Introduction a la biologie moléculaire et a la bio-informatique

Cours de DEA, 2003/2004

Jean-Philippe Vert

Jean-Philippe.Vert@mines.org

Plan

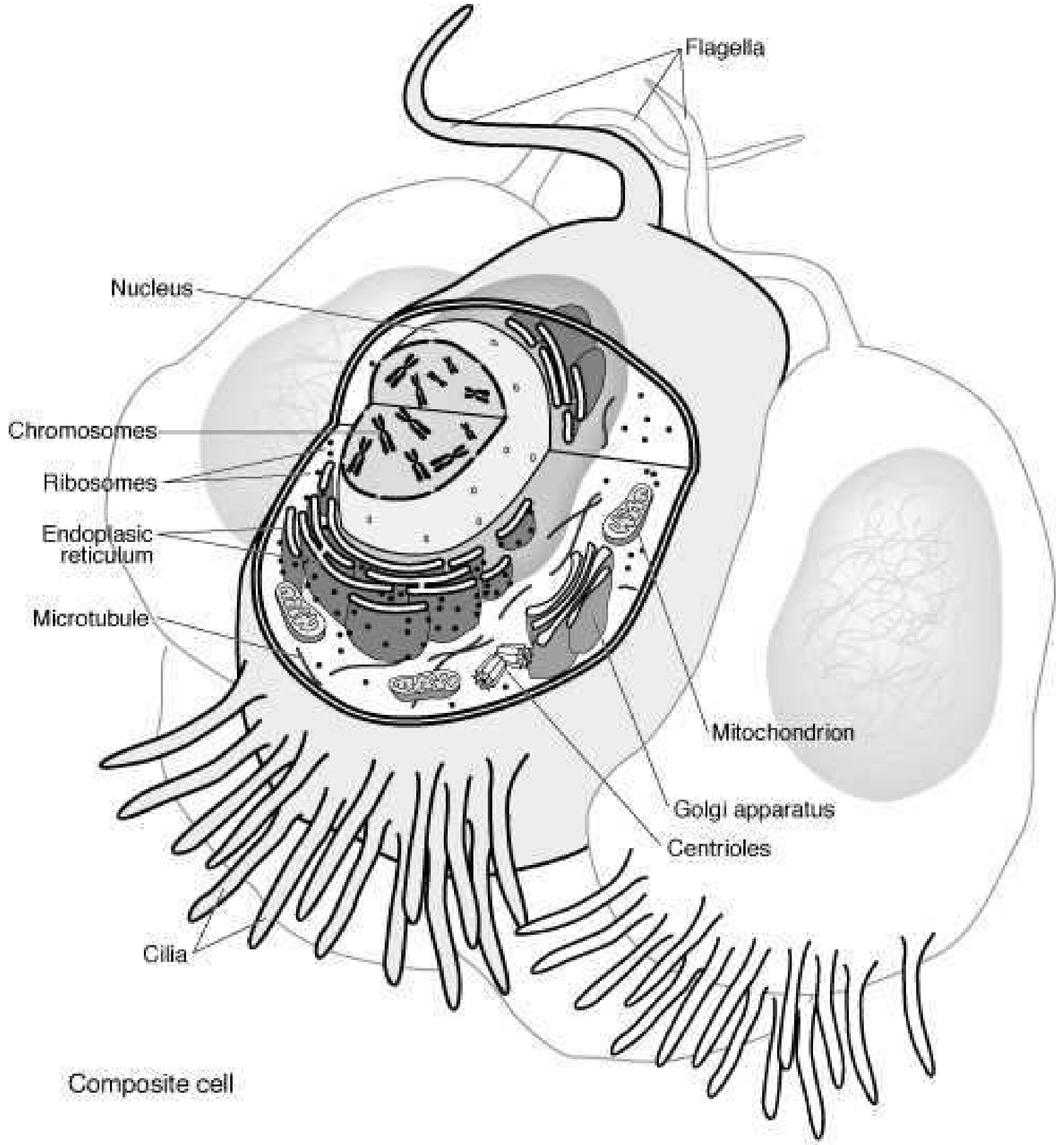
- Organismes et cellules
- Molécules de la vie
- Genes et génomes
- Technologies et données
- Challenges en bio-informatique

Organismes et cellules

Cellules

- Toute 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...)

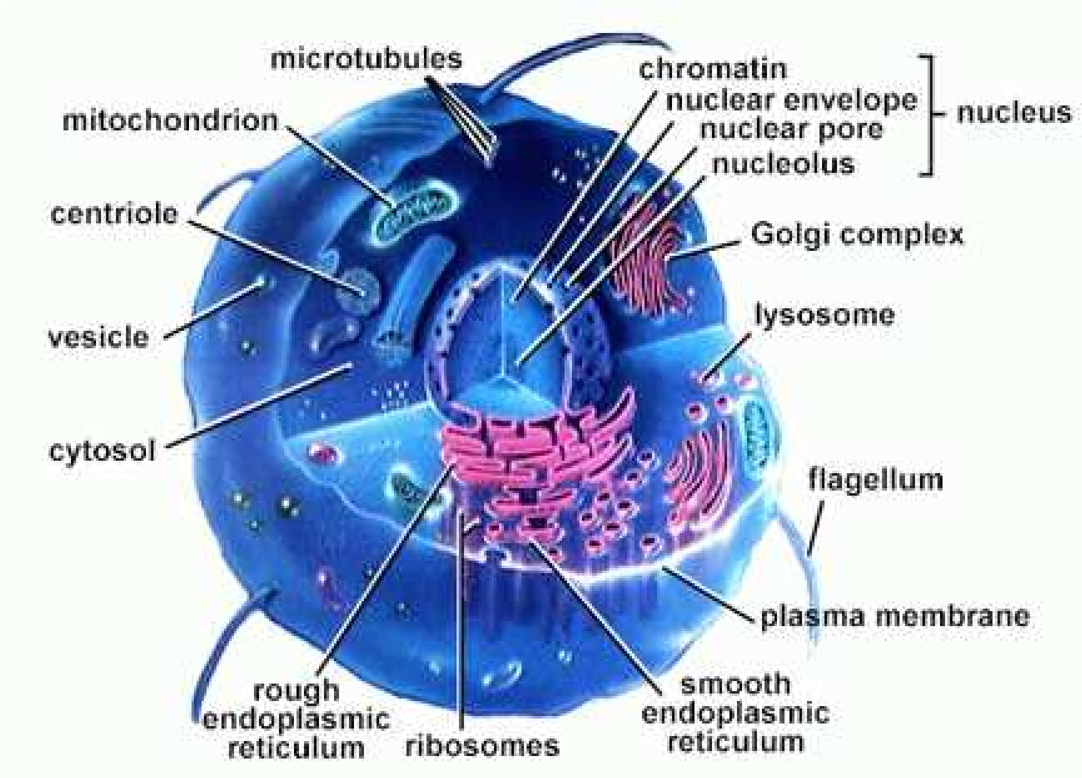
Cellules


Flagella
Nucleus
Chromosomes
Ribosomes
Endoplasic reticulum
Microtubule
Mitochondrion
Golgi apparatus
Centrioles
Cilia
Composite cell

Classification des organismes

- On distingue généralement les *eukaryotes* des *prokaryotes*
- Les prokaryotes (eux-memes subivisés en *bactéries* et *archéens*) sont unicellulaires, de petite taille (typiquemet $1\mu m$), 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 ancetre 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 convertit en molécules utiles et energie
- Elles peuvent réagir a leur environnement

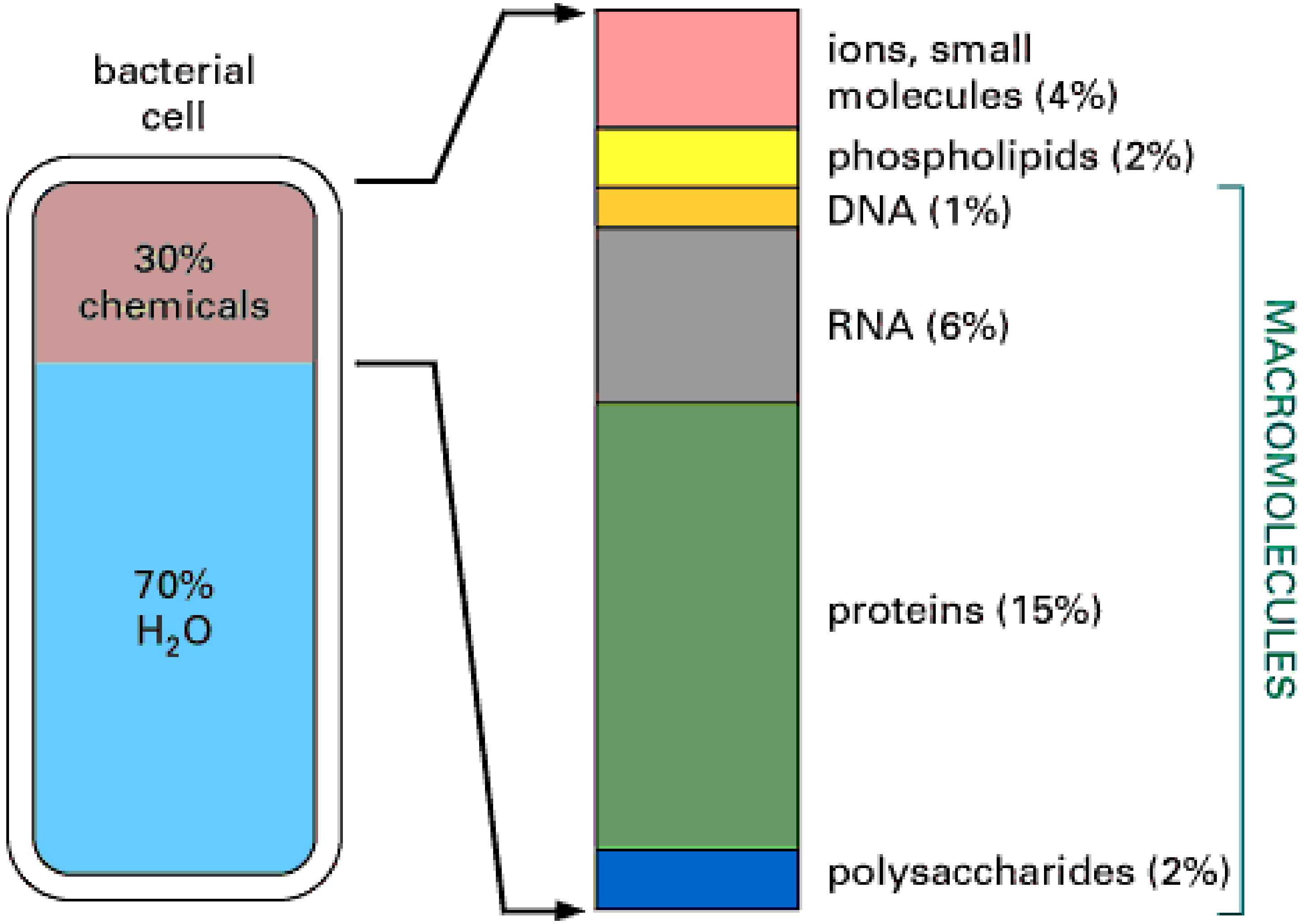
Les molécules de la vie

Types de molécules

On les regroupe en 4 grandes familles:

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

Dans la cellule



Petites molécules

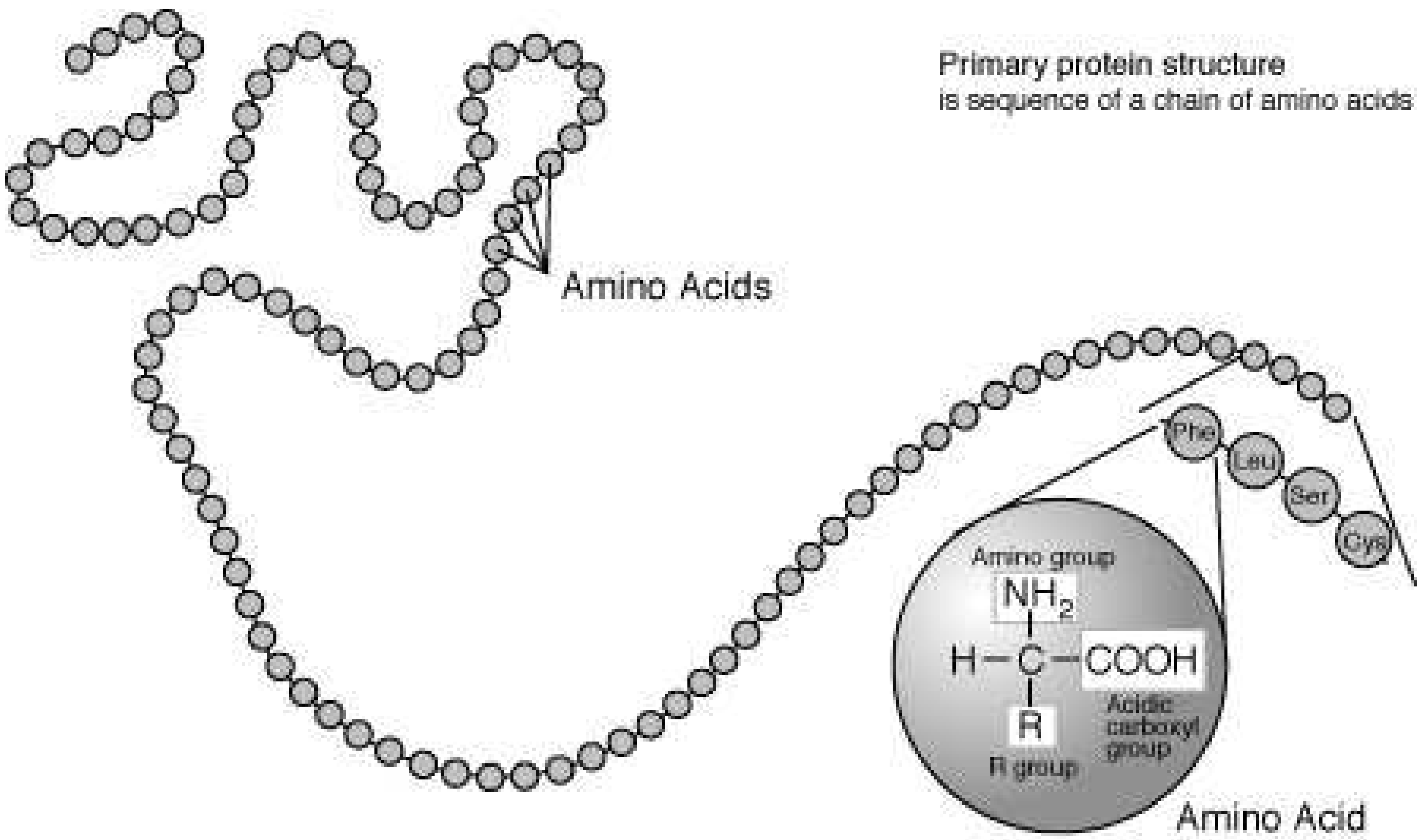
- Petites molécules ayant un role: ATP, $NADPH$ stockent l'energie
- Sucres, lipides (sources d'energie, structure des membranes)
- *Acides aminés* et *nucléotides*, qui sont les blocs de base pour former les *protéines* et l'*ADN/ARN*.

Protéines

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

- *Structural* : ex: le collagene relie les os et les tissus
- *Catalytique*: les *enzymes* catalysent une multitude de réactions biochimique (formant le métabolisme). Ex: la bexokinase 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 systeme nerveux...

Protéine = polymere d'acides aminés



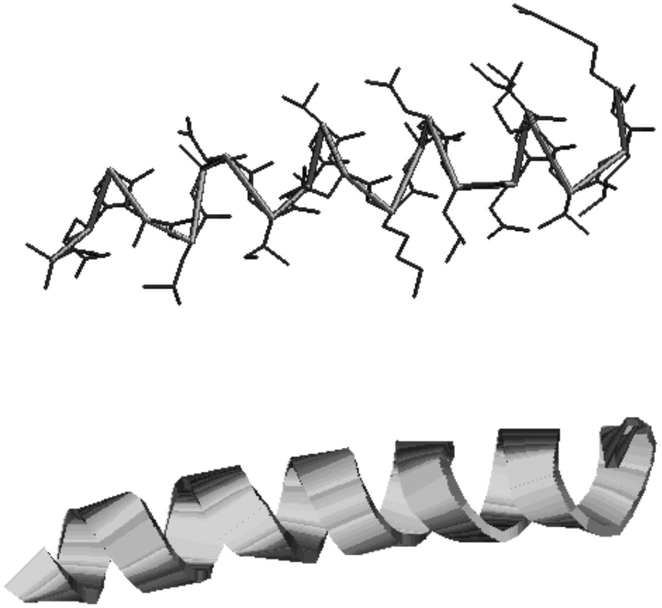
Strucure primaire

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

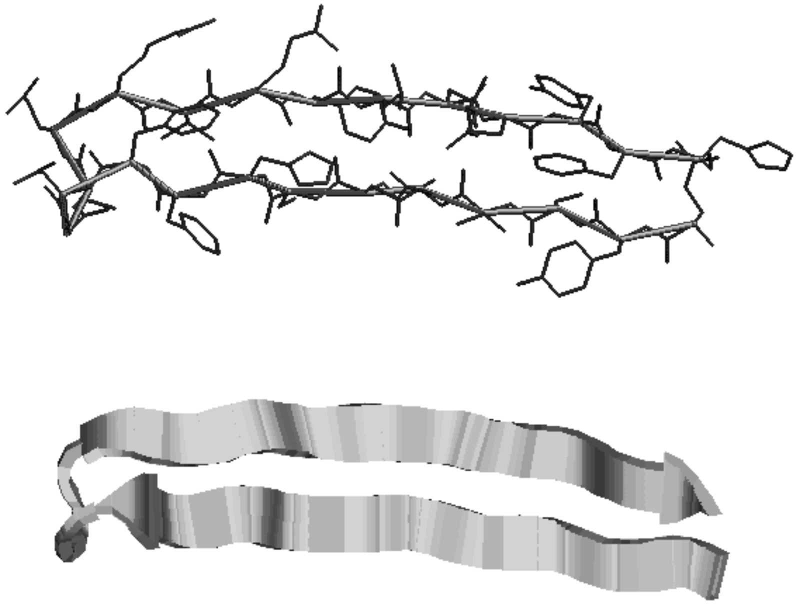
`FVNQHLCGSHLVEALYLVCGERGFFYTPKA`

Structure secondaire

Hélice α

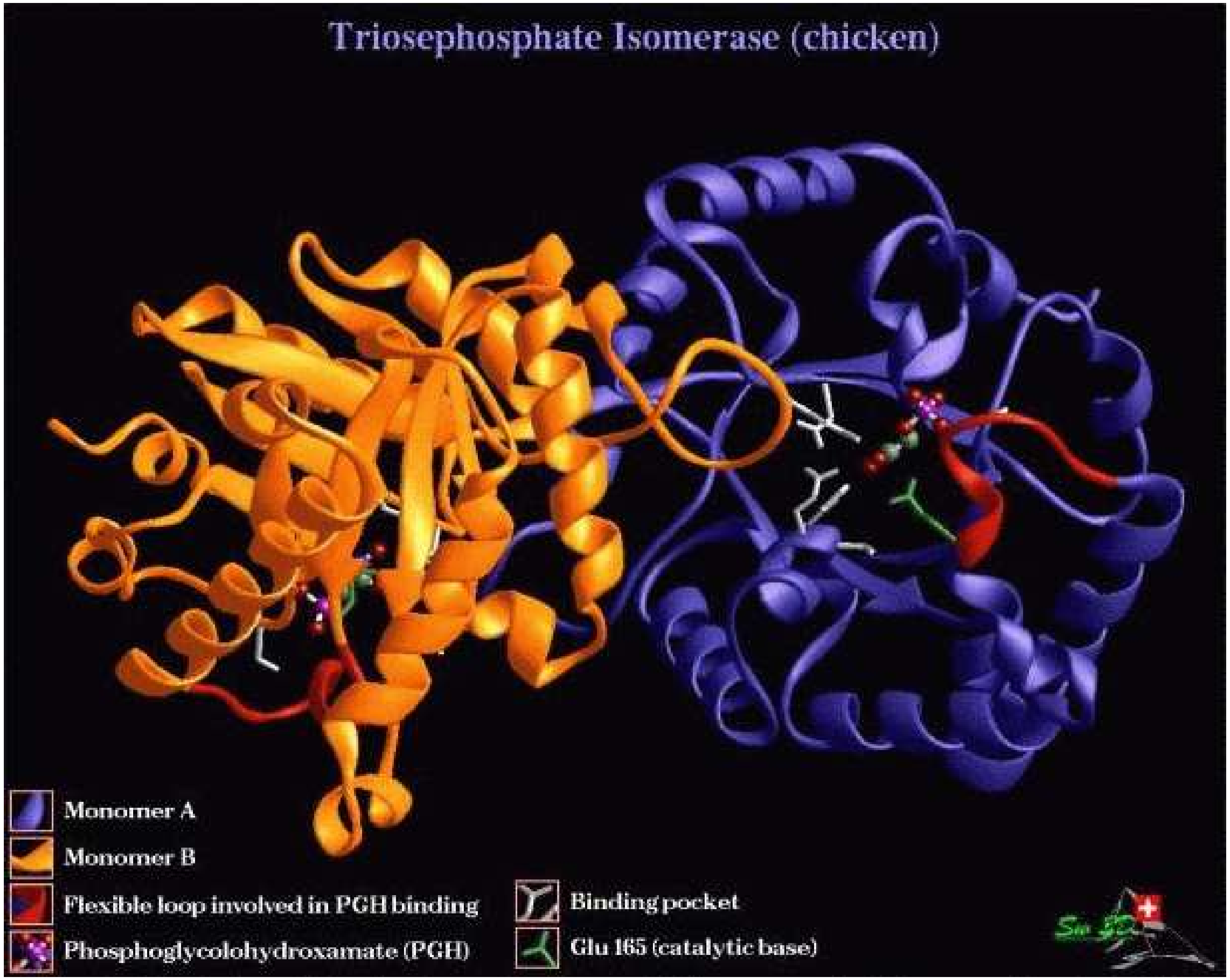


Feuillet β



Structure tertiaire

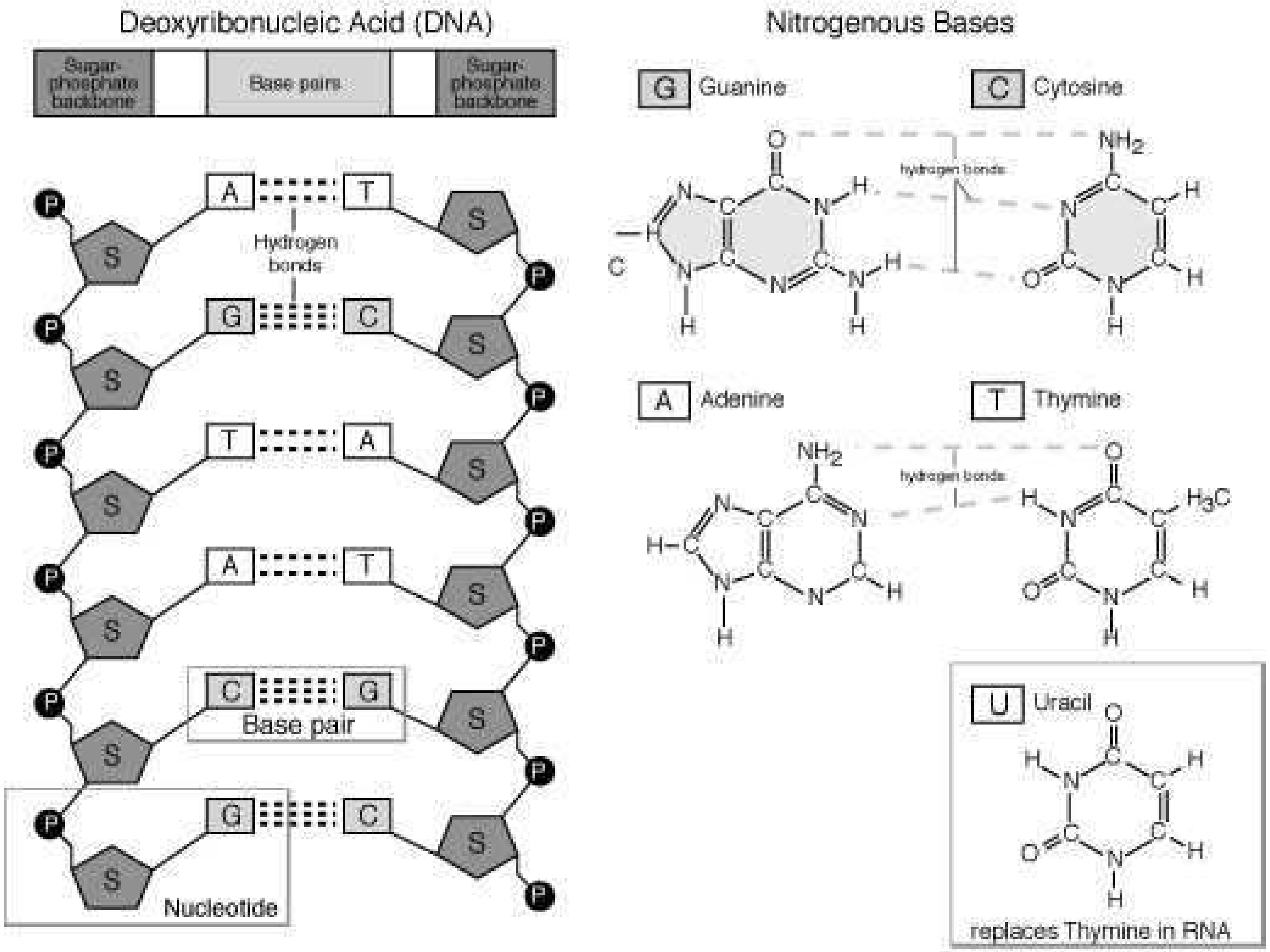
Structure quaternaire



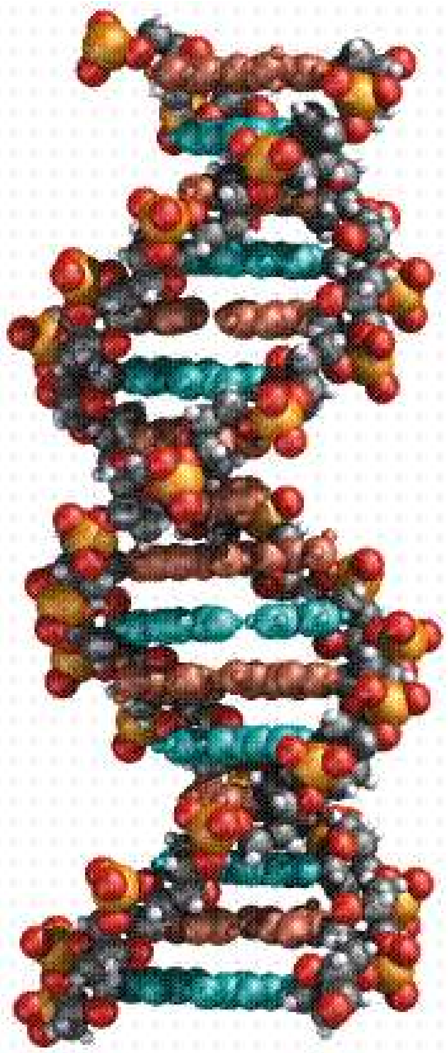
ADN

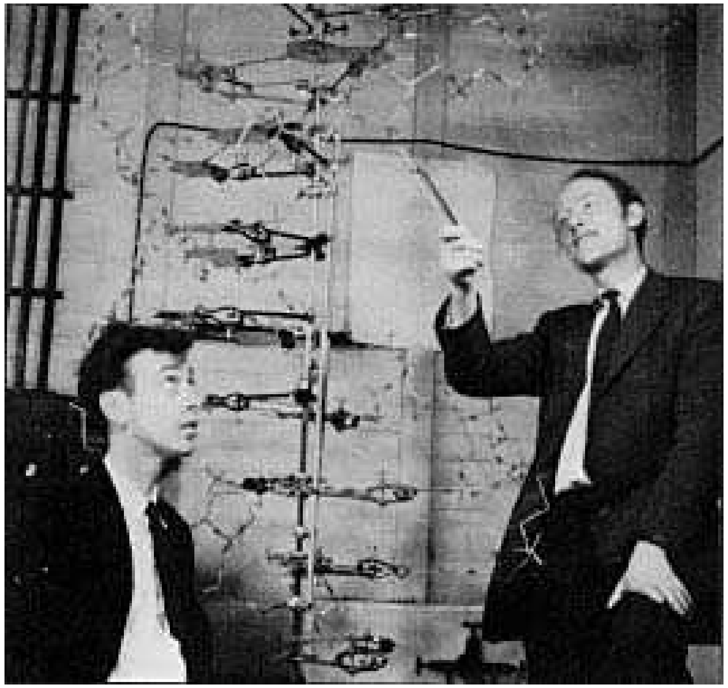
- L'*acide desoxyribonucléique* (ADN) est la molécule, présente dans toutes les cellules, qui contient l'information génétique transmise entre génération.
- L'ADN peut être en *simple brin* ou *double brin*.
- Un brin simple (aussi appelé polynucléotide) est un polymere linéaire composé de 4 *nucléotides*: adénosine (A), cytosine (C), guanine (G) et thymine (T)
- On représente un polynucléotide par une séquence orientée de lettre:
  `5′ -A-T-T-C-A-G-G-C-A-T-T-A-G-C- 3′`

ADN double brin



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

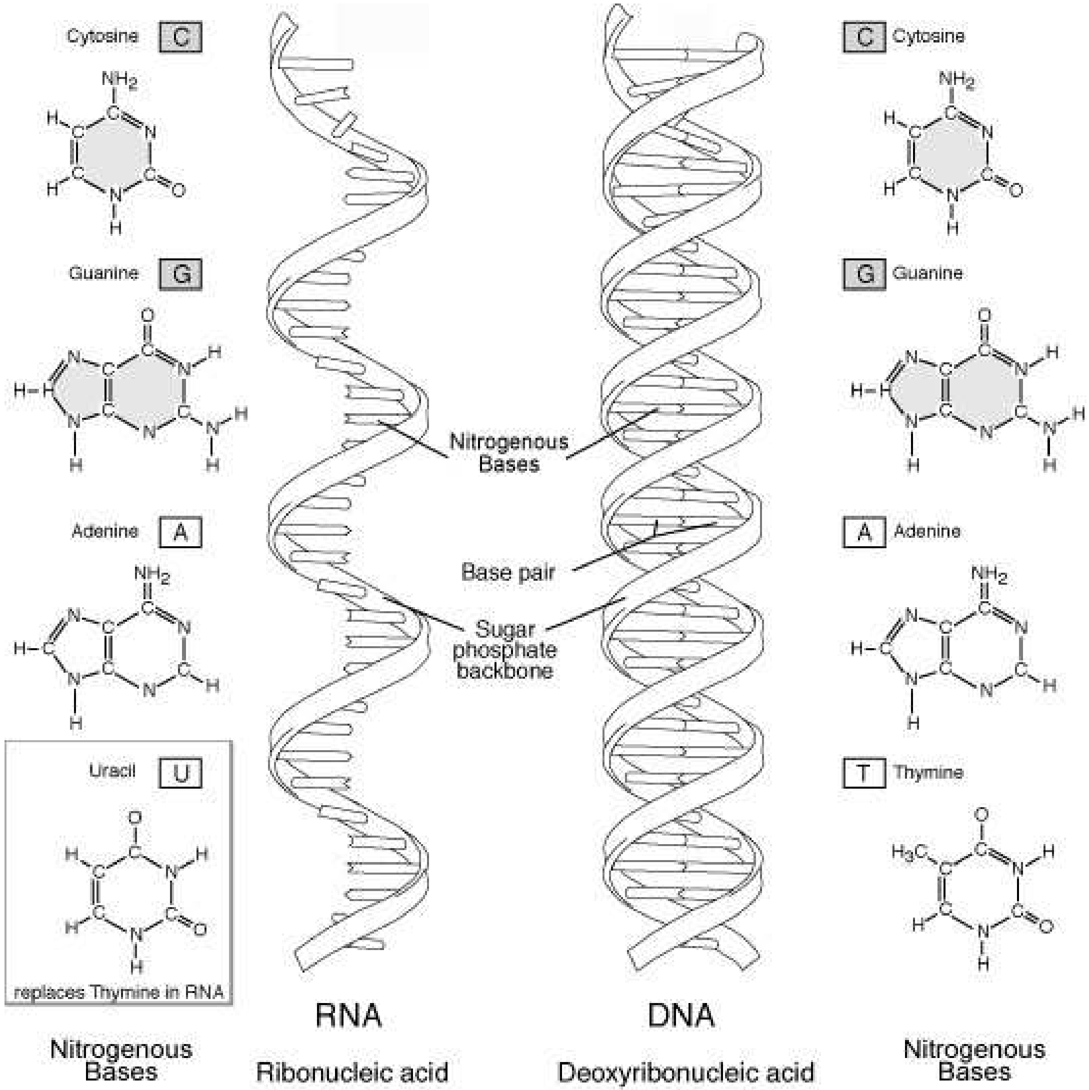




ADN et information

- 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 6.10^8 bits/cm = *75GB/cm*

ARN



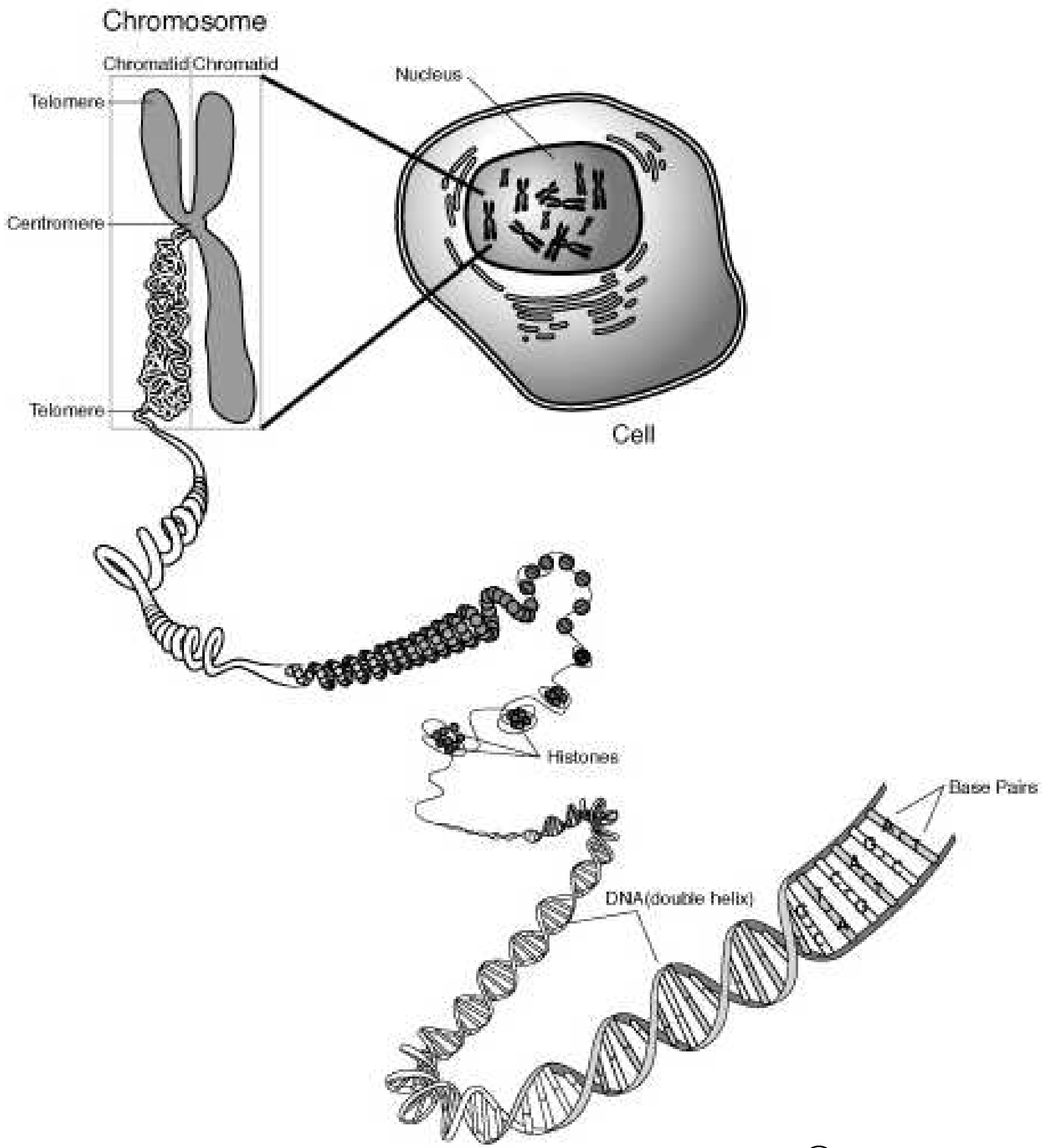
Cytosine
C
Guanine
G
Adenine
A
Uracil
U
replaces Thymine in RNA
Nitrogenous Bases
Nitrogenous Bases
Base pair
Sugar phosphate backbone
RNA
Ribonucleic acid
DNA
Deoxyribonucleic acid
C
Cytosine
G
Guanine
A
Adenine
T
Thymine
Nitrogenous Bases

ARN

- La thymine (T) de l'ADN est remplacée par *l'uracile (U)* dans l'ARN.
- A cause de cela, l'ARN ne forme pas de double-hélice, sauf si le brin complémentaire est de l'ADN complémentaire.

Genes et génomes

ADN et chromosomes

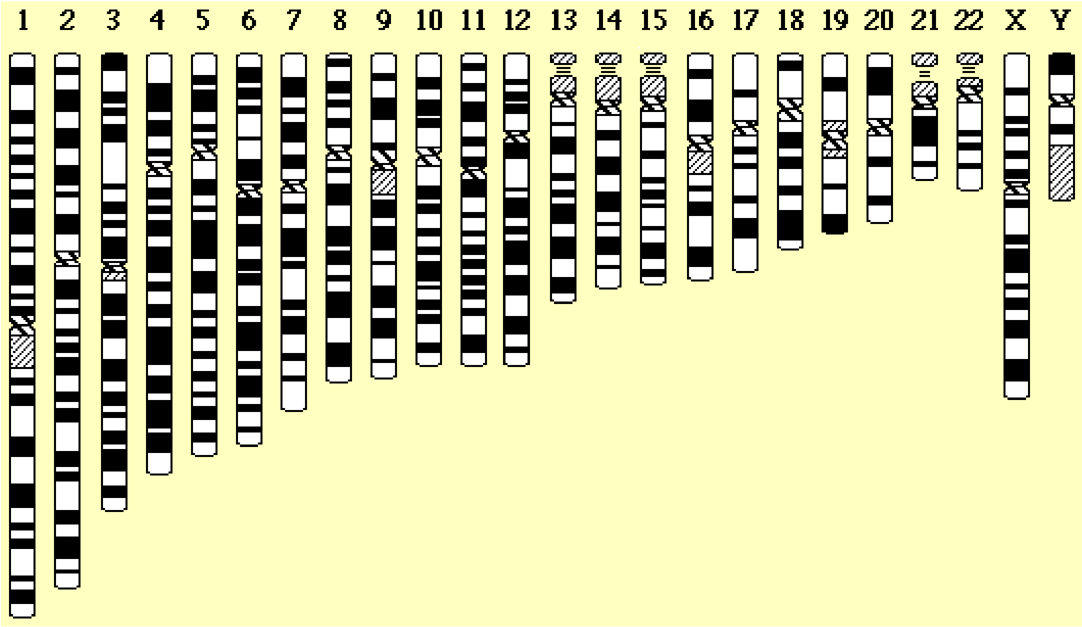


Génome

- Toutes les cellules d'un organisme ont (a peu pres) le meme ADN, appelé *génome*

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

Human genome



Séquencage

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

Gene

Une partie continue d'un brin d'ADN, a partir de laquelle une machinerie moleculaire complexe peut lire de l'information (encodée dans les lettes A,C,G,T) et créer une protéine particuliere

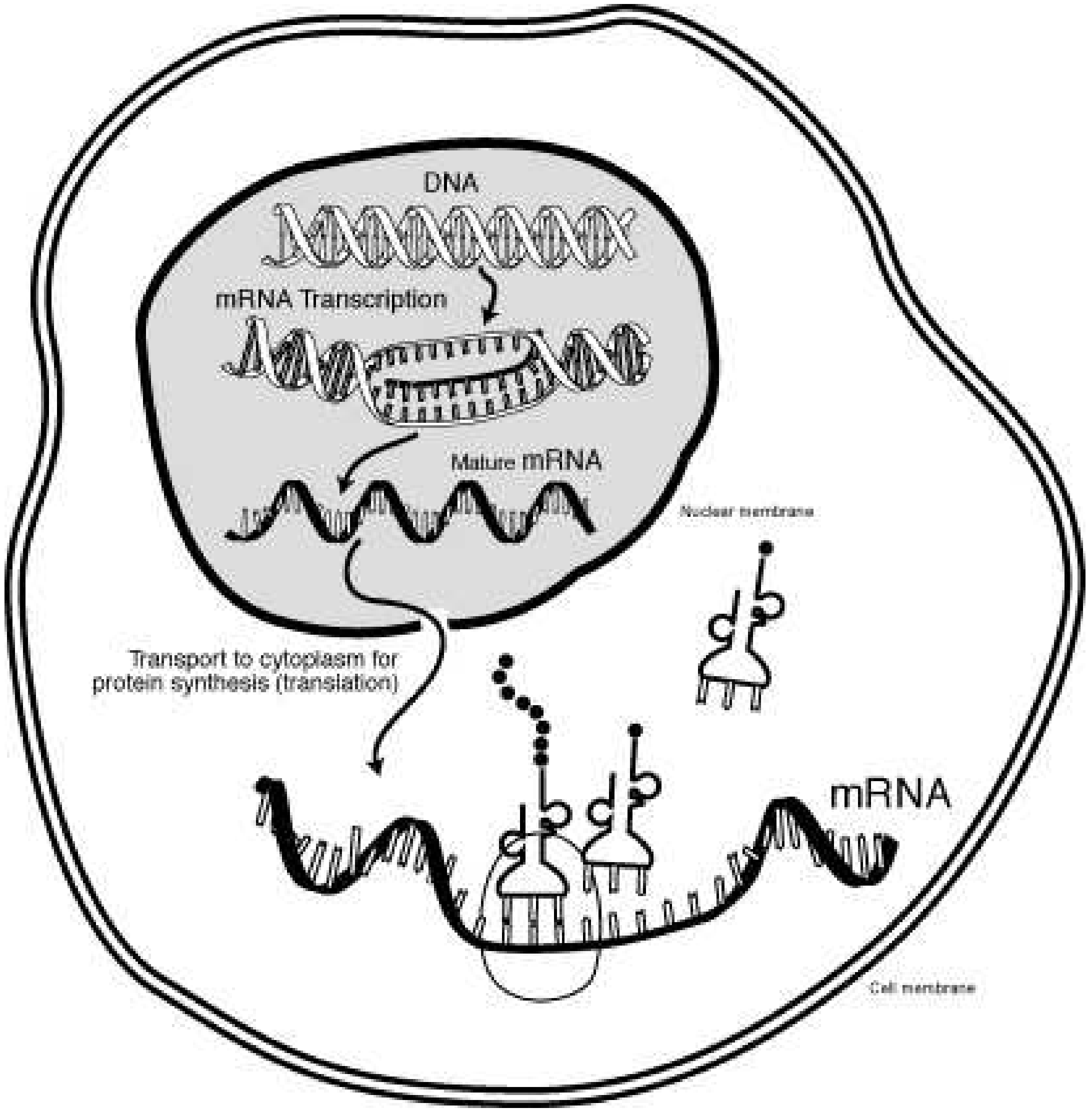
Dogme central

$$\text{DNA} \xrightarrow{\text{transcription}} \text{mRNA} \xrightarrow{\text{translation}} \text{Protein}$$

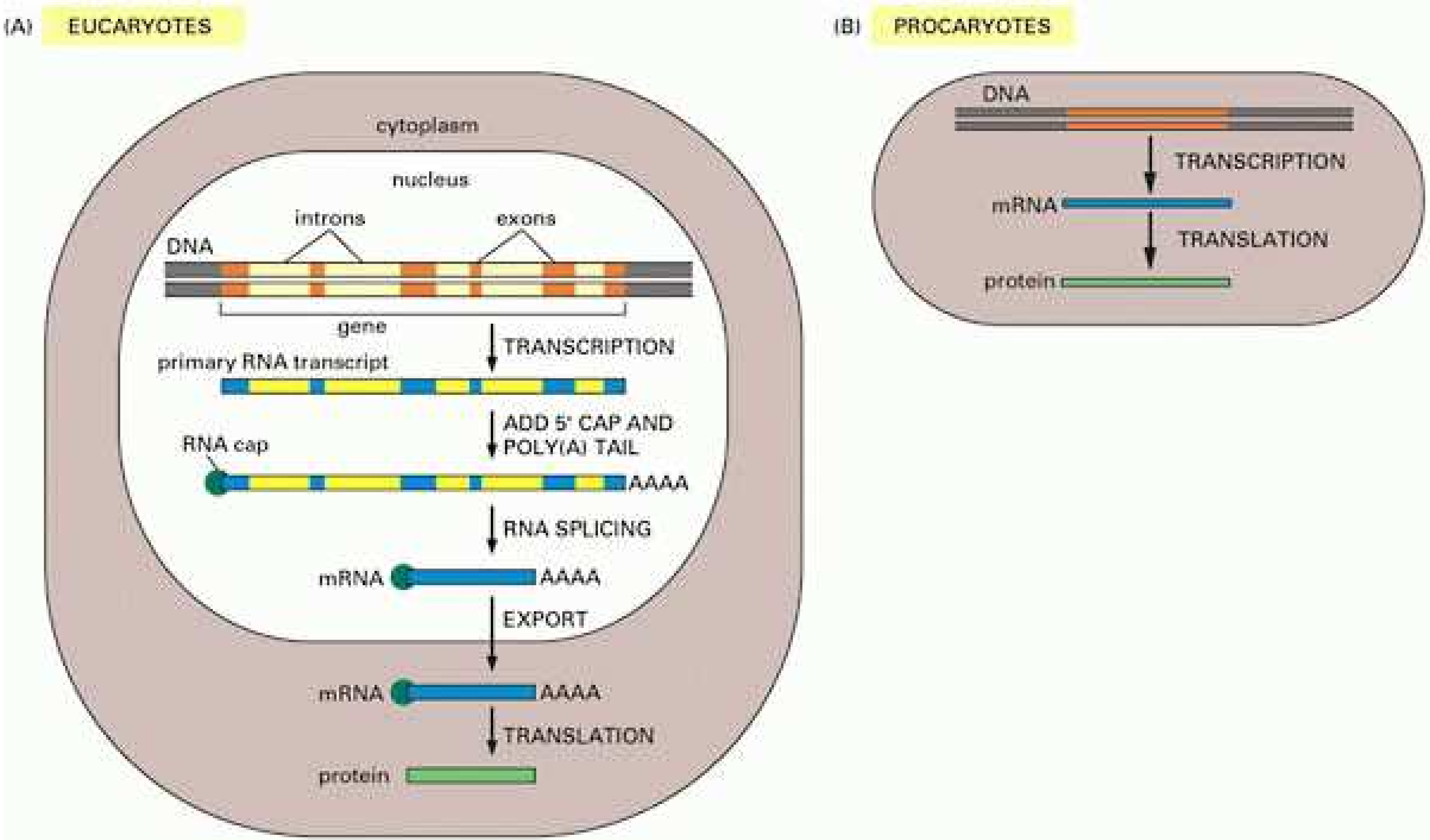
1 nucleotide to 1 nucleotide

1 codon (3 nucleotides) to 1 amino acid according to the ’genetic code’

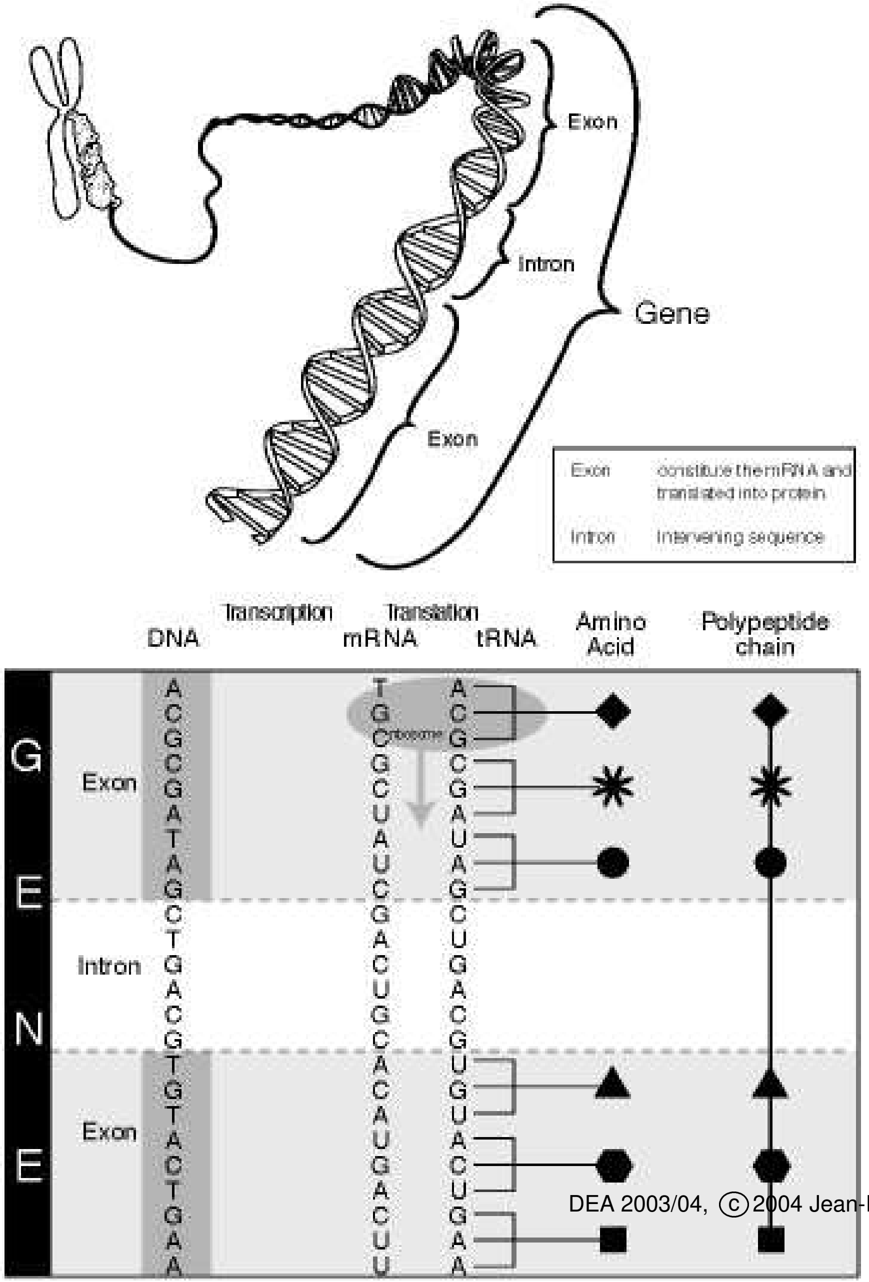
ARN messager



De l'ADN aux protéines

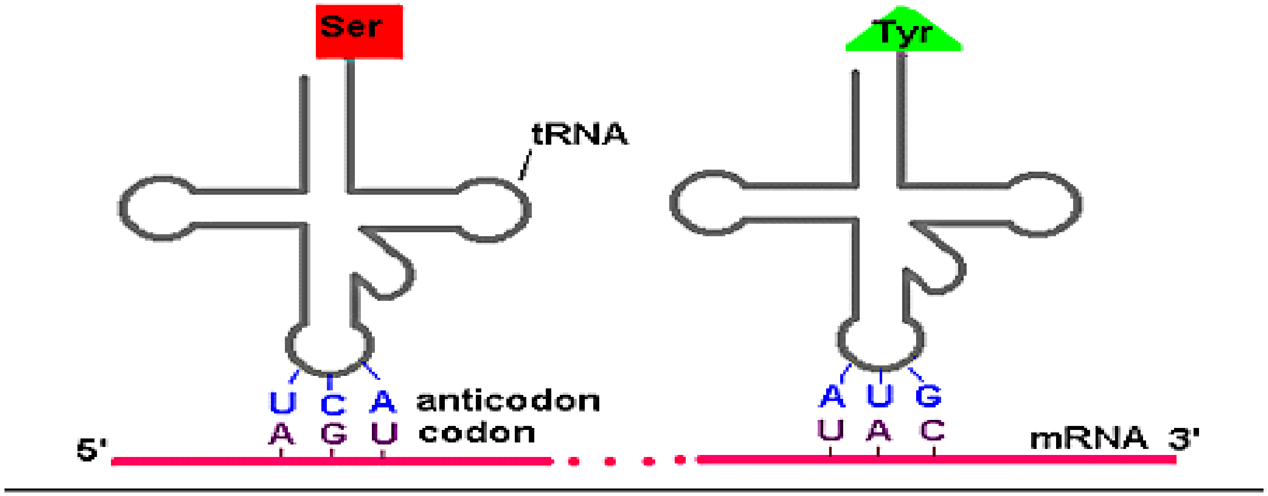


Genes



Exon
Intron
Gene
Exon
Exon
Intron
Intervening sequence
DNA
Transcription
mRNA
Translation
tRNA
Amino Acid
Polypeptide chain
GENE
Exon
Intron
Exon
ribosome

Code génétique

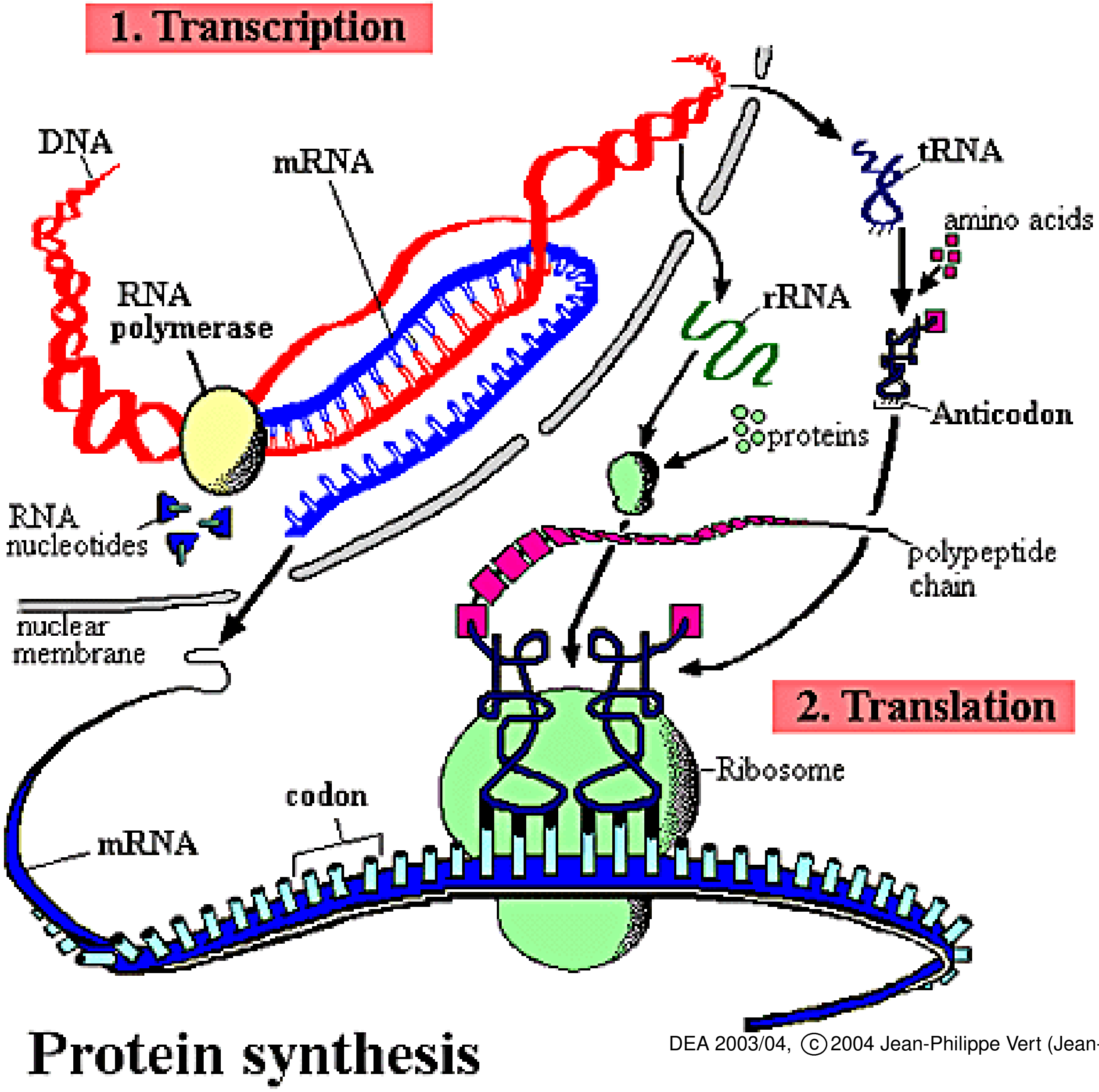


2nd base in codon

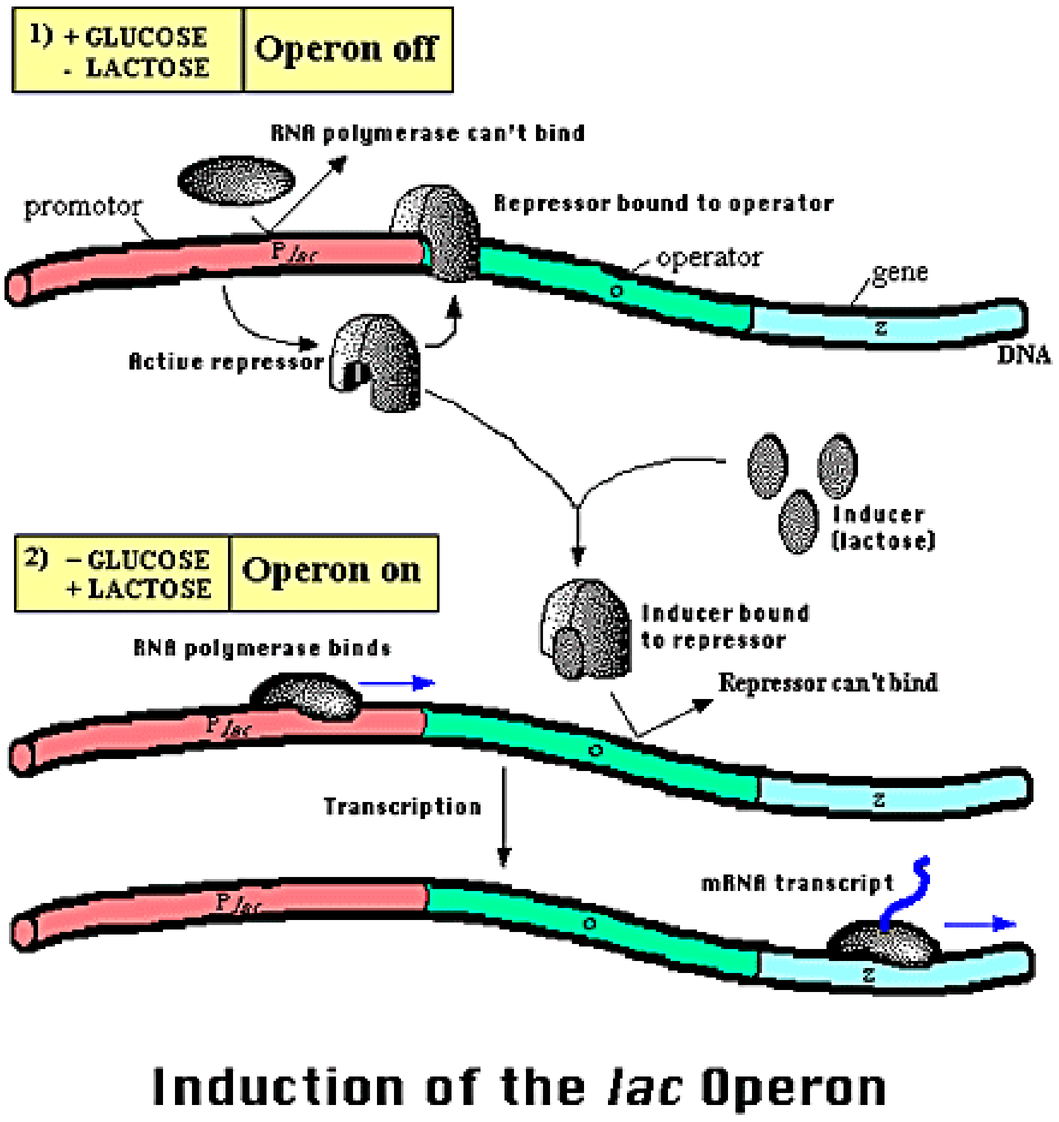
1st base in codon	U	C	A	G	3rd base in codon
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

The Genetic Code

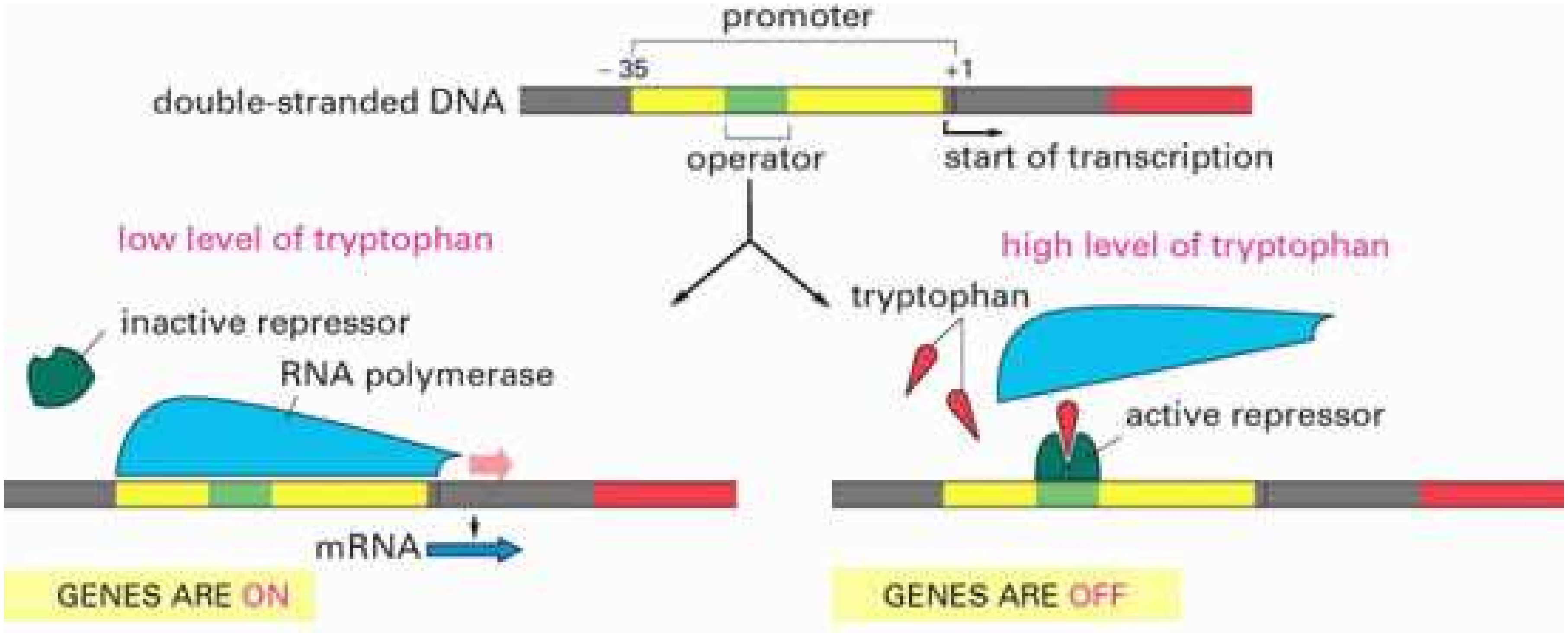
De l’ADN a la protéine



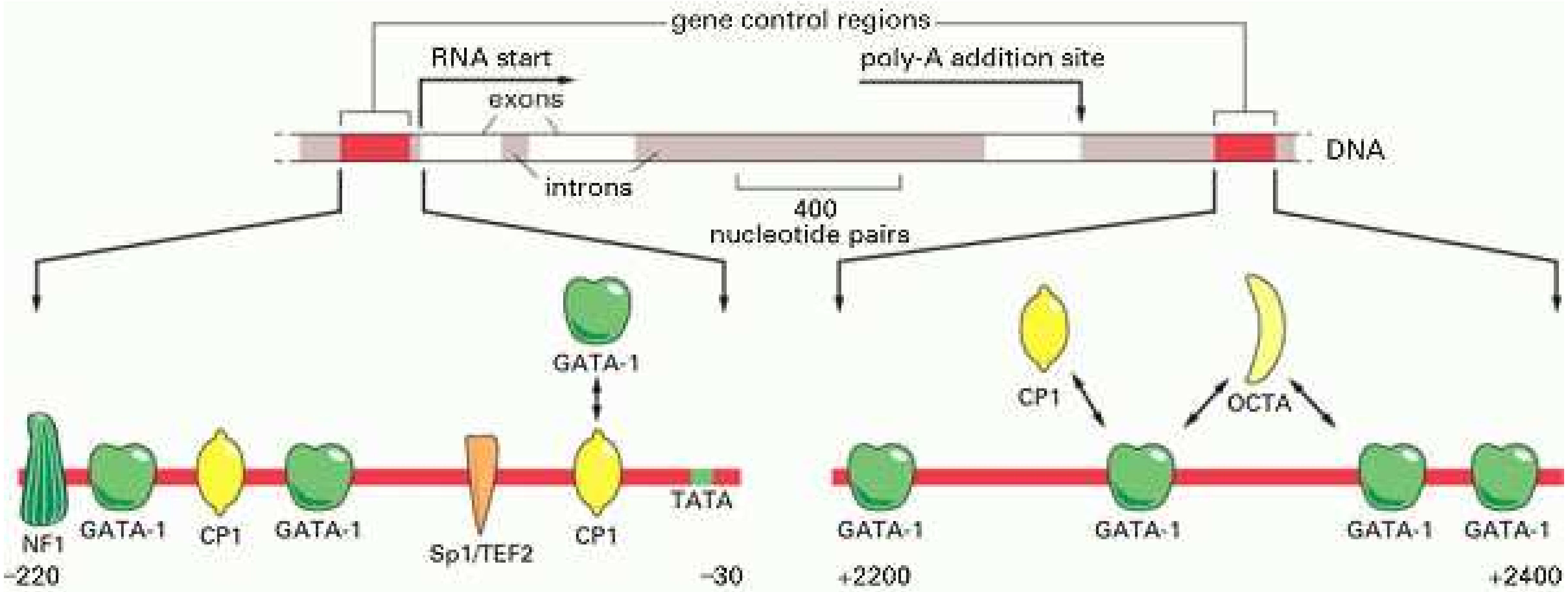
Controle de l'expression (induction)



Controle de l'expression (repression)

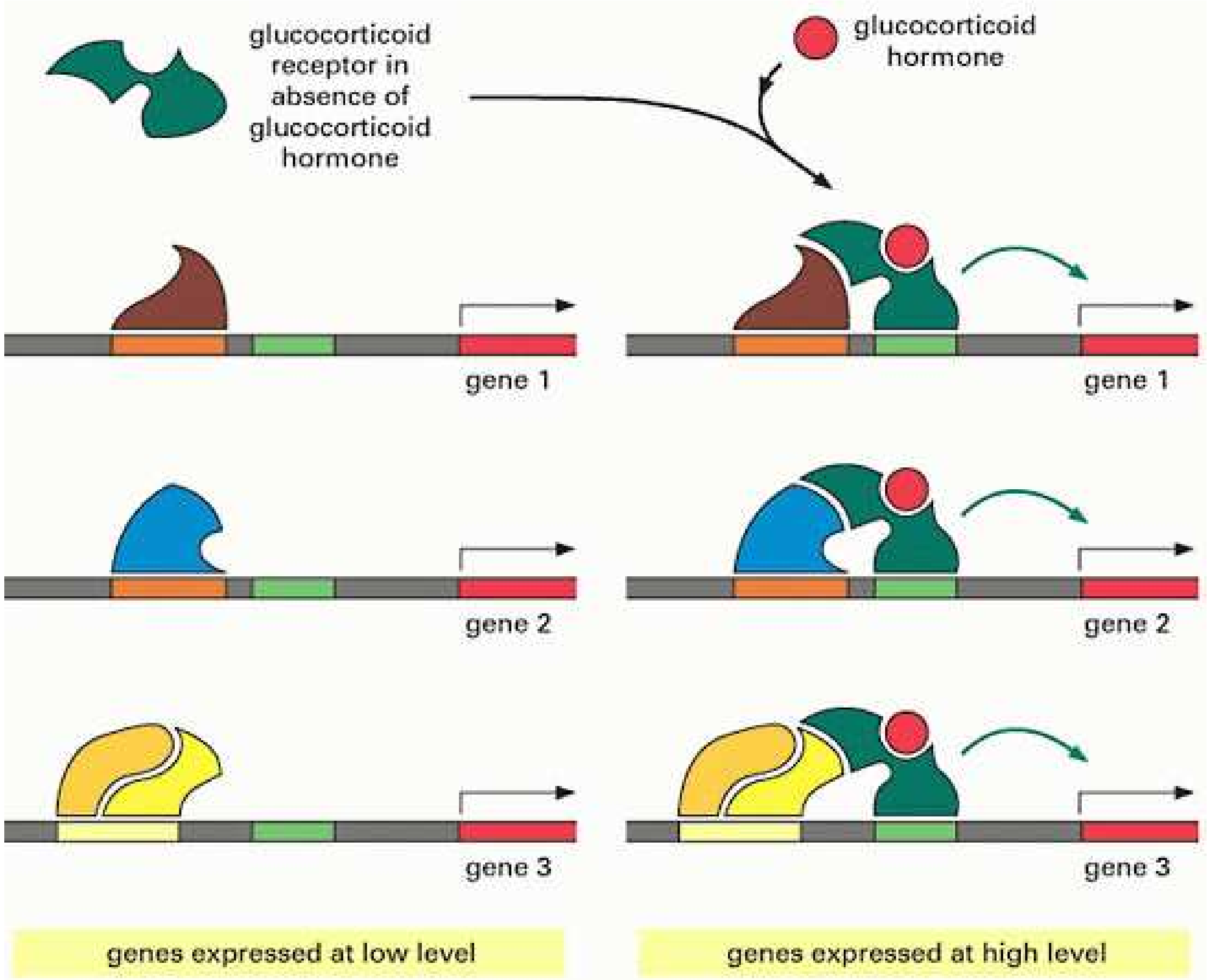


Exemple: B-globine humaine

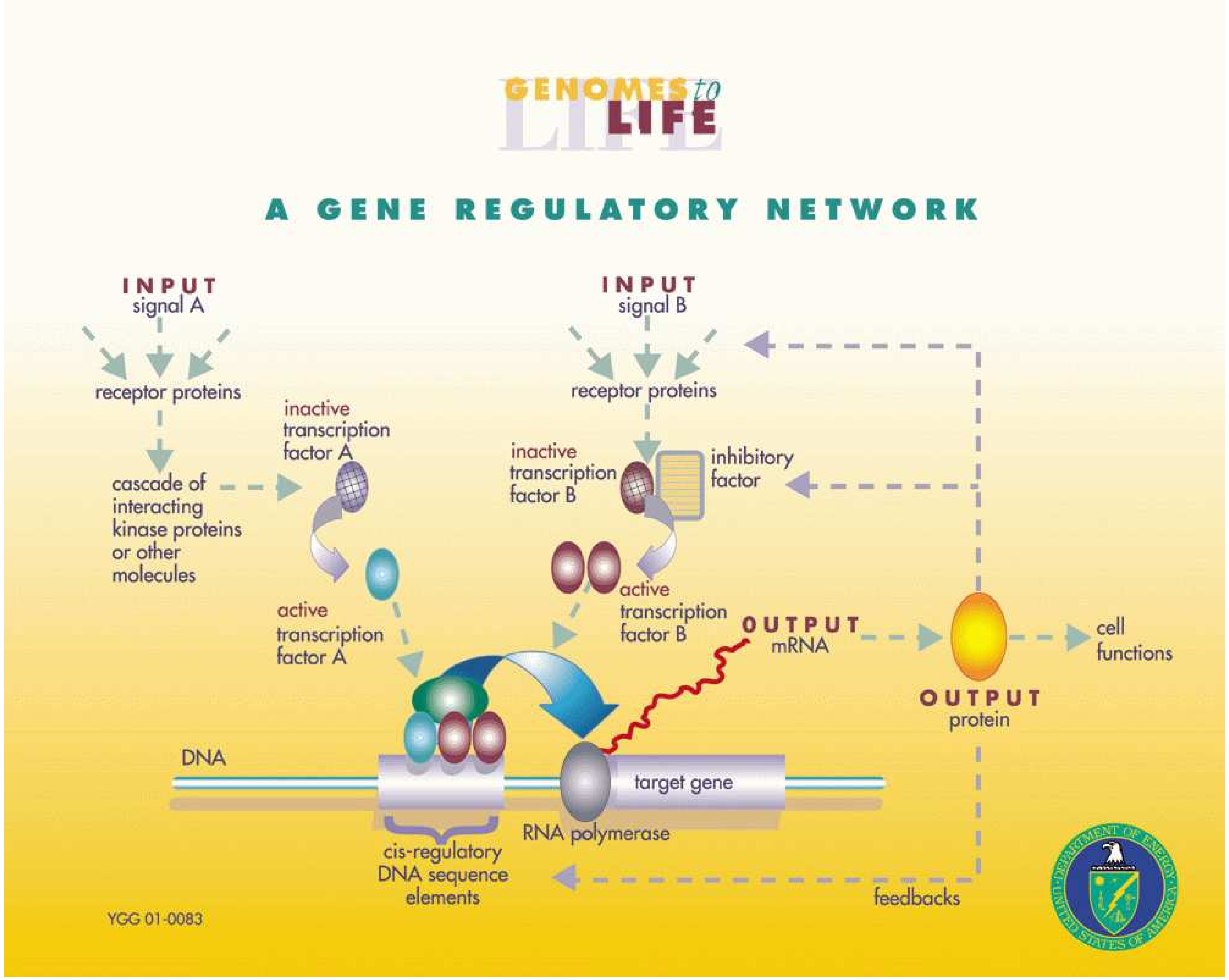


La B-globin jour un role important dans le developpement des cellules rouges du sang. Certaines protéines régulatrices, comme CP1, sont présentes dans de nombreuses cellules, mais d'autres, comme GATA-1, ne se trouvent que dans quelques types de cellules, dont les précurseurs des cellules rouges.

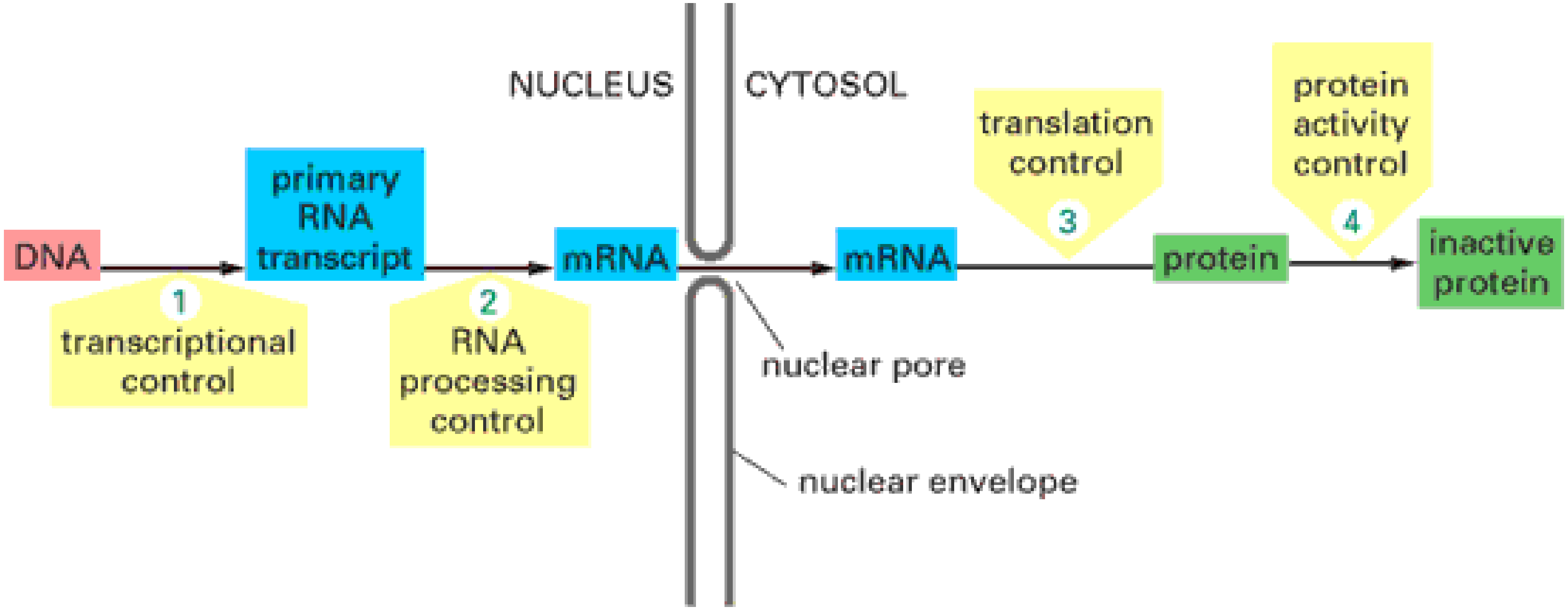
Coordination du controle



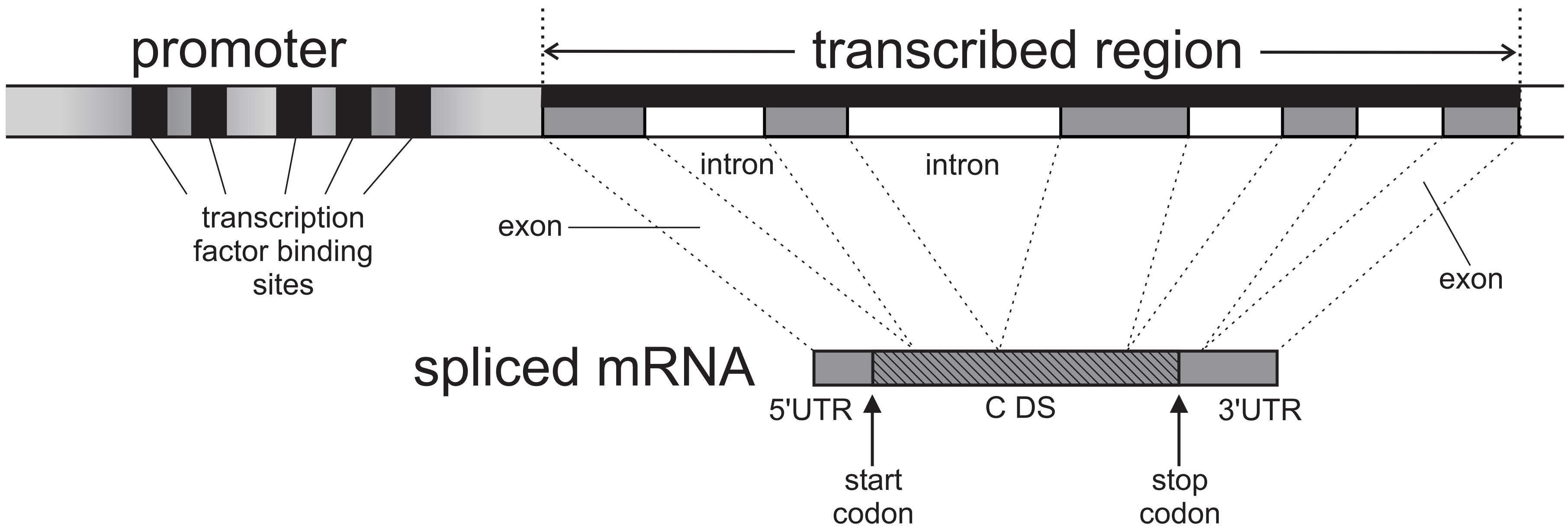
Réseau de régulation



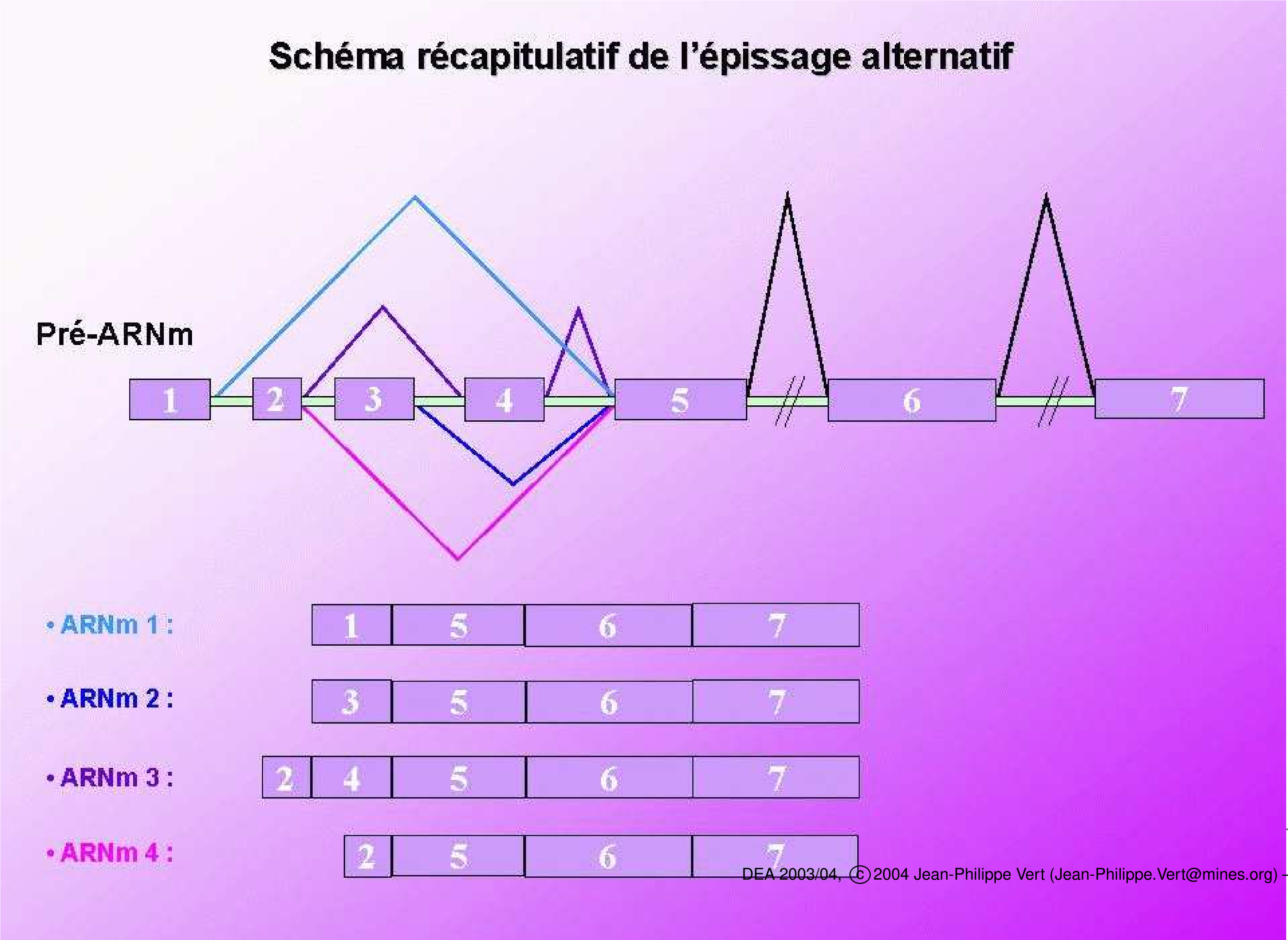
Autres controles: du gene a la protéine active



Epissage


promoter
transcribed region
intron
intron
transcription
factor binding
sites
exon
exon
spliced mRNA
5'UTR
C DS
3'UTR
start
codon
stop
codon

Epissage alternatif



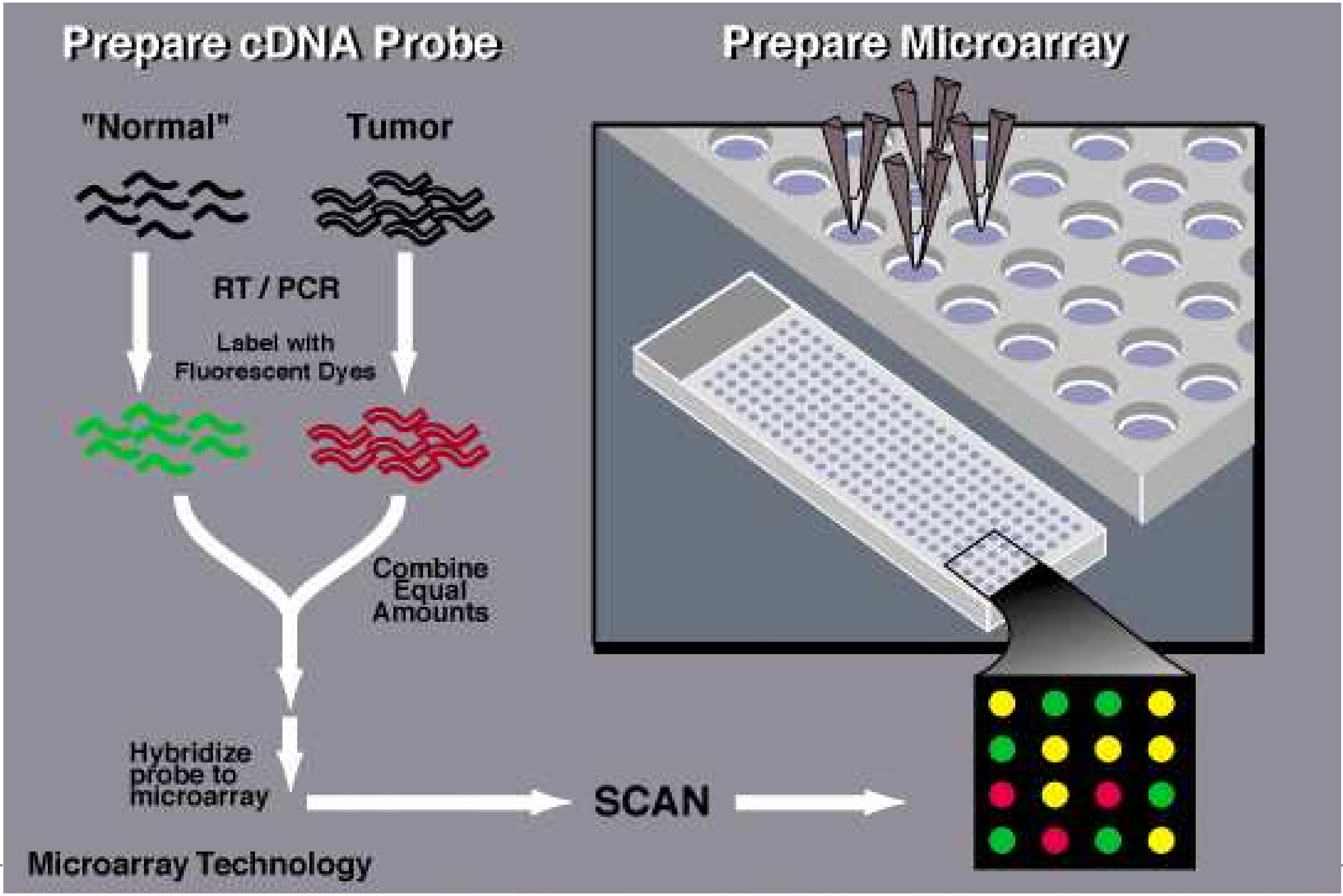
Nouveau dogme

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

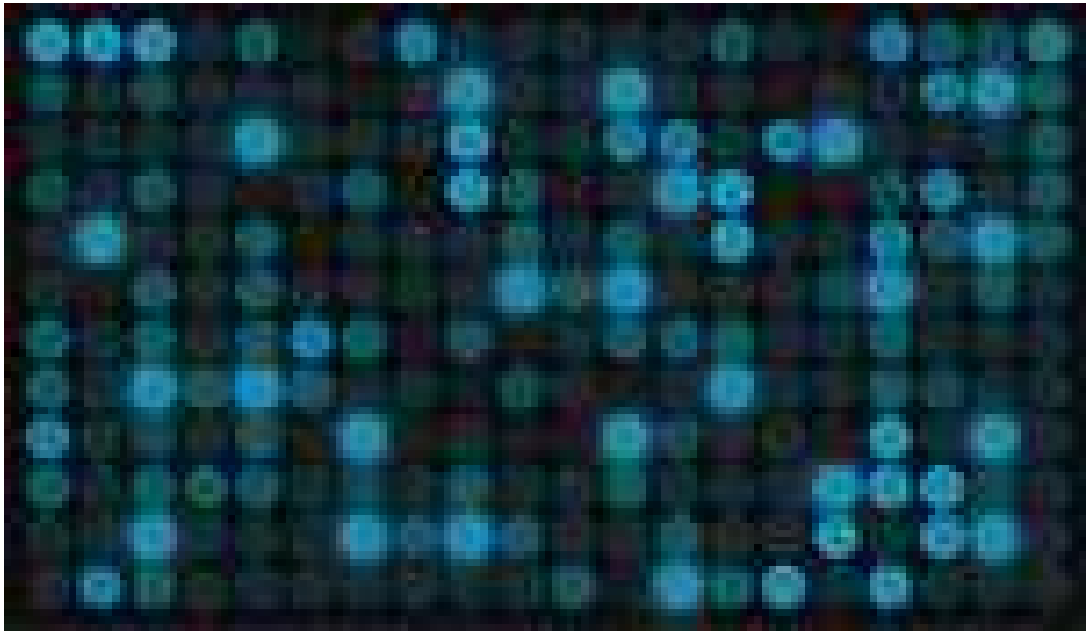
Technologies et données

Sequencer

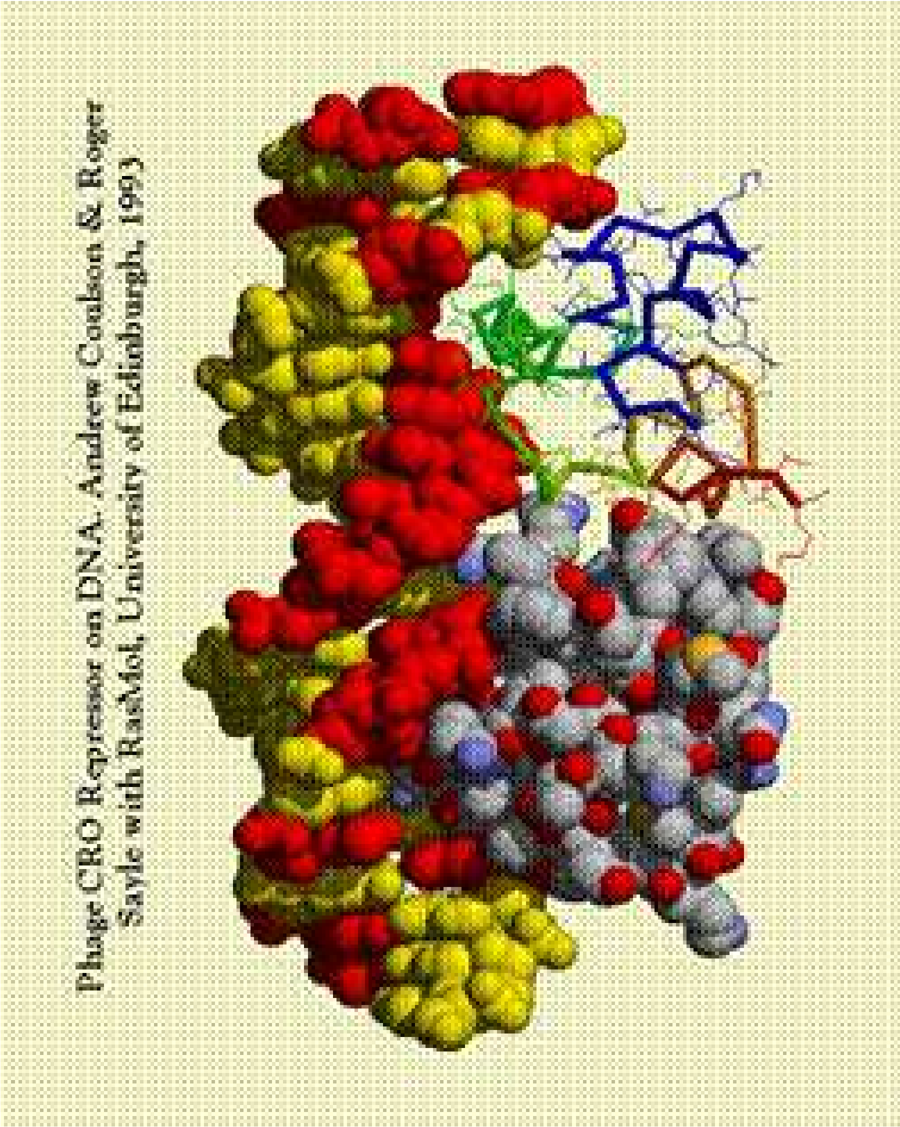
Microarrays



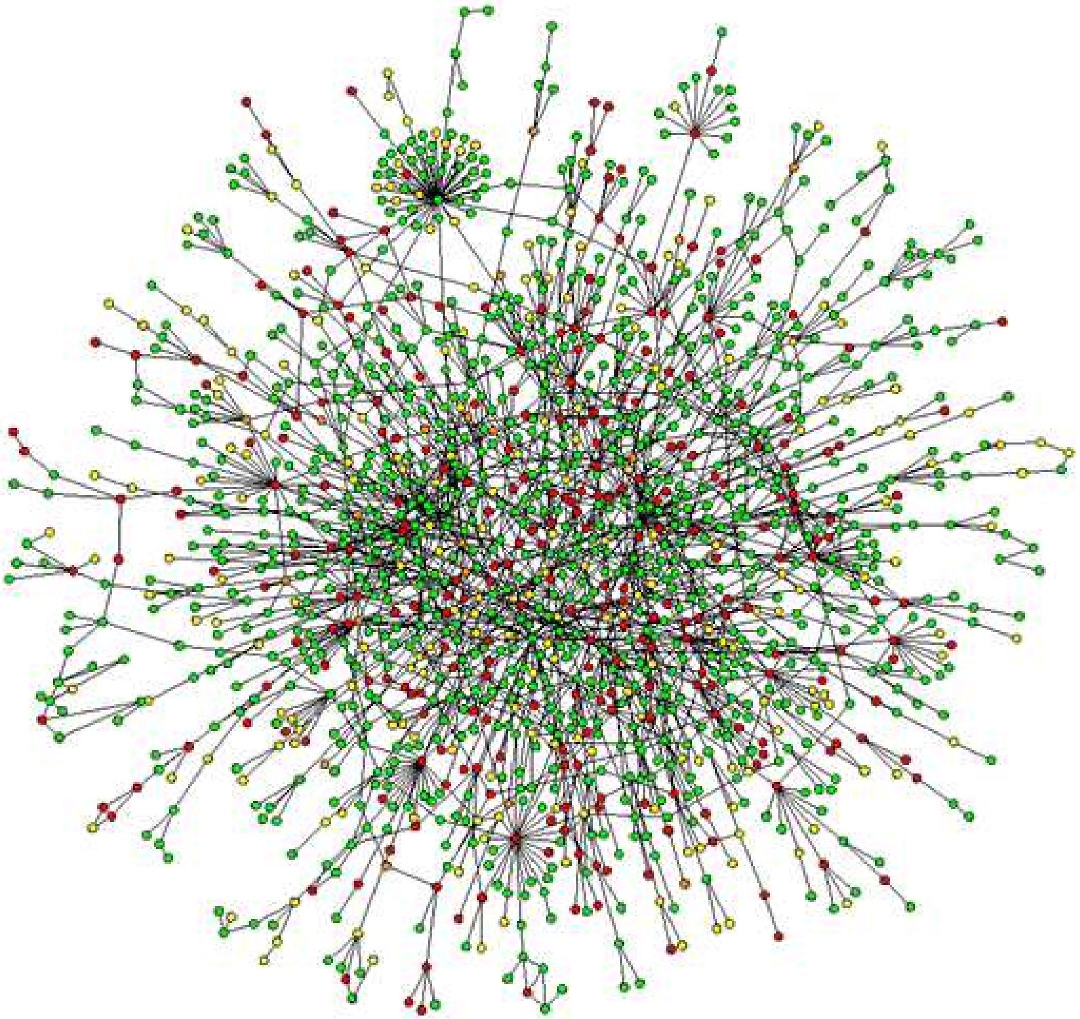
Transcriptome



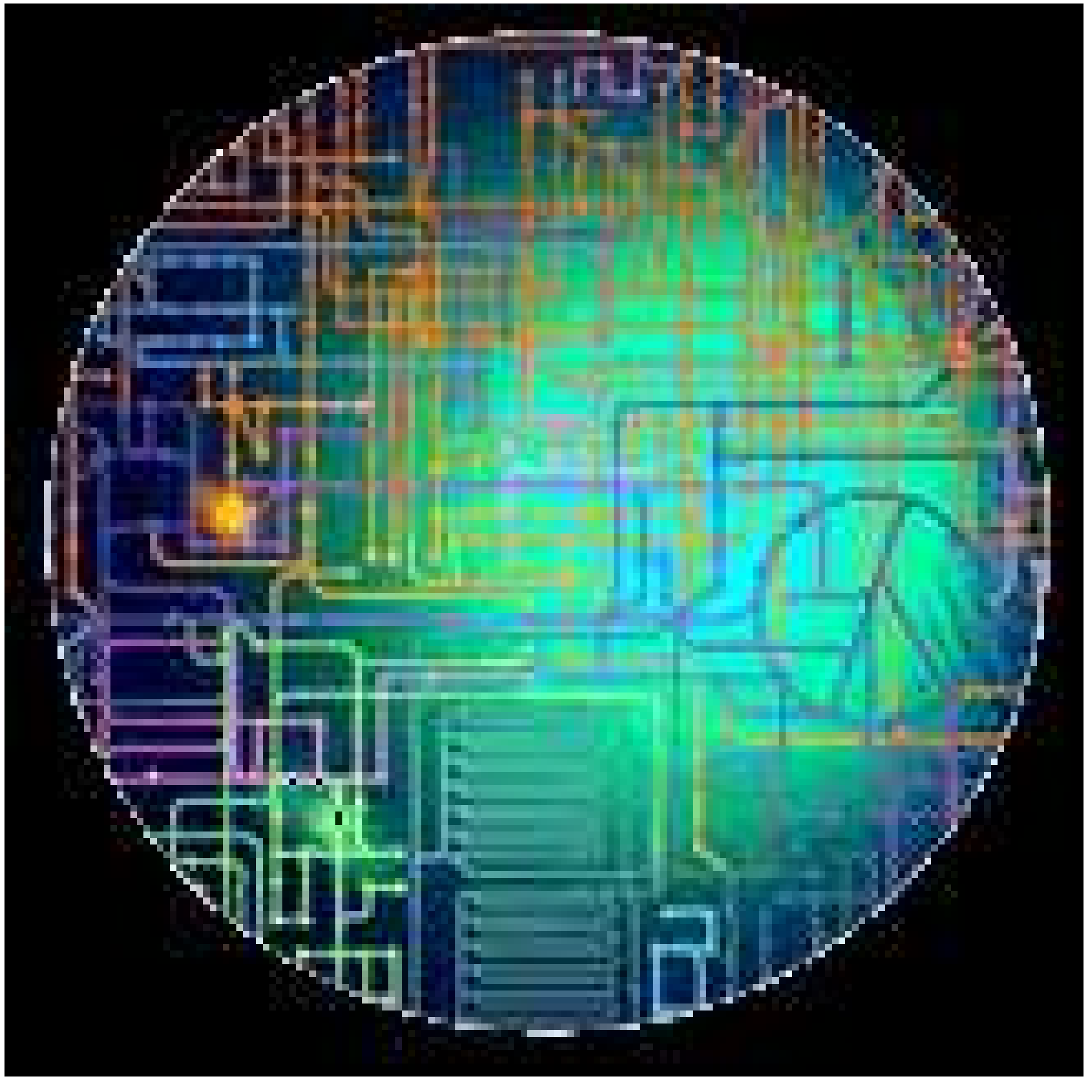
Protéome


Phage CRO Repressor on DNA. Andrew Coulson & Roger
Sayle with RasMol, University of Edinburgh, 1993

Interactome



Metabolome



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	
- **proteins with metabolites**: receptors or enzymes binding ligands	real vectors (binding energies)
- **proteins with DNA**: transcription factors; etc.	binary (bipartite graph)
- **proteins with proteins**: complexes; etc.	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
Protein 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
- ExPASy	`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`	*arabidopsis*
- WormBase	`www.wormbase.org`	*C. elegans*

Protein properties sources

Database	URL (`http://...`)	Remark
Protein 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

Protein properties sources

Molecular interactions and networks

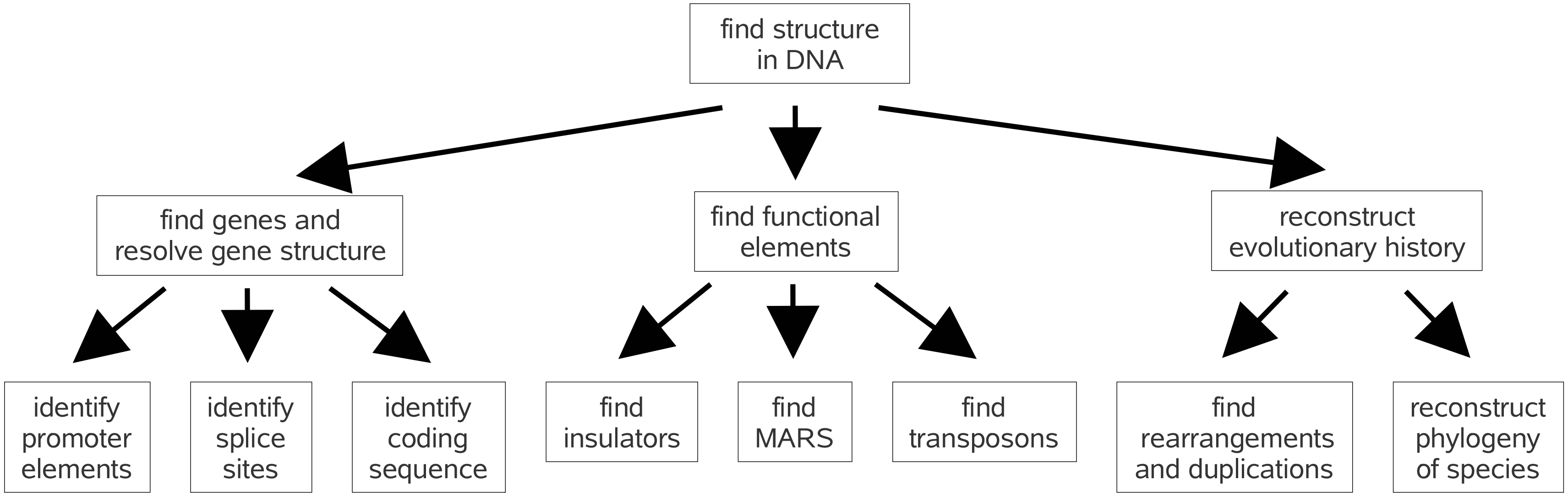
- BIND	`www.bind.ca`	interaction network
- KEGG	`www.genome.ad.jp/kegg/`	metabolic pathways
- DIP	`dip.doe-mbi.ucla.edu`	interacting proteins

Protein properties sources

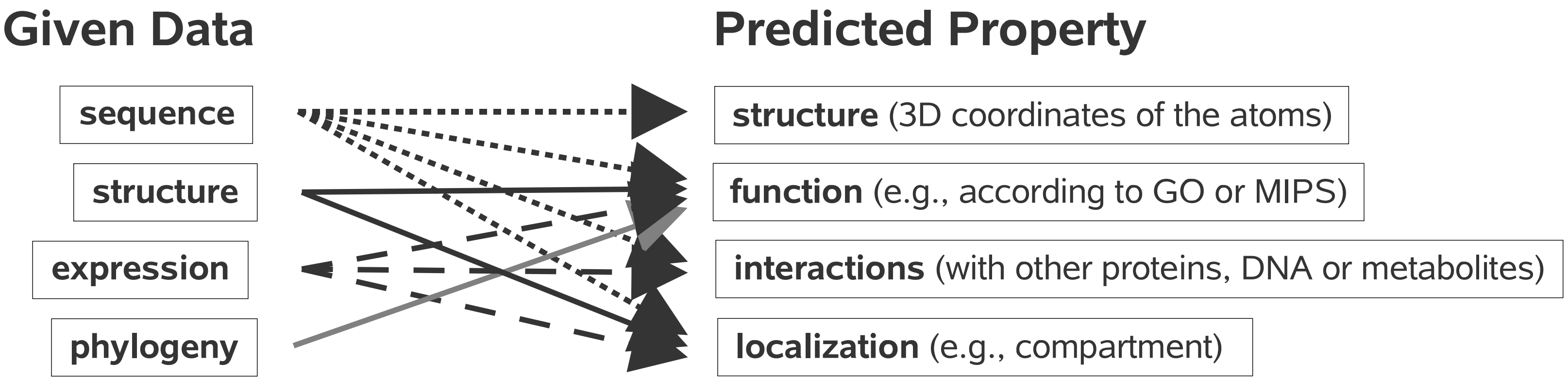
Protein functions		
- GO	`www.geneontology.org`	controlled vocabulary
- EC	`www.chem.qmul.ac.uk/iubmb/enzyme/`	enzyme numbers
- MIPS	`mips.gsf.de/proj/yeast/catalogs/funcat/`	yeast gene functions
Protein expression		
- 2DPAGE	`us.expasy.org/ch2d/`	2D gel electrophoresis data

Challenge en bio-informatique

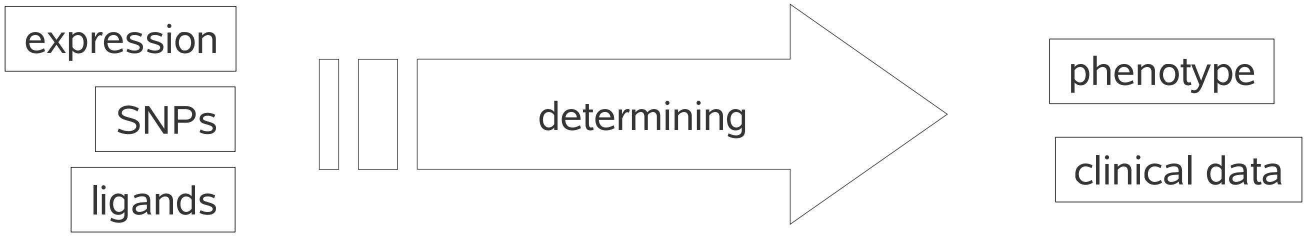
Génomique


find structure in DNA
find genes and resolve gene structure
find functional elements
reconstruct evolutionary history
identify promoter elements
identify splice sites
identify coding sequence
find insulators
find MARS
find transposons
find rearrangements and duplications
reconstruct phylogeny of species

Protéomique



Pharmacogénomique



Caractéristiques

- *Beaucoup* de données...
- mais *beaucoup* de "bruit"
- Données hétérogenes (séquences, structures, vecteurs, graphes...)
- "Small n large p"
- problemes 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* a la biologie moléculaire et la génomique
- *Noyaux positifs*: définition, propriétés, espaces de Hilbert a noyau reproduisant, kernel trick, representer theorem
- *Méthodes a 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. Schoelkops, 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.h`
- .. et quelques images trouvées sur le web