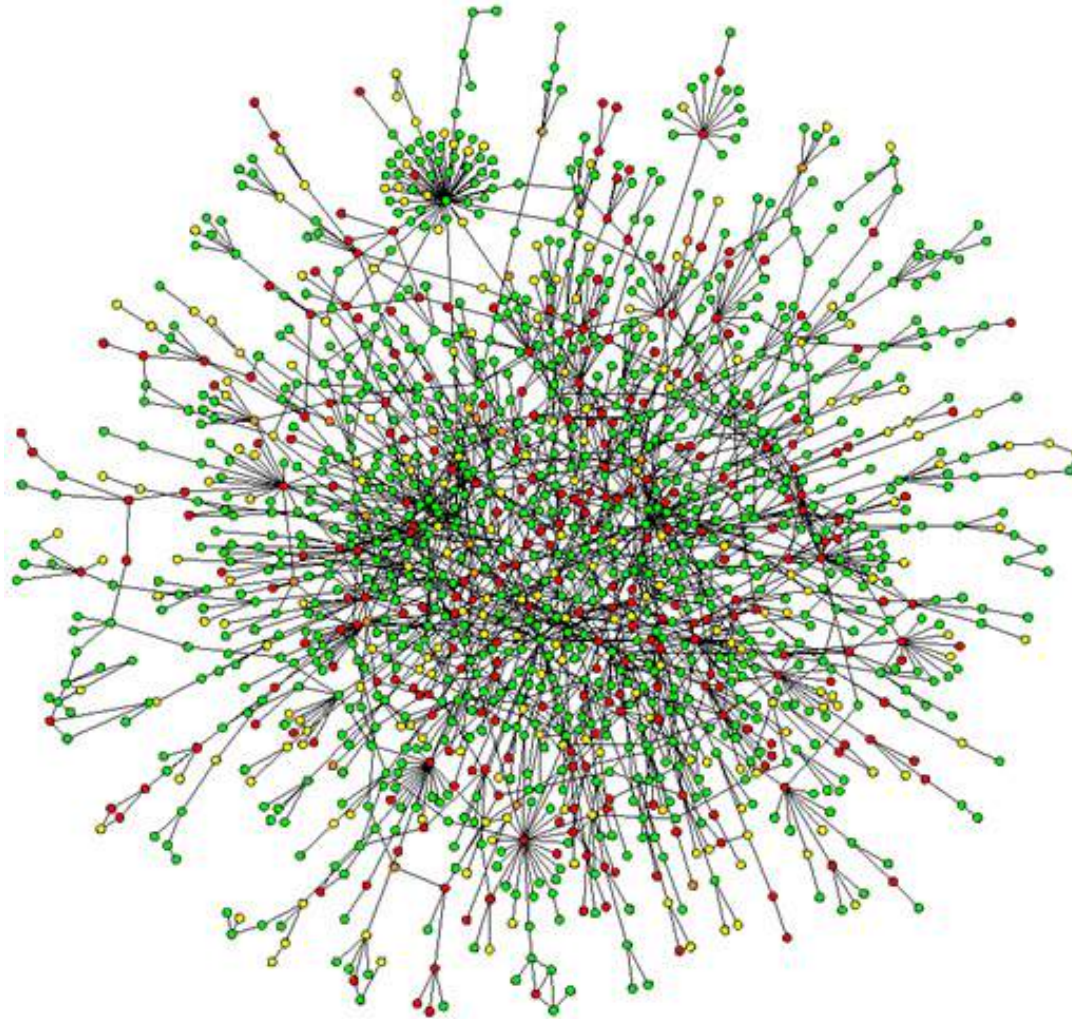
Transcriptome



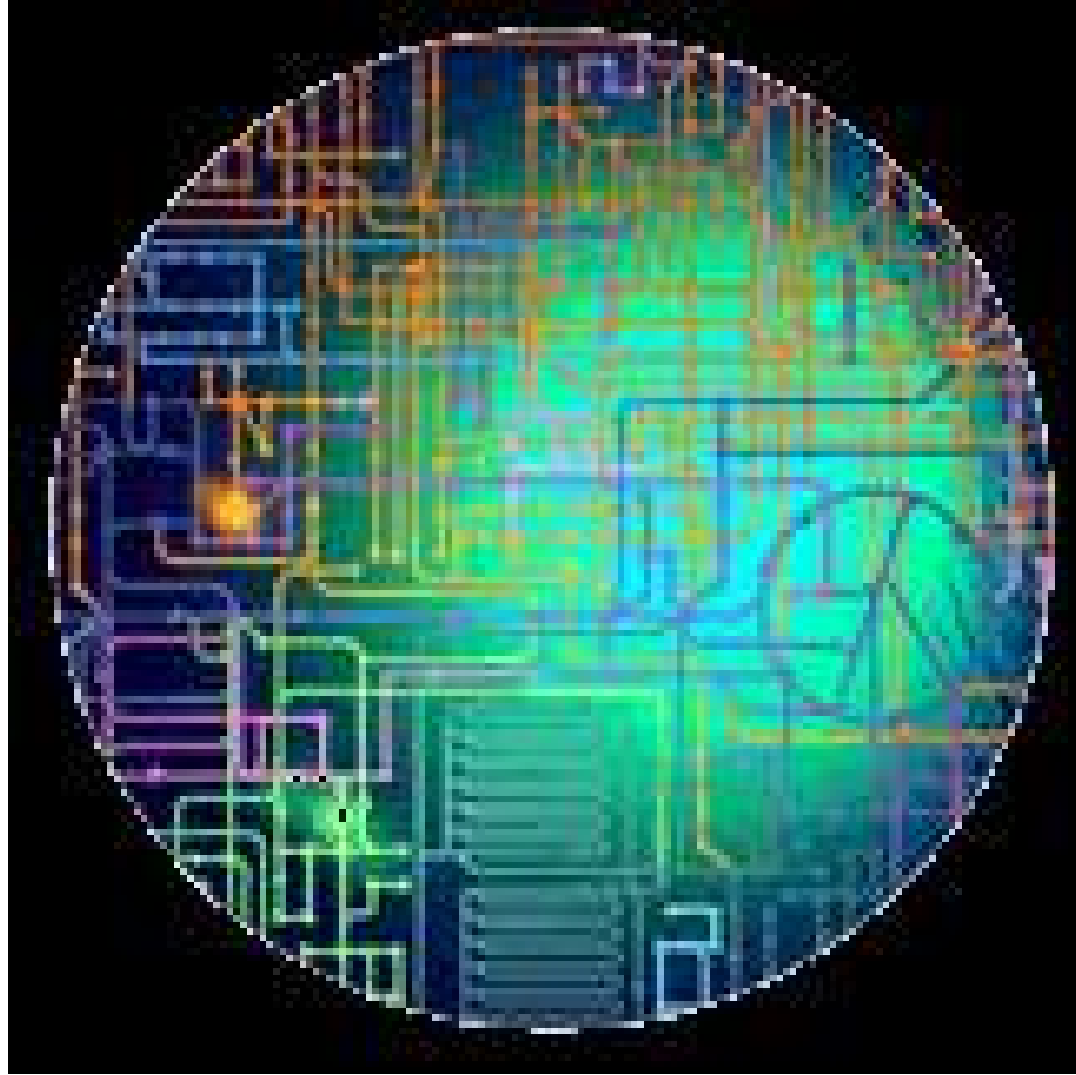
Protéome



Interactome



Métabolome



Data types and representations

Data Type and Details	Representation
Sequences	
- DNA : genome (hereditary information)	string over nucleotides {A,C,G,T}
- full length mRNAs : spliced gene copies	string over ribonucleotides {A,C,G,U}
- ESTs (expressed sequence tags): partial mRNAs	string over ribonucleotides {A,C,G,U}
- proteins	string over amino acids (size 20)

Data types and representations

Data Type and Details	Representation
Structures	
- metabolites: positions and bonds of atoms	labeled graph embedded into 3D-space
- macromolecules (proteins, RNAs, DNA)	labeled graph embedded into 3D-space

Data types and representations

Data Type and Details	Representation
Interactions <ul style="list-style-type: none">- proteins with metabolites: receptors or enzymes binding ligands- proteins with DNA: transcription factors; etc.- proteins with proteins: complexes; etc.	real vectors (binding energies) binary (bipartite graph) binary (graph); Petri-net

Data types and representations

Data Type and Details	Representation
Expression / Localization Data	
- gene expression: abundances of mRNAs	real vectors or matrices
- protein expression: abundances of proteins	real vectors or matrices
- metabolite (small molecule) “expression”: concentrations of metabolites	real vectors or matrices
- protein localization: compartment of presence	categorical

Data types and representations

Data Type and Details	Representation
Cell / Organism Data	
- genotype: single nucleotide polymorphisms	vector of nucleotides {A,C,G,T}
- phenotype: cell type; size; gender; eye color; etc.	vector of real and categorical attributes
- state / clinical data: disease; blood sugar; etc.	vector of real and categorical attributes
- environment: nutrients; temperature; etc.	vector of real and categorical attributes

Data types and representations

Data Type and Details	Representation
Population Data	
- linkage disequilibrium: LOD-scores	real numbers
- pedigrees	certain (tree-like) graphs
- phylogenies: “pedigree of species”	trees or generalizations of trees

Data types and representations

Data Type and Details	Representation
Scientific Texts - Texts: articles, abstracts, web-pages	natural language texts (in English)

Sequence sources

Database	URL (http://...)	Remark
Nucleotide sequence databases		
- DDBJ	www.ddbj.nig.ac.jp	these three databases ...
- EMBL	www.ebi.ac.uk/embl/	... synchronize their ...
- GenBank	www.ncbi.nlm.nih.gov	... contents daily
Protine sequence databases		
- SwissProt	www.expasy.org/sprot/	curated
- TrEMBL	www.expasy.org/sprot/	not curated

Sequence sources

(Some) Sequence motif databases

- eMotif `motif.stanford.edu/emotif/` protein regular expression patterns
- SMART `smart.embl-heidelberg.de/` protein domain HMMs
- TRANSFAC `transfac.gbf.de/TRANSFAC/` transcription factor binding sites

Sequence sources

General portals

- EBI `www.ebi.ac.uk` European Bioinformatics Institute
- Entrez `www.ncbi.nlm.nih.gov/Entrez/` U.S. National Bioinf. Institute
- Ex-PASy `www.expasy.org` Expert Protein Analysis System
- SRS `srs.ebi.ac.uk` Sequence Retrieval System

Expression sources

Database	URL (http://...)	Remark
General databases		
- ArrayExpress	www.ebi.ac.uk/arrayexpress/	by the EBI
- GEO	www.ncbi.nlm.nih.gov/geo/	by the NCBI
Organism specific databases		
- MGI GXD	www.informatics.jax.org	mouse
- TAIR Microarray	www.arabidopsis.org	<i>arabidopsis</i>
- WormBase	www.wormbase.org	<i>C. elegans</i>

Protéine properties sources

Database	URL (http://...)	Remark
Protine structures		
- PDB	www.rcsb.org/pdb/	3D structures
- SCOP	scop.mrc-lmb.cam.ac.uk/scop/	structural classification
- CATH	www.biochem.ucl.ac.uk/bsm/cath/	structural classification

Protéine properties sources

Molecular interactions and networks

- BIND `www.bind.ca`

interaction
network

- KEGG `www.genome.ad.jp/kegg/`

metabolic
pathways

- DIP `dip.doe-mbi.ucla.edu`

interacting
proteins

Protéine properties sources

Protine functions

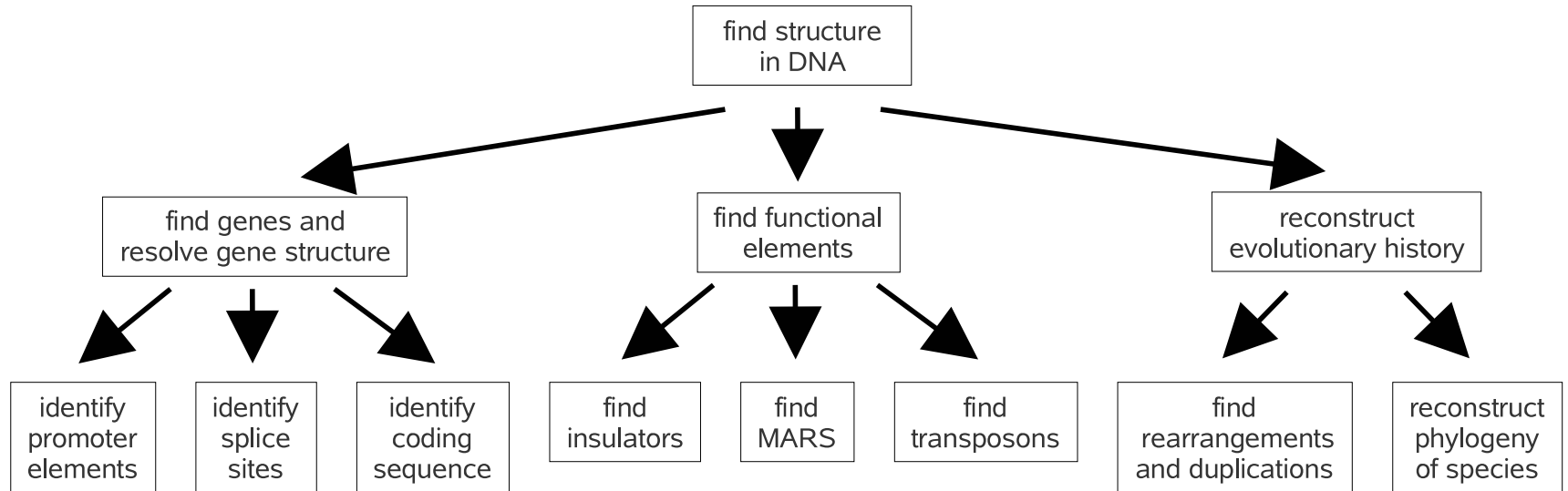
- GO www.geneontology.org controlled vocabulary
- EC www.chem.qmul.ac.uk/iubmb/enzyme/ enzyme numbers
- MIPS mips.gsf.de/proj/yeast/catalogs/funcat/ yeast gene functions

Protine expression

- us.expasy.org/ch2d/ 2D gel electrophoresis data

Challenge en bio-informatique

Génomique



Protéomique

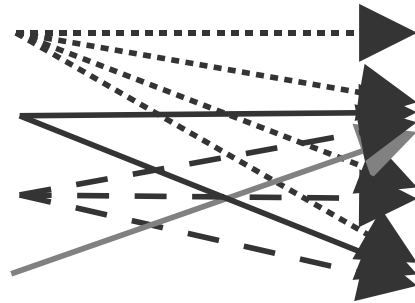
Given Data

sequence

structure

expression

phylogeny



Predicted Property

structure (3D coordinates of the atoms)

function (e.g., according to GO or MIPS)

interactions (with other proteins, DNA or metabolites)

localization (e.g., compartment)

Pharmacogénomique



Caractéristiques

- *Beaucoup* de données...
- mais *beaucoup* de "bruit"
- Données hétérogènes (séquences, structures, vecteurs, graphes...)
- "Small n large p"
- problèmes souvent *mal posés* (data mining)

Pour vous motiver

- discipline nouvelle (les données n'existaient pas il y a 10 ans)
- application (therapeutique, biologie fondamentale)
- besoin de math/info de plus en plus pointu (voir évolution récente du domaine)
- peu de spécialistes...

But du cours

Proposer une théorie et des outils pour

- *représenter* les données dans un cadre mathématique cohérent...
- ...avec des *méthodes* d'analyse performantes...
- ...en *pleine expansion* actuellement.

Contenu du cours

- *Introduction* à la biologie moléculaire et à la génomique
- *Noyaux positifs*: définition, propriétés, espaces de Hilbert à noyau reproduisant, kernel trick, théorème du représentant
- *Méthodes à noyau*: kernel PCA, SVM, LS-SVM, kernel CCA
- *Noyaux*: pour séquences, pour graphes, noyau de diffusion, noyau de convolution, noyau de semi-groupe
- *Applications*: classification de séquences, inférence sur des graphes, sélection de genes

Crédits

Source des images et tables

- Alex Zien, "A primer on molecular biology", *Kernel Methods in Computational Biology* (B. Schölkopf, K. Tsuda, J.-P. Vert ed.), MIT Press, 2004
- Image gallery:
<http://www.accessexcellence.org/AB/GG/>
- A quick introduction to elements of biology - cells, molecules, genes, functional genomics, microarrays, by Alvis Brazma, Helen Parkinson, Thomas Schlitt, Mohammadreza Shojatalab
http://www.ebi.ac.uk/microarray/biology_intro
- .. et quelques images trouvées sur le web