

Introduction à la bioinformatique (UE SSV3U15)

Chapitre 3. Du gène au génome

Jacques van Helden (Aix-Marseille Université)

ORCID [0000-0002-8799-8584](https://orcid.org/0000-0002-8799-8584)

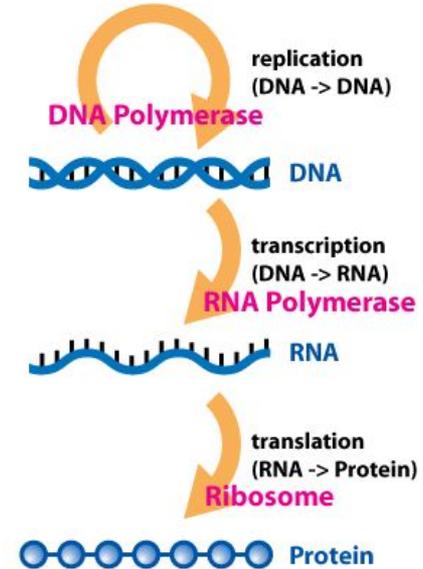
1. Les voies de l'information génétique
2. Structure d'un gène
3. Disponibilité des génomes
4. Composition et organisation des génomes
5. Annotation des génomes : où sont les gènes ?
6. Annotation des génomes : que font les gènes ?
 - Assignation de fonction par similarité de séquences
 - Un élément structurant des génomes: la régulation
 - Génomique comparative
 - Coupable par association
 - La Gene Ontology – Définir et structurer les termes d'annotation des gènes et de leurs produits

Les voies de l'information génétique

DNA makes RNA makes protein

L'ADN est le support de l'information génétique, et ceci de deux façons

- **Hérédité**, via la réplication
- **Information fonctionnelle**
 - ADN → [transcription] → ARN
 - ARN → [traduction] → protéine
- Les protéines et certains ARN sont les effecteurs moléculaires des fonctions biologiques

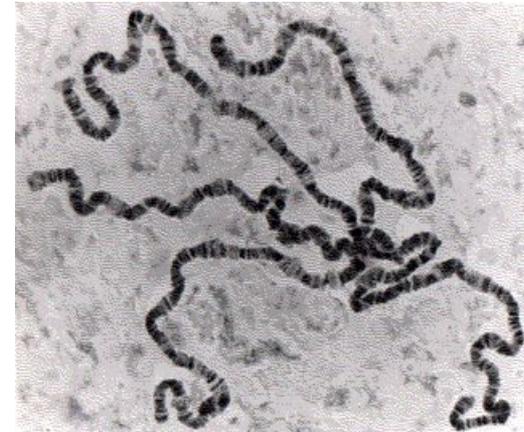


Les chromosomes constituent le support physique de l'hérédité

- En 1915, dans un livre intitulé Mécanismes de l'hérédité Mendélienne, Thomas Hunt Morgan formule la théorie chromosomique de l'hérédité.
- Ses observations
 - Les 4 groupes de liaison génétiques de la drosophile correspondent aux 4 chromosomes.
 - Les chromosomes sont porteurs des caractères transmis de façon héréditaire.
 - Sur chaque chromosome, les gènes sont ordonnés de façon linéaire.
- Il en déduit que les chromosomes sont le support physiques des caractères héréditaires.



http://news.bbc.co.uk/olmedia/440000/images/_443673_drosogene.jpg



<http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/P/Polytene.jpg>

Caryotype humain

Chez l'humain, les noyaux des cellules somatiques comportent 23 paires de chromosomes.

Les cellules somatiques sont diploïdes: chaque cellule comporte 2 copies de chaque chromosome (1 maternelle et 1 paternelle).

Photo de chromosomes étalés



Ces mêmes chromosomes regroupés pour mettre en évidence les paires homologues



Les chromosomes sont essentiellement composés d'ADN

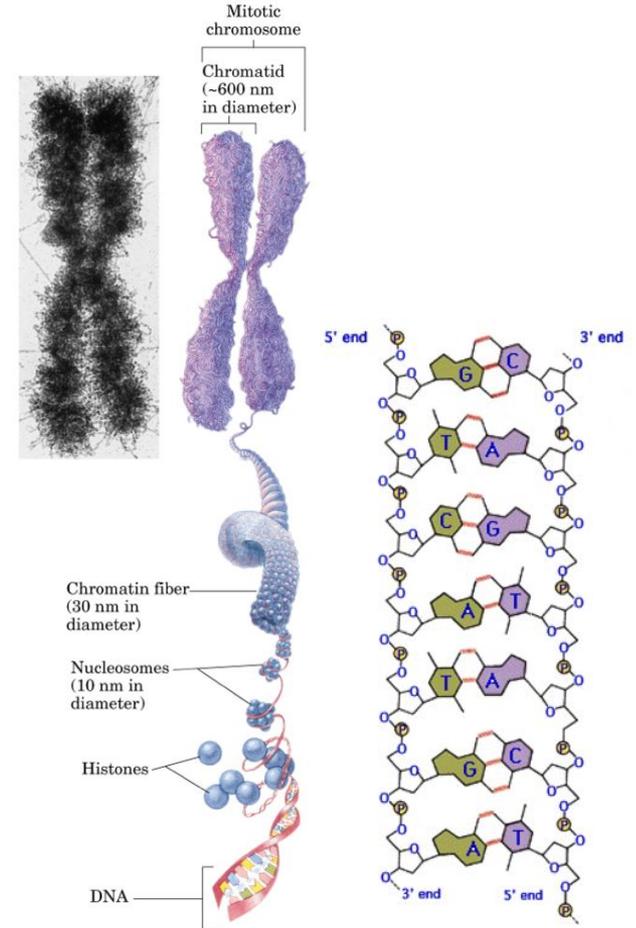
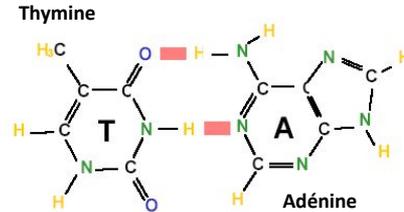
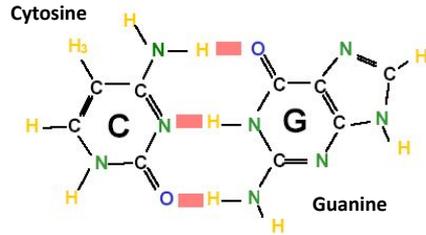
Chaque chromosome contient une chaîne extrêmement longue d'acide désoxyribonucléique (ADN).

L'ADN est composé d'une double hélice, qui porte 4 types de bases azotées.

- A Adénine
- C Cytosine
- G Guanine
- T Thymine

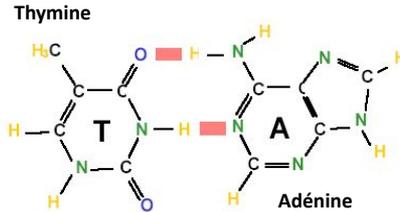
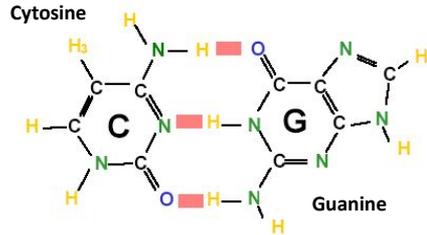
L'information génétique réside dans la succession de ces bases azotées.

Ces bases azotées sont appariées de façon spécifique dans la structure en double hélice.



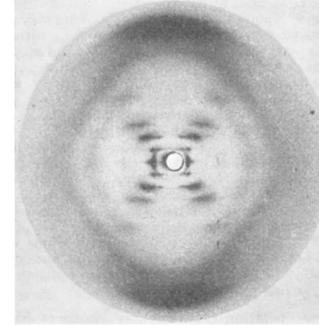
Structure de l'ADN - la double hélice

- En 1953, Watson et Crick proposent un modèle pour la structure B de l'ADN, inspiré par la structure cristallographique caractérisée par Rosalind Franklin.
- L'ADN est une double hélice, dont chacun des deux "montants" est formé d'une chaîne de désoxyribose (un sucre) unis par des groupes phosphate.
- Chaque "barreau" est formé par une paire de nucléotides liés par des ponts hydrogènes.
 - guanine \longleftrightarrow Cytosine (3 ponts hydrogène).
 - Adénine \longleftrightarrow Thymine (2 ponts hydrogène).



- Franklin, R.E. and Gosling, R.G. (1953) Molecular configuration in sodium thymonucleate. doi.org/10.1038/171740a0
- WATSON, J.D. and CRICK, F.H. (1953a) The structure of DNA. Cold Spring Harb Symp Quant Biol, 18, 123–131. doi.org/10.1101/sqb.1953.018.01.020
- Watson, J. and Crick, F. (1953b) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature, 171, 737–738. doi.org/10.1038/171737a0
- WATSON, J.D. and CRICK, F.H. (1953c) Genetical implications of the structure of deoxyribonucleic acid. Nature, 171, 964–967. doi.org/10.1038/171964b0

Image cristallographique l'ADN, par diffraction de rayons X (R.E. Franklin and R. Gosling, 1953)



Sodium deoxyribose nucleate from calf thymus. Structure B



Modèle de la structure de l'ADN (Watson and Crick, 1953b)



Implications de la structure de l'ADN

Dès 1953, Watson et Crick discutent de l'impact de leur modèle pour comprendre les mécanismes de répllication de l'information génétique.

- *Il n'a pas échappé à notre attention que l'appariement spécifique que nous avons postulé suggère immédiatement un mécanisme possible de copie pour le matériel génétique. (Watson & Crick, 1953b)*
- *Notre modèle d'acide désoxyribonucléique constitue en fait une paire de modèles, chacun étant complémentaire de l'autre. Nous imaginons qu'avant la duplication, les liaisons hydrogène sont rompues et que les deux chaînes se déroulent et se séparent. Chaque chaîne sert alors de modèle pour la formation, sur elle-même, d'une nouvelle chaîne complémentaire, de sorte qu'on obtient finalement deux paires de chaînes, alors que nous n'en avions qu'une auparavant. De plus, la séquence des paires de bases aura été dupliquée exactement. (Watson & Crick, 1953c)*

GENETICAL IMPLICATIONS OF THE STRUCTURE OF DEOXYRIBONUCLEIC ACID

By J. D. WATSON and F. H. C. CRICK

Medical Research Council Unit for the Study of the Molecular Structure of Biological Systems, Cavendish Laboratory, Cambridge

THE importance of deoxyribonucleic acid (DNA) within living cells is undisputed. It is found in all dividing cells, largely if not entirely in the nucleus, where it is an essential constituent of the chromosomes. Many lines of evidence indicate that it is the carrier of a part of (if not all) the genetic specificity of the chromosomes and thus of the gene itself.

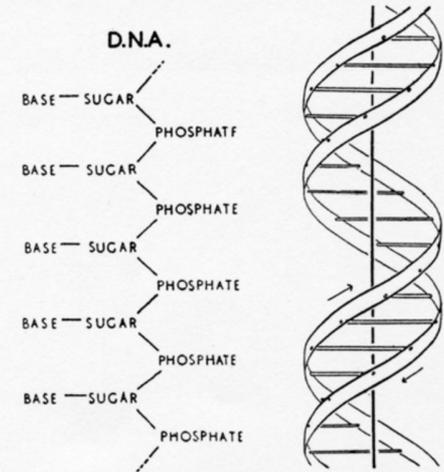


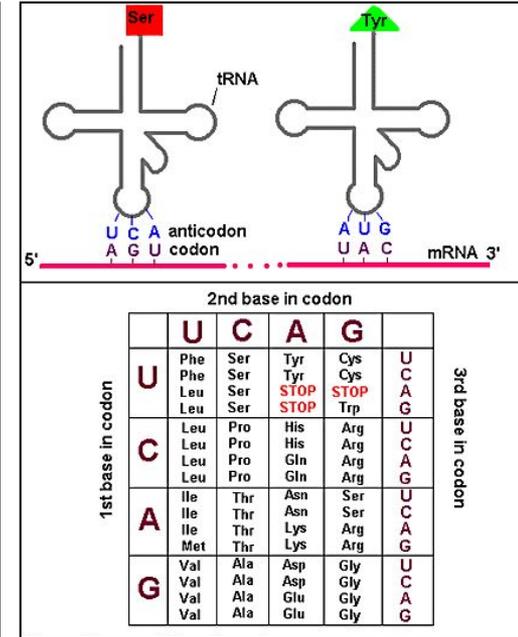
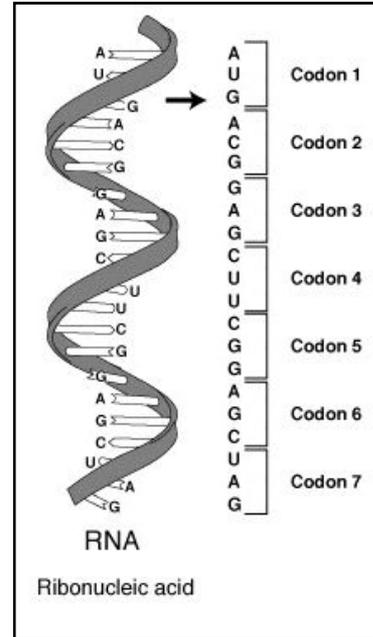
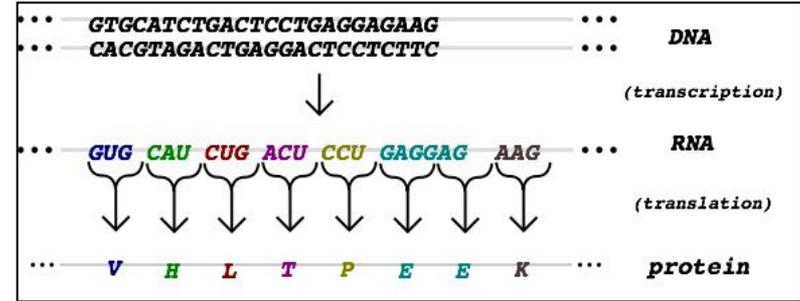
Fig. 1. Chemical formula of a single chain of deoxyribonucleic acid

Fig. 2. This figure is purely diagrammatic. The two ribbons symbolize the two phosphate-sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

Le code génétique - Concepts de base

Bref rappel de concepts-clés vus lors des cours de biologie moléculaire

- Le **code génétique** a été élucidé en 1961.
- Traduction:** les protéines sont synthétisées sur modèle de l'ARN.
- Contrairement à la transcription, il n'y a **pas de correspondance de un à un** entre nucléotides et acides aminés. En effet, l'ARN ne comporte que 4 nucléotides distincts (adénine, uracile, guanine et cytosine), tandis que les protéines sont formées de 20 acides aminés distincts.
- Codons:** chaque acide aminé est spécifié par une succession de 3 nucléotides
- Dégénérescence (redondance) du code:** Il y a 64 triplets de nucléotides possibles mais 20 acides aminés. **Plusieurs codons spécifient le même acide aminé.**



Le “dogme central”

- Le « dogme central » a été formulé en 1958 par Francis Crick. Je recommande également de lire cette discussion ultérieure (Crick, 1970).
- On le résume souvent de la façon suivante
“DNA makes RNA makes protein”
“L’ADN fait l’ARN fait la protéine”
- Cette phrase est très subtile (syntaxiquement et sémantiquement), mais souvent mal comprise. Le dogme ne se réduit pas à cette formule concise. Il énonce les transferts d’information qui sont possibles (schéma du haut) ou impossibles (schéma du bas) entre les séquences d’acides nucléiques et celles des protéines.
- Le “dogme” a souvent été critiqué par des gens qui n’avaient pas lu sa formulation exacte, en évoquant par exemple
 - La transcription réverse (“RNA makes DNA”)
 - Les modifications des prions (“protein changes protein”)
- La formulation de Crick est pourtant sans ambiguïté, et elle a conservé toute sa validité.
- Il ne s’agit pas d’un dogme mais d’une **théorie scientifique** rationnelle et logique. L’impossibilité de transfert de protéine à acide nucléique résulte directement de la dégénérescence du code.

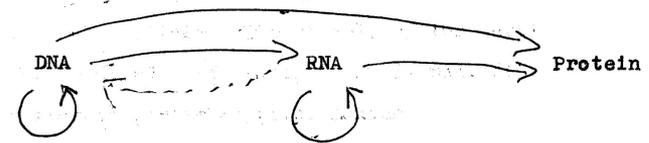
Le dogme central stipule que, une fois que l’ « information » est passée dans la protéine elle ne peut pas en ressortir. Plus précisément, le transfert d’information serait possible d’acide nucléique à acide nucléique, ou d’acide nucléique à protéine, mais le transfert de protéine à protéine, ou de protéine à acide nucléique est impossible. Information signifie ici la détermination précise de la séquence, soit des bases dans l’acide nucléique, soit des résidus aminoacides dans la protéine.

Crick, F. H. (1958). On protein synthesis. Symp Soc Exp Biol 12, 138-63.

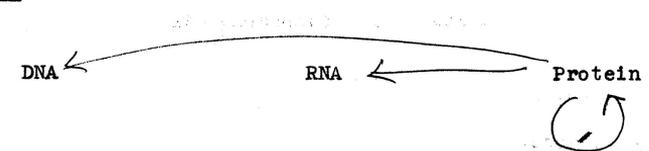
The Doctrine of the Triad.

The Central Dogma: "Once information has got into a protein it can't get out again". Information here means the sequence of the amino acid residues, or other sequences related to it.

That is, we may be able to have



but never



where the arrows show the transfer of information.

Le dogme central a-t-il été réfuté ?

On a à plusieurs reprises affirmé que le dogme central avait été réfuté :

- découverte de la transcription réverse.
- découverte du prion.

En 1970 Crick publie une clarification, pour rappeler ce que dit le dogme central, et explique pourquoi la réverse transcription ne le réfute pas.

Il distingue 3 classes de transfert d'information.

- Transferts pour lesquels on dispose d'indications directes ou indirectes (flèches pleines) : DNA \rightarrow DNA, DNA \rightarrow RNA, RNA \rightarrow Protein, RNA \rightarrow RNA.
- Possibles mais sans aucune indication d'existence (pointillés) : RNA \rightarrow DNA, DNA \rightarrow Protein.
- Très invraisemblables : Protein \rightarrow Protein, Protein \rightarrow DNA, protein \rightarrow RNA.

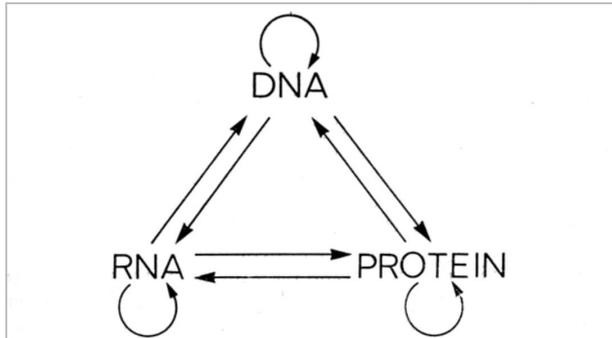


Fig. 1. The arrows show all the possible simple transfers between the three families of polymers. They represent the directional flow of detailed sequence information.

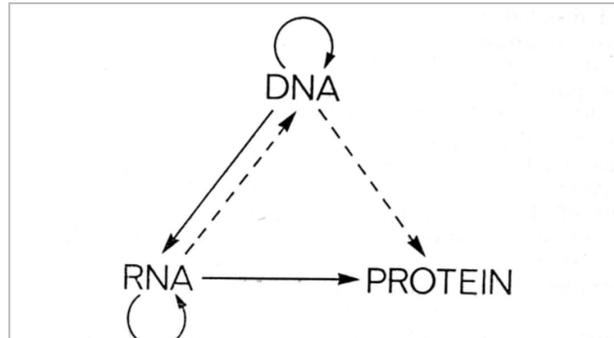


Fig. 2. The arrows show the situation as it seemed in 1958. Solid arrows represent probable transfers, dotted arrows possible transfers. The absent arrows (compare Fig. 1) represent the impossible transfers postulated by the central dogma. They are the three possible arrows starting from protein.

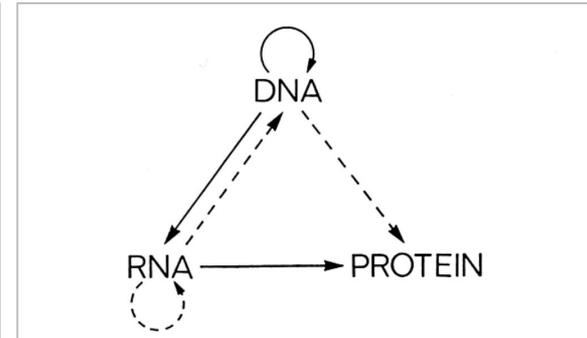
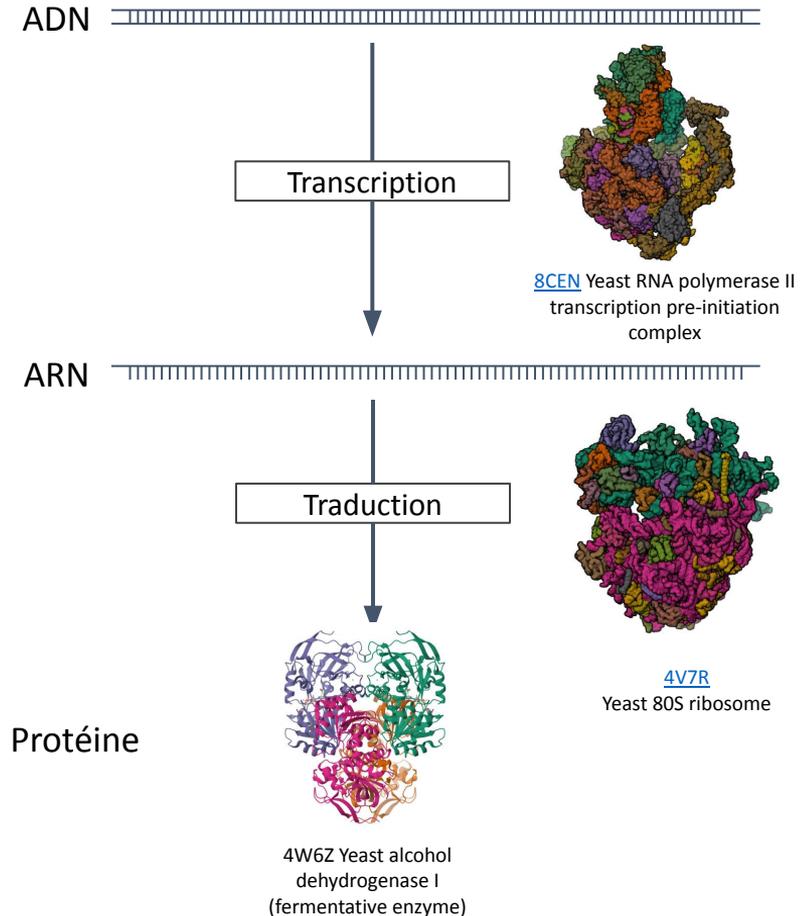


Fig. 3. A tentative classification for the present day. Solid arrows show general transfers; dotted arrows show special transfers. Again, the absent arrows are the undetected transfers specified by the central dogma.

Structure d'un gène

Le cas simple : l'ADN fait l'ARN fait la protéine

- Le modèle de base (et un peu trop simpliste) de l'expression des gènes repose sur une relation simple
 - **Transcription** : synthèse d'une molécule d'ARN sur modèle, à partir d'une région de l'ADN
 - **Traduction** : synthèse d'un polypeptide à partir de l'ARN messager (mRNA)

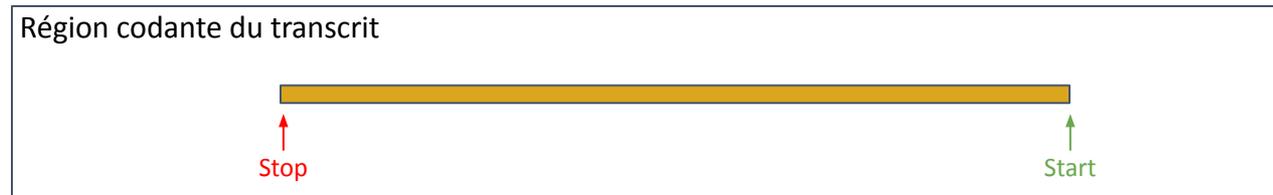
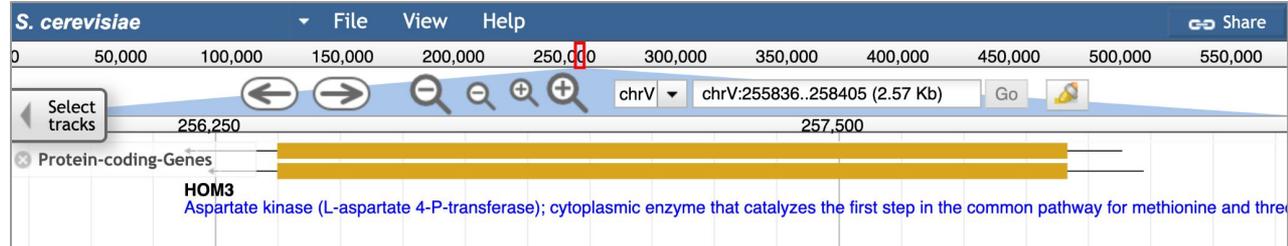


Transcripts alternatifs

Le navigateur de génomes yeastgenomes.org permet de visualiser des régions génomiques et leurs annotations (indication de tous les éléments qu'on y détecte).

Le gène **HOM3** code pour l'enzyme **aspartate kinase**, qui catalyse la première étape de la biosynthèse de l'homosérine.

- La ligne noire (partiellement marquée par la boîte ocre) indique l'étendue du transcrit.
- La flèche indique le sens de la transcription.
- Pour ce gène, il existe deux transcrits alternatifs, qui diffèrent par le site d'initiation de la transcription (Transcription Start Site, TSS) et par le site de terminaison (Transcription Termination Site, TTS)
- Le rectangle ocre indique la région codante, qui s'étend du codon start au codon stop.



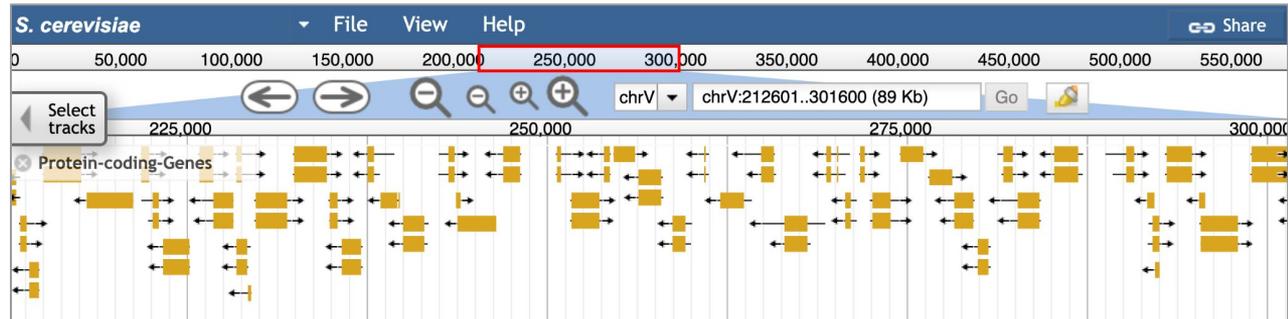
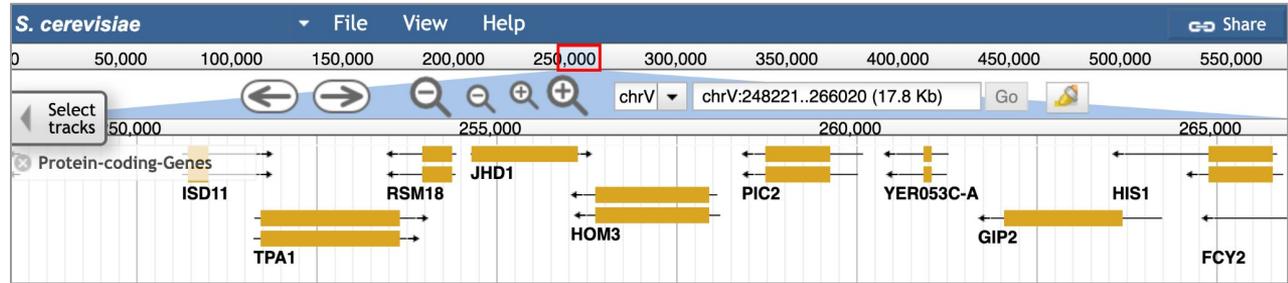
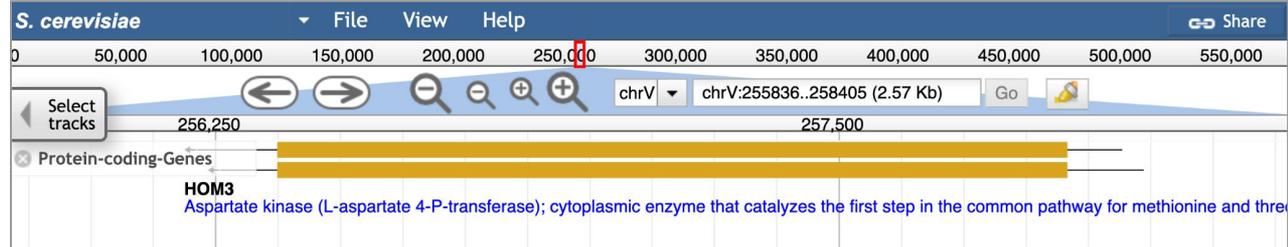
Disposition des gènes dans une région génomique de levure

En dézoomant, on peut observer la disposition des gènes dans la région génomique avoisinante.

Orientation : sur l'un ou l'autre brin, sans logique apparente.

Notation des brins

- + = D (direct) = W (Watson)
- - = R (réverse) = C (Crick)



<https://jbrowse.yeastgenome.org/?loc=chrV%3A255836..258405&tracks=Protein-Coding-Genes%2CNon-coding-RNA-Genes>

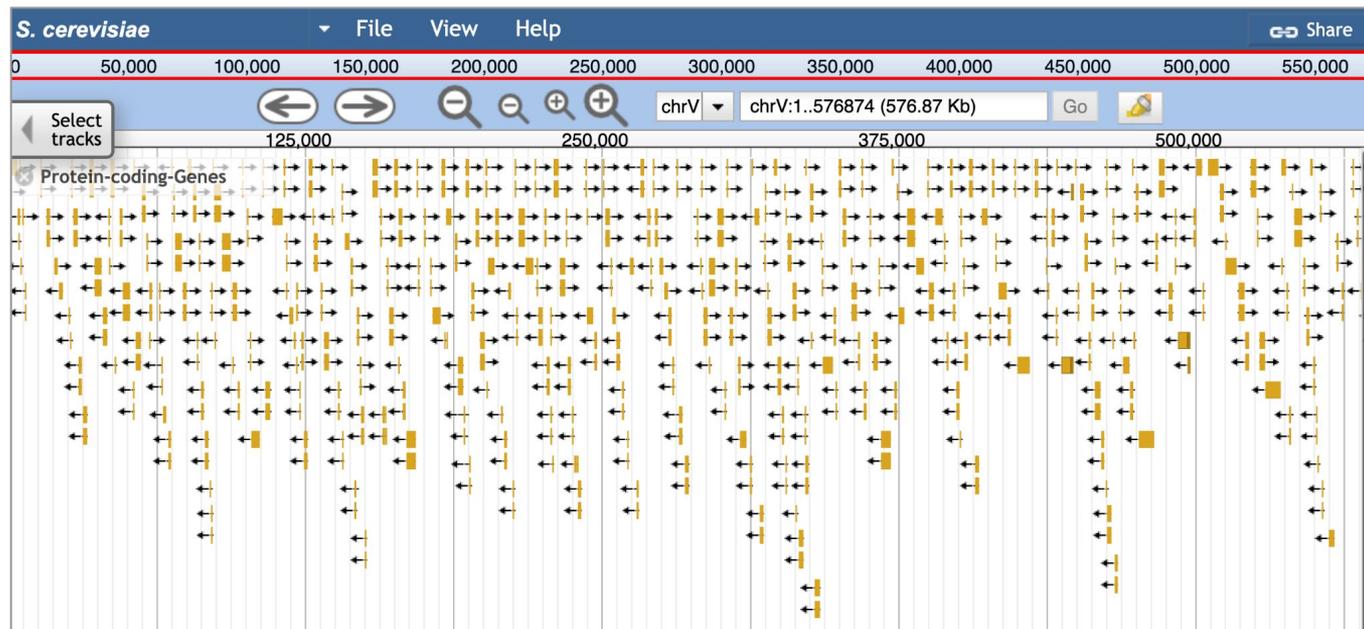
<https://jbrowse.yeastgenome.org/?loc=chrV%3A248221..266020&tracks=Protein-Coding-Genes%2CNon-coding-RNA-Genes>

<https://jbrowse.yeastgenome.org/?loc=chrV%3A212601..301600&tracks=Protein-Coding-Genes%2CNon-coding-RNA-Genes>

Disposition des gènes sur un chromosome de levure

On voit ici la disposition des gènes codants sur l'ensemble du cinquième chromosome (chrV) de levure.

- Longueur totale du chromosome : 576 874 bases.
- Nombre de gènes codants: 289
- Densité moyenne : 1 gène / 2kb



Les gènes non-codants

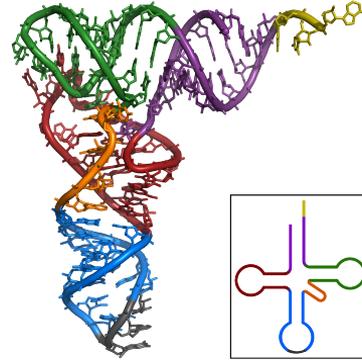
Eviter l'erreur fréquente qui consiste à ne prendre en considération que les gènes codants.

Les ARN ne font pas que servir de modèle à la synthèse des protéines.

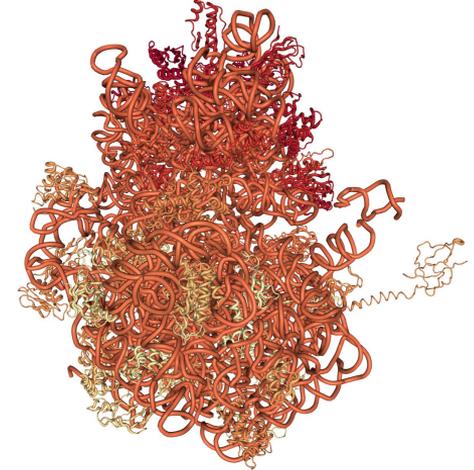
Il existe des gènes qui sont transcrits mais pas traduits.

- tRNA : ARN de transfert
- rRNA : ANR ribosomique
 - Le ribosome est un assemblage complexe d'ARN et de protéines
- lncRNA : long non-coding RNA (lncRNA)
- microRNA : petits ARN impliqués dans la régulation de l'expression des gènes

tRNA



Ribosome



PDB 4V6C. Crystal structure of the *E. coli* 70S ribosome in an intermediate state of ratcheting

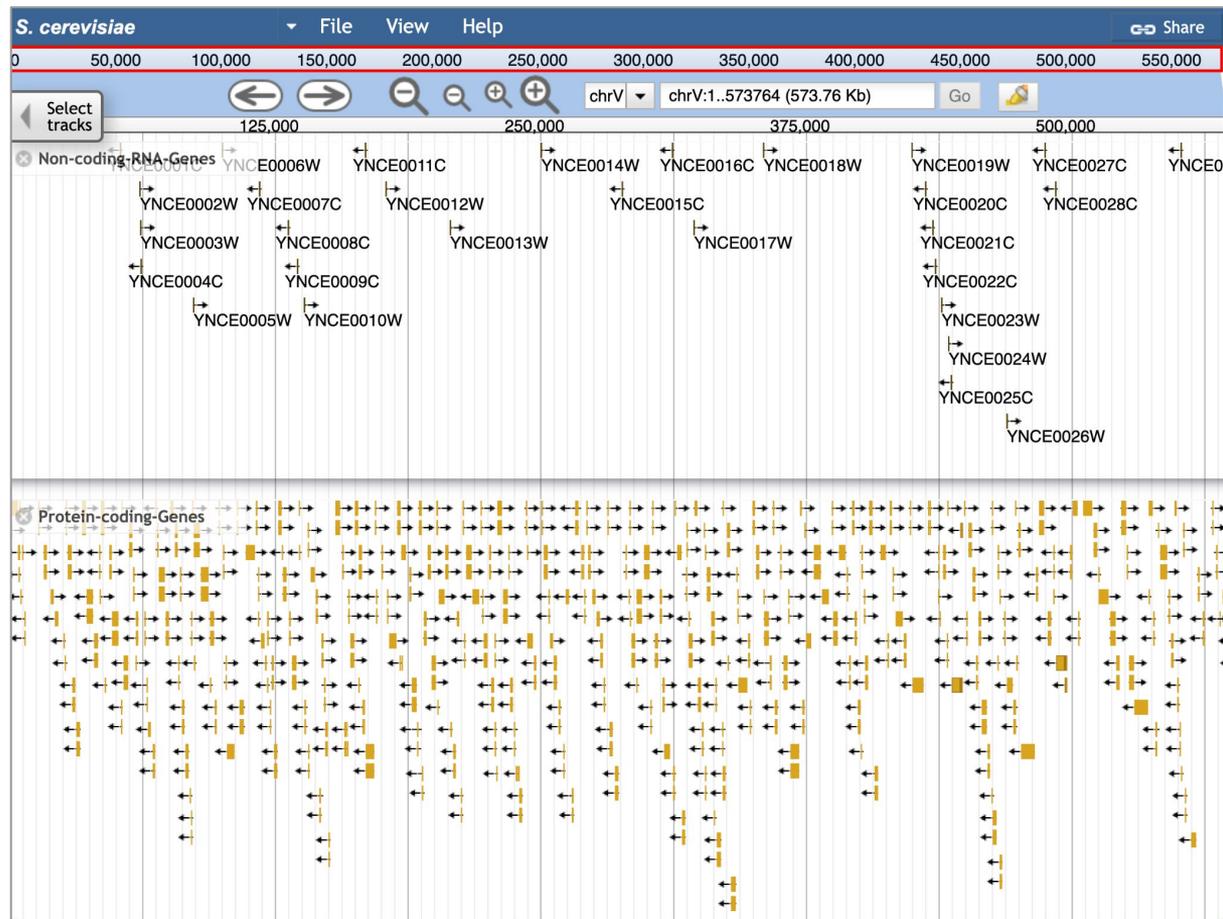
Les gènes non-codants

Le navigateur de génomes permet de sélectionner différentes **pistes d'annotation (annotation tracks)**.

Le chromosome V de la levure inclut 28 gènes non-codants (haut de la figure).

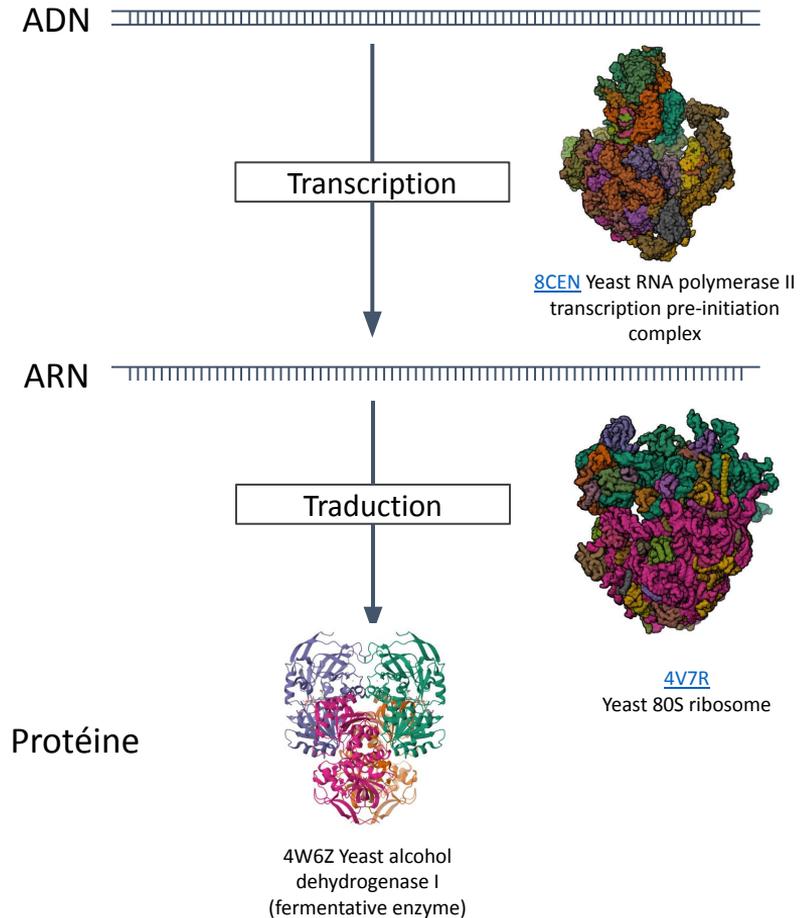
Ces gènes sont transcrits, et produisent des ARN non codants avec différentes fonctions:

- ARN de transfert (**tRNA**), 20 gènes sur le chromosome V
- snRNA: small nuclear RNA
- régulation d'autres gènes



Revenons au cas simple : l'ADN fait l'ARN fait la protéine

- Revenons au modèle de base (un peu trop simpliste)
 - **Transcription** : synthèse d'une molécule d'ARN sur modèle, à partir d'une région de l'ADN
 - **Traduction** : synthèse d'un polypeptide à partir de l'ARN messager (mRNA)



Pas si simple : l'ADN fait l'ARN primaire fait l'ARN mature fait la protéine

- D'après Uniprot, la myoglobine compte 154 acides aminés (Uniprot [MYG_HUMAN](#)).
- En principe il suffirait donc d'un ARN de 154 codons = $154 \times 3 = 459$ nucléotides pour fournir l'information nécessaire à la traduction.
- Cependant, le UCSC genome browser indique que le gène occupe ~17kb (piste [UCSC RefSeq](#))
 - Comment expliquer la différence ?
 - Comment lire et interpréter les informations du navigateur de génome ?

P02144 · MYG_HUMAN

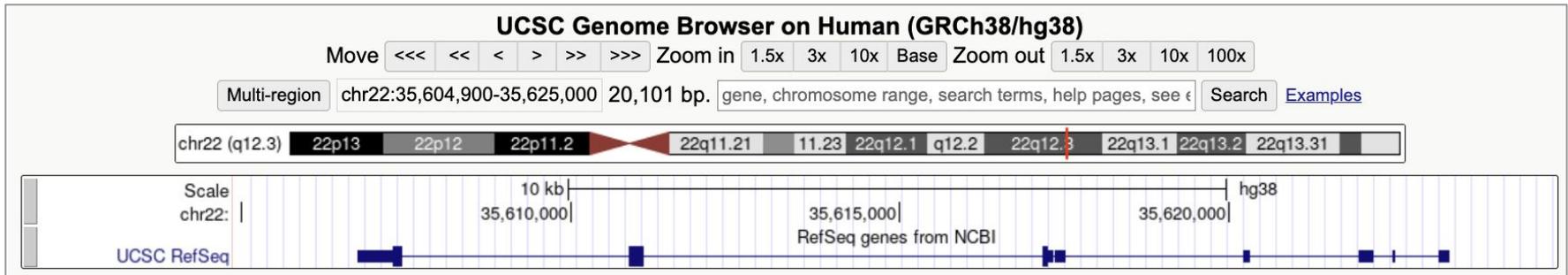
Protein ¹	Myoglobin	Amino acids	154 (go to sequence)
Gene ¹	MB	Protein existence ¹	Evidence at protein level
Status ¹	UniProtKB reviewed (Swiss-Prot)	Annotation score ²	5/5
Organism ¹	Homo sapiens (Human)		

Entry Variant viewer Feature viewer Genomic coordinates Publications External links His

Tools Download Add Add a publication Entry feedback

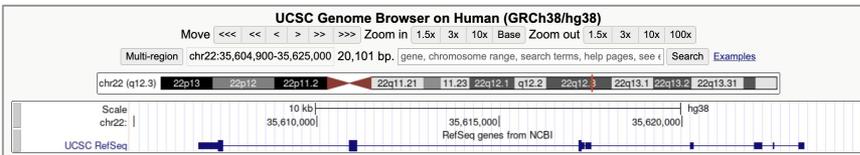
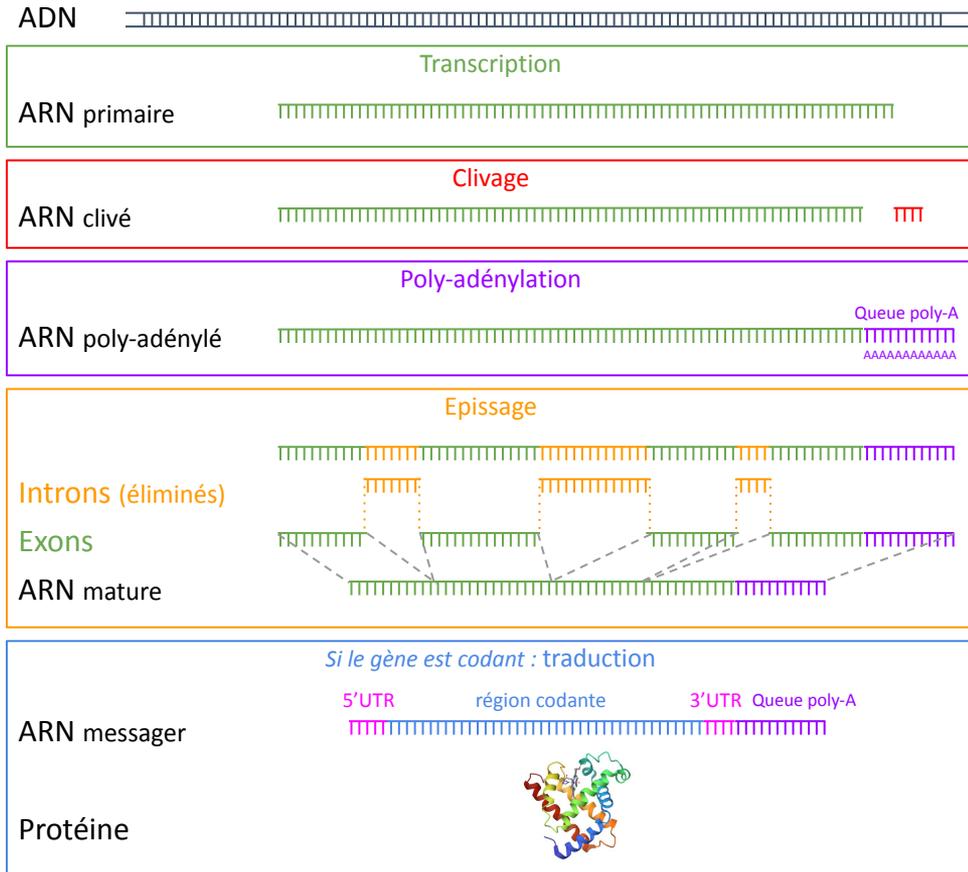
Function¹

Monomeric heme protein which primary function is to store oxygen and facilitate its diffusion within muscle tissues. Reversibly binds oxygen through a pentacoordinated heme iron and enables its timely and efficient release as needed during periods of heightened demand (PubMed:30918256, PubMed:34679218). Depending on the oxidative conditions of tissues and cells, and in addition to its ability to bind oxygen, it also has a nitrite reductase activity whereby it regulates the production of bioactive nitric oxide (PubMed:32891753). Under stress conditions, like hypoxia and anoxia, it also protects cells against reactive oxygen species thanks to its pseudoperoxidase activity (PubMed:34679218). [3 Publications](#)

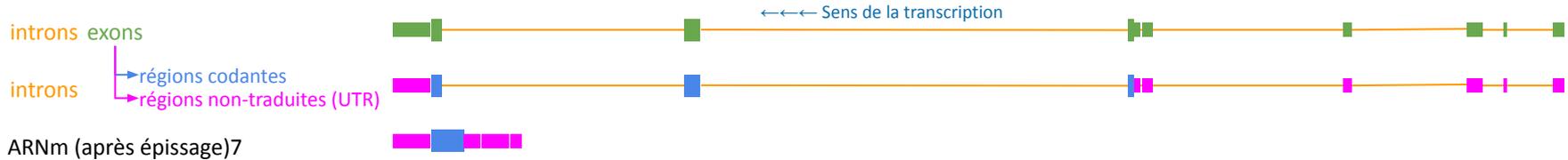
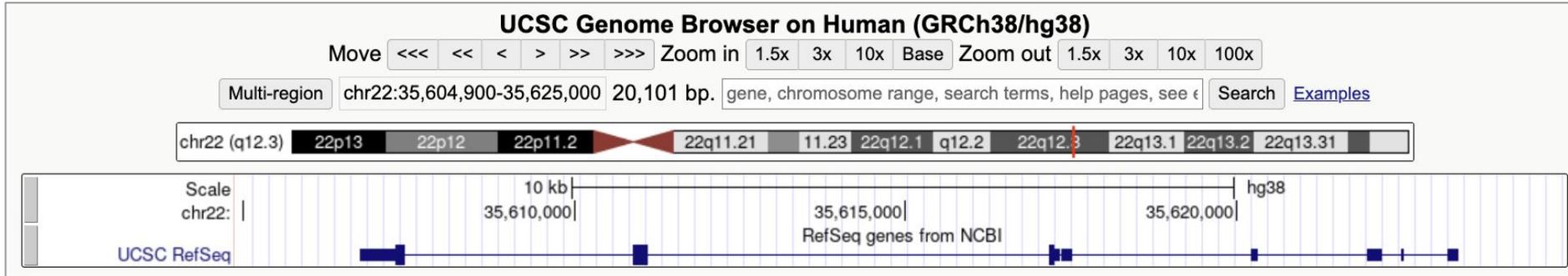


Pas si simple : l'ADN fait l'ARN primaire fait l'ARN mature fait la protéine

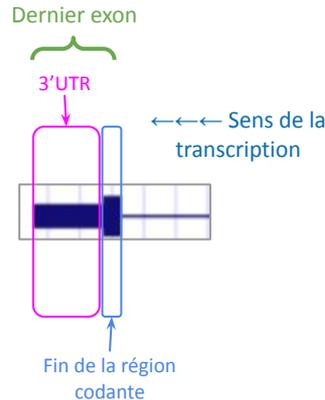
- Schéma adapté en incluant la **maturation de l'ARN**
- **Transcription** : synthèse d'une molécule d'ARN sur modèle, à partir d'une région de l'ADN.
 - Sites alternatifs d'initiation et de terminaison → transcrits multiples pour un gène
- **Clivage et poly-adénylation** : dans la région 3', l'ARN primaire est clivé (coupé), et une queue poly-A y est ajoutée (stabilisation de l'ARN). Cette queue polyA stabilise l'ARN.
- **Épissage** : élimination de certains segments de l'ARN ("introns") et raboutage des autres segments ("exons").
 - Sites alternatifs d'épissage → transcrits multiples pour un gène
- **Traduction** : synthèse d'un polypeptide à partir de la **partie codante** de l'ARN messager (mRNA).
- Note: les **régions non traduites (untranslated regions, UTR)** aux extrémités 5' et 3' de l'ARNm jouent un rôle dans la stabilité de l'ARN et dans la régulation de la traduction.



Pas si simple : l'ADN fait l'ARN primaire fait l'ARN mature fait la protéine



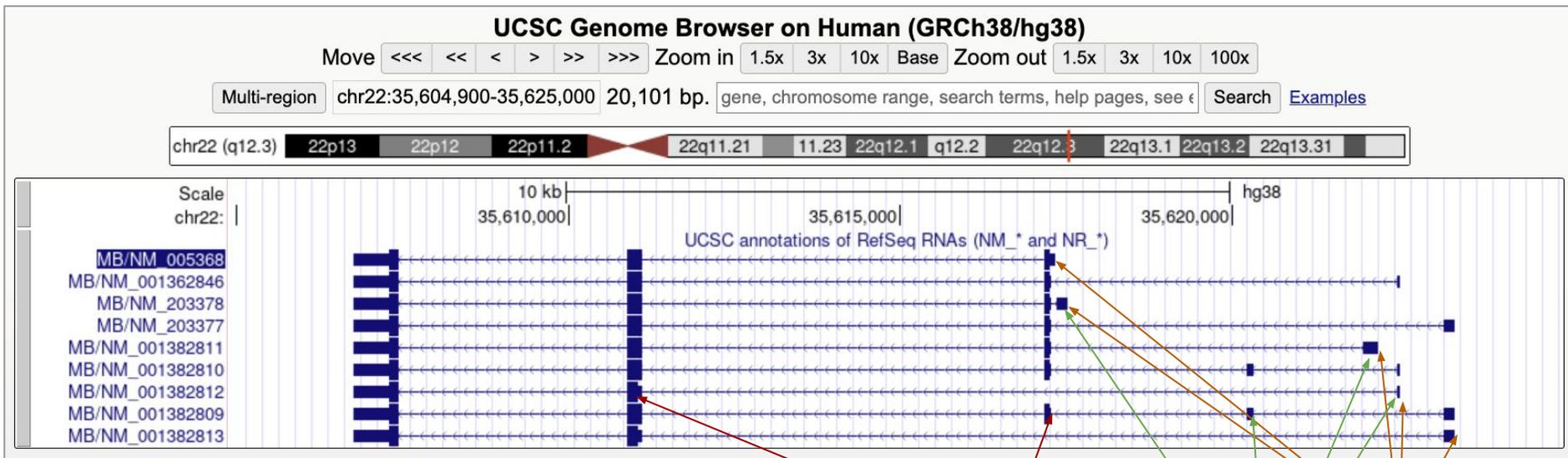
Cette figure montre la représentation “dense” du gène MB, codant pour la myoglobine (vue complète diapo suivante). Nous avons recoloré le schéma pour indiquer les différents types de régions géniques.



Remarques

- Ce gène est transcrit sur le brin réverse (de droite à gauche)
- Le dernier exon (le plus à gauche) inclut la fin de la région codante suivie du 3' UTR
- Le 5' UTR s'étend sur plusieurs exons (les 5 exons les plus à droite + la moitié du 6ème)
- L'ARN messager est beaucoup plus petit que le transcrit primaire
- La partie codante de cet ARNm couvre moins de la moitié de sa longueur

Pas si simple : transcrits alternatifs



Codons start alternatifs

Exons facultatifs
TSS alternatifs

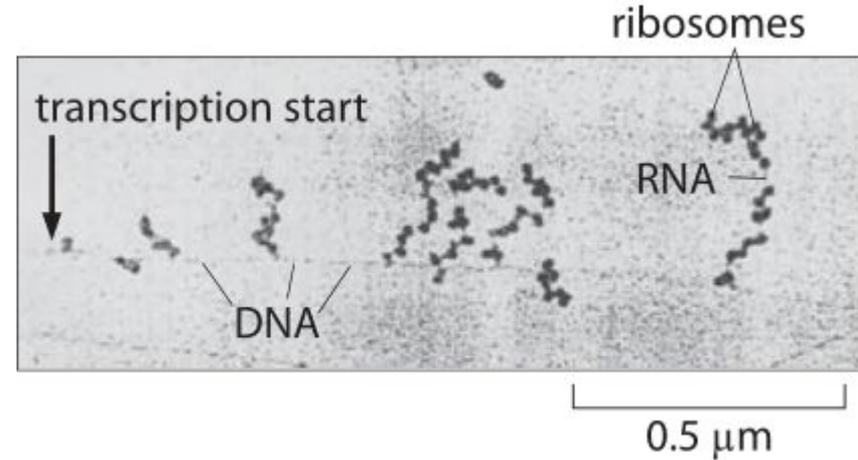
La vue complète indique que le gène de la myoglobine (MB) a au moins 9 transcrits alternatifs.

- TSS alternatifs (sites d'initiation de la transcription, transcription start sites)
- Exons facultatifs (présents dans certains échantillons, absents dans d'autres)
- Codons start alternatifs

Le transcrit du haut est majoritaire.

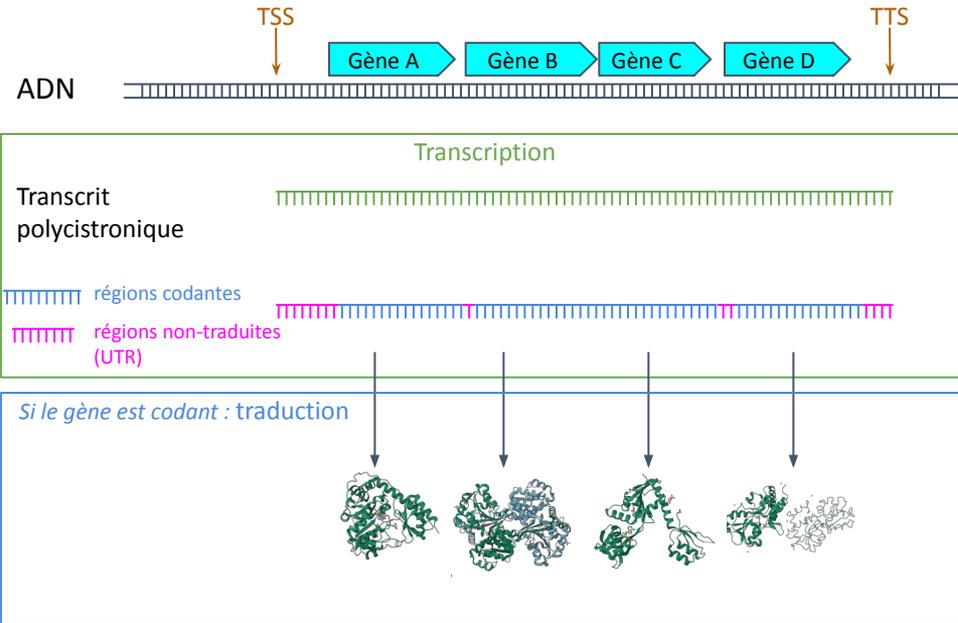
Transcription et traduction simultanées chez les bactéries

- Chez les eucaryotes, la transcription et la traduction se font séparément: dans le noyau pour la traduction, et dans le cytoplasme pour la traduction.
- Chez les prokaryotes, la transcription et la traduction se passent au même endroit, et simultanément.
- Figure: photo en microscopie électronique d'un morceau de génome bactérien (DNA) avec
 - plusieurs sites de transcription active (RNA),
 - sur chaque ARN, plusieurs sites de traduction active (ribosomes)



Génomés bactériens – Opéron

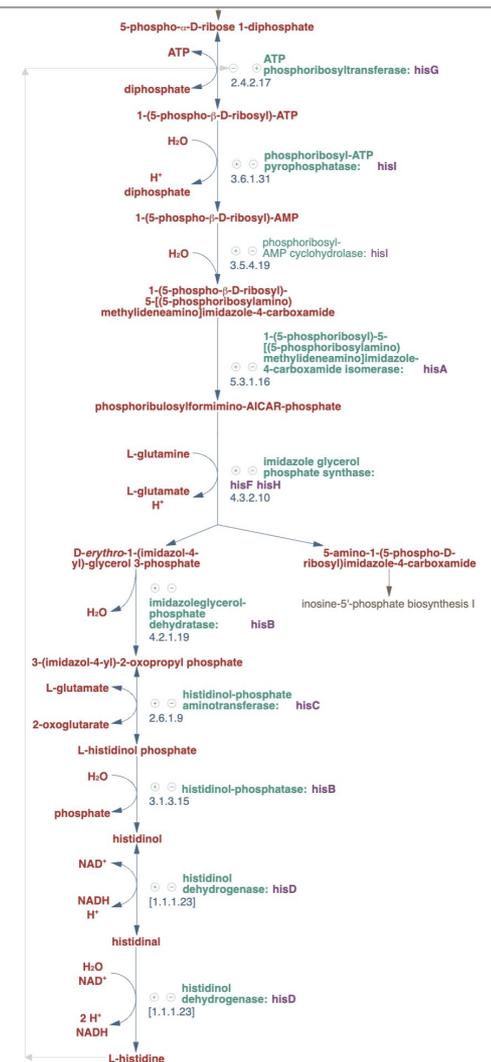
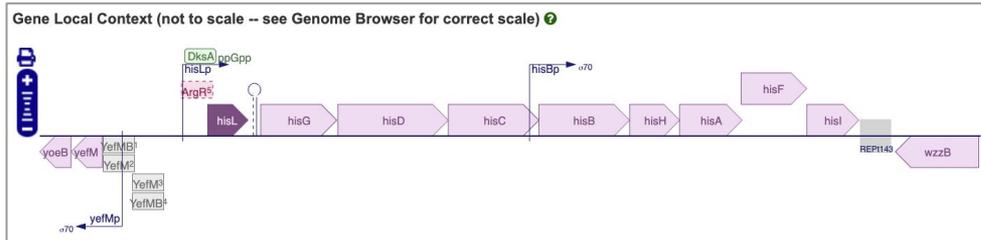
- Chez les prokaryotes, une unité de transcription peut couvrir un ou plusieurs gènes.
- **Opéron**: transcrit incluant plusieurs gènes



Exemple: l'opéron histidine d'*Escherichia coli*

Figure ci-dessous : structure de l'opéron histidine d'*Escherichia coli* extraite de la base de connaissances EcoCyc (ecocyc.org).

Figure de droite: voie métabolique de biosynthèse de la L-histidine



Exemple: l'opéron histidine d'*Escherichia coli*

Figure du haut: structure d'un opéron d'*Escherichia coli* extraite de la base de connaissances EcoCyc (ecocyc.org).

Figure du bas: localisation (mapping) des fragments de lecture d'ARN (RNA-seq transcriptomique) dans la région génomique correspondante.

- La hauteur des profils est proportionnelle au nombre de fragments de lecture localisés à chaque position.
- La couleur et l'orientation verticale indiquent le brin de lecture direct (vert, haut) ou réverse (violet, bas).
- On note un continuum de lectures sur toute la longueur de l'opéron (avec des disparités quantitatives).
- Noter aussi le gène b3207 (yrbL), transcrit séparément.

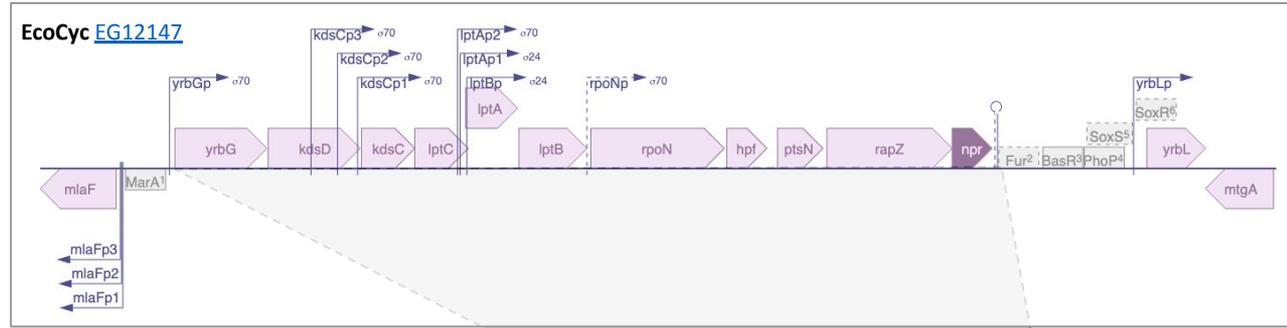
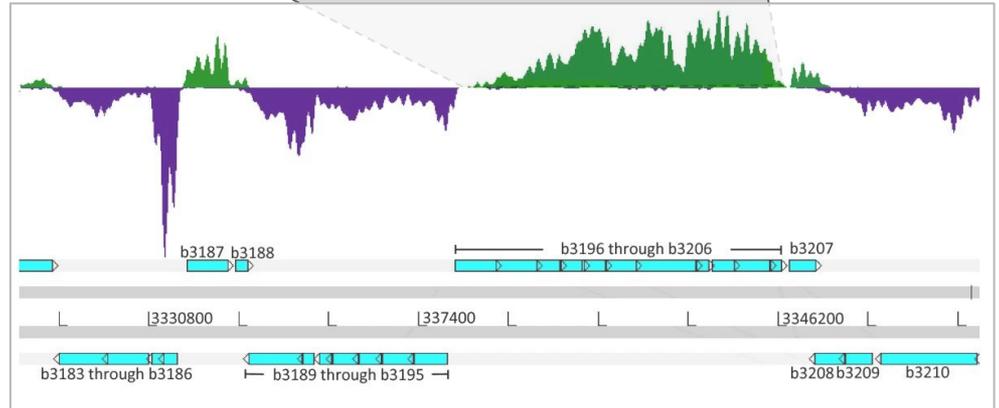


Figure de [Giannoukos et al. \(2012\)](#).

- Vert: régions transcrites sur le brin positif
- Violet: régions transcrites sur le brin négatif
- Cyan: gènes (la flèche indique l'orientation)
- "b3196 through b3206": identifiants des gènes délimitant l'opéron

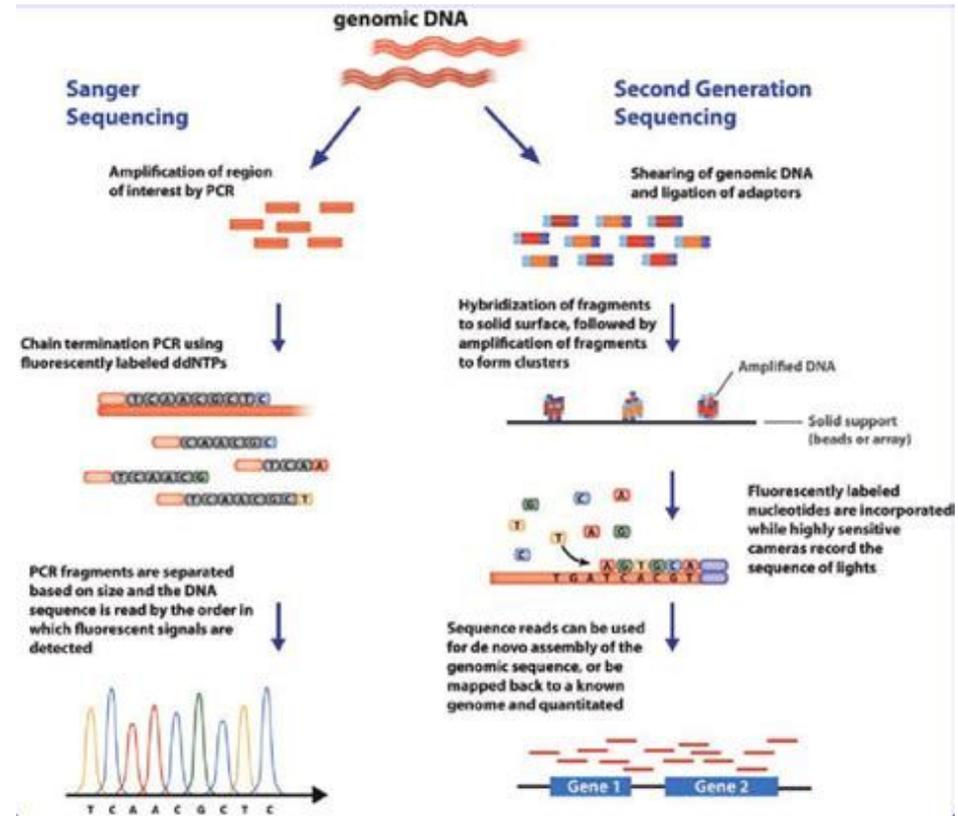


- EcoCyc : ecocyc.org/gene?orgid=ECOLI&id=EG12147#TU
- Giannoukos, G. et al. Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. *Genome Biol* 13, r23 (2012). doi.org/10.1186/gb-2012-13-3-r23

Disponibilité des génomes

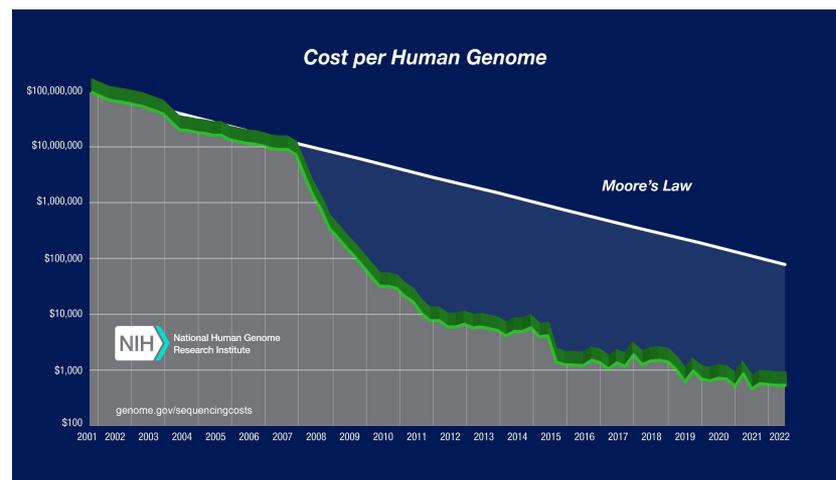
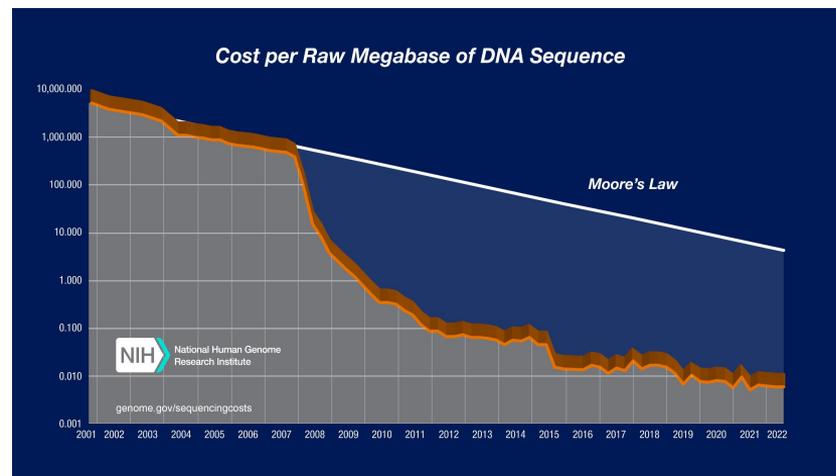
Séquençage massivement parallèle

- De 1977 à 2007, la méthode de Sanger était la seule façon de séquencer l'ADN (partie gauche de la figure)
- Durant les années 1990-2000, cette méthode a été utilisée pour les premiers projets de séquençage génomique, qui ont suscité des améliorations techniques (robotisation, informatisation)
- En 2007, plusieurs compagnies proposent une stratégie radicalement différente: le **séquençage massivement parallèle**.
- Cette approche produit des millions de petits fragments de séquences (typiquement 36 à 300bp), qu'il faut ensuite analyser, avec différentes approches possibles
 - Localisation sur un génome de référence s'il existe
 - Assemblage de novo s'il n'y a pas de génome de référence



Du gène au génome

- 1990-2000 : premiers projets de séquençage du génome d'organismes modèles: bactéries, levure du boulanger, drosophile, nématode, arabette, et ... "le" génome humain
- 2001 : première publication d'un génome humain
- 2007 : technologies de séquençage massivement parallèle ("Next Generation Sequencing", NGS)
 - De 2001 à 2007: les coûts diminuent en suivant la loi de Moore (décroissance exponentielle)
 - 2008; diminution brutale des coûts du séquençage
 - Depuis 2011: réduction plus modérée des coûts

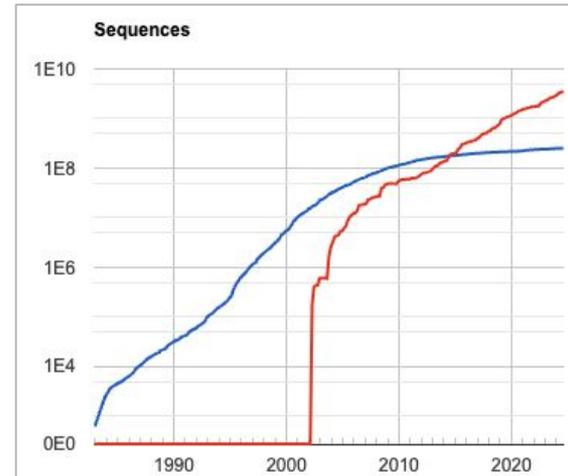
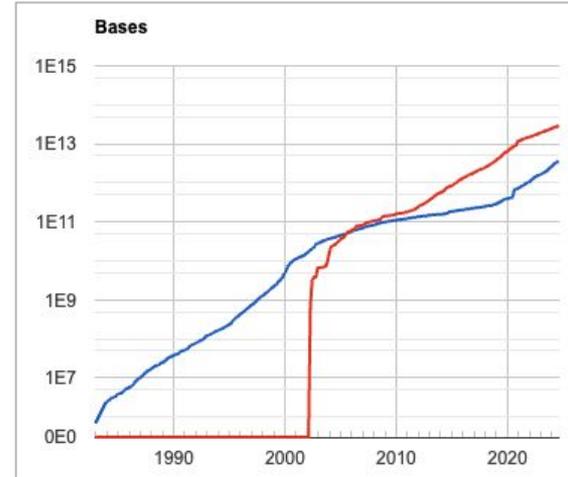


Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcostsdata. Accessed 2024-09-04.

<https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data>

Disponibilité des séquences d'ADN

- Les séquences de macromolécules qui font l'objet de publications scientifiques sont systématiquement déposées dans des entrepôts de données internationaux, et rendues accessibles au public
 - Une exception: les séquences génomiques associées à des échantillons humains (voir cours sur la médecine génomique)
- Le nombre de séquences disponibles depuis 1980 montre une croissance exponentielle (linéaire sur un axe logarithmique).
 - Taux d'augmentation: de 1990 à 2020, x 1.48/an
- Avant 2002, il s'agissait de séquences individuelles de gènes ou de fragments génomiques (courbe bleue, Genbank).
- A partir de 2002, le séquençage de génomes complets prend le pas (courbe rouge).



Disponibilité des génomes

- Avant les années 1990, le séquençage de l'ADN représentait un travail important. Un doctorant pouvait passer une partie significative de sa thèse à séquencer quelques kilobases afin de caractériser un seul gène.
- Les « projets génomes » ont stimulé le développement de méthodes de séquençage automatique, qui ont suscité des progrès technologiques impressionnants.
- Nous disposons aujourd'hui (septembre 2024) de plusieurs centaines de milliers de génomes complètement séquencés, en libre accès.

Remarques

- Le degré de finition de ces génomes varie d'un groupe à l'autre
- Un grand nombre de génomes additionnels ont été séquencés par des compagnies, et ne sont pas accessibles au public.

Génomes complets disponibles au NCBI

Date 31/07/2024

Source <https://ftp.ncbi.nlm.nih.gov/genomes/>

Nombre	Groupe
367 675	Bactéries
14 975	Virus
2 171	Archées
588	Fungi (levures, champignons)
404	Métazoaires – Invertébrés
399	Métazoaires – Autres vertébrés
220	Métazoaires – Mammifères
171	Plantes
83	Protozoaires
386 686	Total

Composition et organisation des génomes

De la génomique à la génomique fonctionnelle

Le séquençage ne constitue qu'une toute première étape pour l'analyse des génomes.

Au terme d'un projet de séquençage, on obtient un "texte" formé des 4 lettres A, C, G, T (une par nucléotide), et il reste un énorme travail de décryptage pour pouvoir interpréter ce texte.

L'exemple ci-dessous montre un fragment de 1000 nucléotides du génome humain.

```
...CGATGCTCAAACATTTCAATTTTTAGGTCAAAAATGCCTTAGGTTTAGCACAGCAATGTAGGTGCCAAACTC
ATCGCAGTGAATTGCAGGCGGGAGCAACAAGGACGCCTGCCTCCTTTCTGCCTGCTTTTTGCAATAGTCCGATTTGA
GAAGGGGACCCACGAGAGACACAAAATGCACGCCCCACGCCACATCCTTTTTACCCGCAATGGGTTAAGACTGTC
AACAGGCAGGCCACCTCGCAGCGTCCGCGGAGTTGCAGGCCCGCCCGCCAGGGTGTGGCGCTGTCCCTGGCCG
TGGCGGGGGAGGAGGGGGCGCGCGGGCCGAGGAGGGGGCGCGGGGGCGGGCGGGGGCGAGCGGAGGGCGAGTGGA
GGACCGGTAGACGCGCCCGGTCCCGCCCTGCCGCTGCTCCGCCGAGTCCGCCGCTCCAGTCTATCCGGCACTAGGA
ACAGCCCCGAGCGGGGAGACGGTCCCGCCATGTCTGCGGCCATGAGGGAGAGGTTTCGACCGGTTCCCTGCACGAGAA
GAACTGCATGACTGACCTTCTGGCCAAAGCTCGAGGCCAAAACCGGCGTGAACAGGAGCTTCATCGCTCTTGGTGGGT
GGCCGGGGTTCGCGCCCGCTGGTAGGGCCACGGGAGCCGCGCTGCCCGAGTGTGGGGAAGGAAGCAGGGAGAG
ACTCGGGAAAGGTGGAGTCCGAGACAGACGGGCAAGCAGCATATTCAGGGATCAGGCTGGCCCTCCGGAAAGCGTG
GGCATCGGAGGACCCCGGGGGCTGCCAAGCTGAGGGTCCCGGGGCTGGAGGGCAGCTCCGGCCCGGGCGCTGG
CAGCTGGAAGGGCCAGCGCTGACGTATGTCTGCCCGCGGCCCGGCCCTATTCCTGCTGTCTGCGCGGTGGGCG
GGGACGGCGGGGGCCCTGCGGGCGGGCGGTTGACGGAGGTACCCGCTCTACCCGACCTCCGTGGAGCTCCGCC
GGAG....
```

Le génome complet comporte 3 milliards de nucléotides, 3 millions de fois plus grand.

Les premières questions qui se posent au terme du séquençage =

1. Où sont localisés les gènes ?
2. Quelle est la fonction de ces gènes ?



[Drew Sheneman, New Jersey -- The Newark Star Ledger](#)

Annotation des génomes : où sont les gènes ?

A partir de la séquence « brute » d'un génome comment prédire la position des gènes ?

- Présence de **phases ouverte de lecture (longues régions sans codon stop)** indiquent des régions vraisemblablement codantes.
- **Fréquences de codons** sont caractéristiques des régions codantes.
- Fréquences des oligonucléotides.
 - Par exemple, les fréquences d'hexanucléotides diffèrent entre régions codantes et non-codantes.
- Présence de **signaux**
 - Chez les procaryotes: juste avant une région codante, on trouve parfois un motif appelé « boîte de Shine-Delgarno » (AGGAGGU), qui favorise la liaison du ribosome à l'ARN
 - Chez les eucaryotes, on peut détecter des signaux d'épissage qui indiquent les débuts et fins des exons
- **Recherche de similarité** avec des gènes connus.
 - Comparaison d'une séquence génomique avec tout ce qui a été préalablement séquencé → détection de correspondances avec séquences déjà connues.
- **Génomique comparative** : comparaison entre génomes entiers
- **Transcriptome** : localisation ("mapping") de toutes les régions génomiques transcrites dans différents tissus

Cadres ouverts de lecture (open reading frame)

Une séquence nucléique (ADN ou ARN) peut être parcourue en avançant de triplet en triplet de nucléotide, selon trois **cadres de lecture**, ou **phases de lecture**, selon qu'on parte du premier, du deuxième ou du troisième nucléotide de la séquence.

Pour les séquences d'ADN, il y a donc 6 cadres de lecture (3 sur chaque brin).

Un **cadre de lecture ouvert** (**open reading frame**, **ORF**) est un segment de séquence nucléique qui n'est pas interrompu par un **codon stop** (TAG, TGA ou TAA) dans une phase de lecture donnée, et est donc "ouvert" à la traduction (Sieber et al. 2018).

Quand on dispose d'un génome ou d'un fragment de génome, les séquences codantes (**coding sequences**, **CDS**) peuvent être identifiées en cherchant le cadre ouvert de lecture le plus long à partir d'un **codon start** potentiel (ATG) et du prochain codon stop.

Difficultés

- Tous les codons ATG ne sont pas des codons start, il existe des méthionines internes aux protéines. On prend donc généralement en compte le **plus long cadre de lecture** (depuis le codon start le plus éloigné en amont du codon stop)
- Chez les eucaryotes, les introns n'ont pas forcément une longueur multiple de 3, une protéine peut donc combiner des **cadres ouverts de lecture situés sur différentes phases** de la séquence génomique.
- Il existe des **codons start alternatifs** (exemple, chez Escherichia coli, ATG=85%, GTG=7,6%, TTG=1.2%, ...)

```
1.  ATG  CAA TGG GGA AAT GTT ACC AGG TCC GAA CTT ATT GAG GTA AGA CAG ATT  TAA
2.  A TGC AAT GGG GAA  ATG  TTA CCA GGT CCG AAC TTA TTG AGG  TAA  GAC AGA TTT AA
3.  AT GCA  ATG  GGG AAA TGT TAC CAG GTC CGA ACT TAT  TGA  GGT AAG ACA GAT TTA A
```

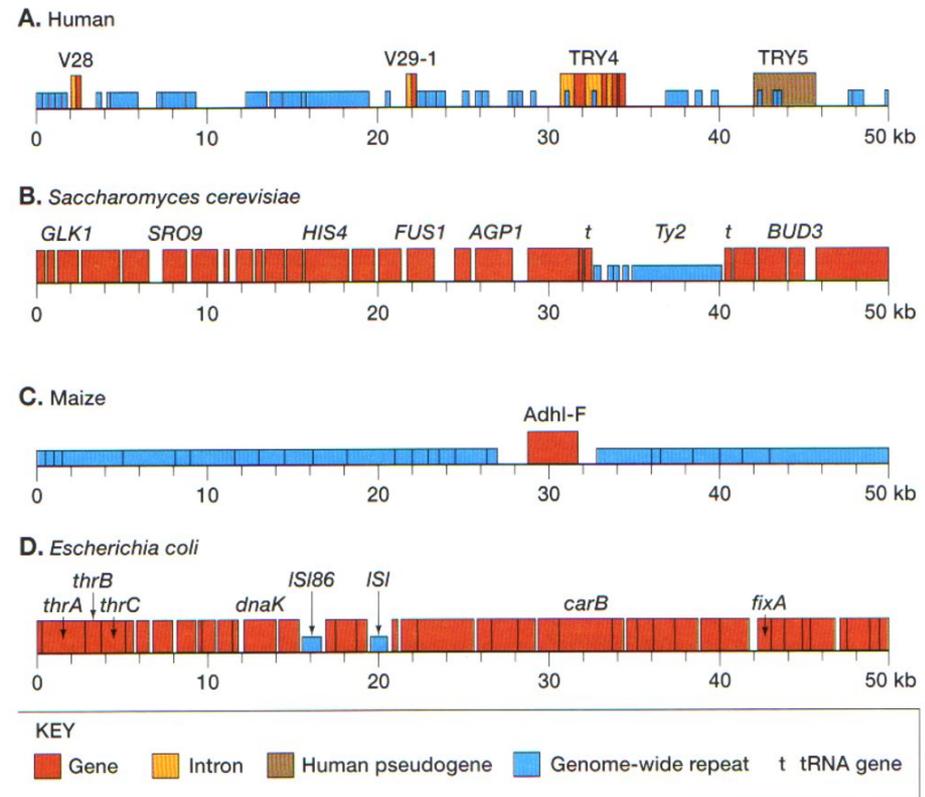
Taille et composition des génomes

Nom d'espèce	Nom commun	Année de publication	Taille du génome Mb	Nombre de gènes	Distance moyenne Kb	Fraction codante %	Fraction non-codante %	Fraction répétitive %	Fraction transcrite %
Bactérie									
<i>Mycoplasma genitalium</i>	<i>Mycoplasma</i>	1995	0,6	481	1,2	90	10		
<i>Haemophilus influenzae</i>		1995	1,8	1 717	1,0	86	14		
<i>Escherichia coli</i>	Entérobactérie	1997	4,6	4 289	1,1	87	13		
Levures									
<i>Saccharomyces cerevisiae</i>	Levure du boulanger	1996	12	6 286	1,9	72	28		
Animaux									
<i>Caenorhabditis elegans</i>	Ver nématode	1998	97	19 000	5	27	73		
<i>Drosophila melanogaster</i>	Mouche à vinaigre	2000	165	16 000	10	15	85		
<i>Ciona intestinalis</i>			174	14 180	12				
<i>Danio rerio</i>	Poisson zèbre		1 527	18 957	81				
<i>Xenopus laevis</i>	Xénope (amphibien)		1 511	18 023	84				
<i>Gallus gallus</i>	Poule		2 961	16 736	177				
<i>Ornithorynchus anatinus</i>	Ornithorynque		1 918	17 951	107				
<i>Mus musculus</i>	Souris	2002	3 421	23 493	146				
<i>Pan troglodytes</i>	Chimpanzé		2 929	20 829	141				
<i>Homo sapiens</i>	Humain	2001	3 200	21 528	149	2	98	46	28
Plantes									
<i>Arabidopsis thaliana</i>	Arabette	2001	120	27 000	4	30	70		
<i>Oryza sativa</i>	Riz		390	37 544	10				
<i>Zea mays</i>	Mais		2 500	50 000	50			50	
<i>Triticum aestivum</i>	Blé		16 000						
<i>Lilium</i>	Lys		120 000						
<i>Psilotum nudum</i>			250 000						

Structuration des génomes

La structure des génomes dépend fortement du groupe taxonomique

- Bactéries (*Escherichia coli*)
 - génomes compacts
 - majorité codante
 - Organisation en opérons
- Levures (*Saccharomyces cerevisiae*)
 - Régulation séparée pour chaque gène
 - Exons / introns occasionnels ou fréquents selon espèce
- Métazoaires – animaux pluricellulaires (ex: humain)
 - Majorité non codante
 - Éléments répétitifs
 - Structure complexe des gènes (exons / exons, éléments de régulation)
- Plantes (ex: maïs)
 - Même type de complexité que chez les métazoaires



Annotations fonctionnelles : que font les gènes ?

Assignment de fonction par similarité de séquences

Alignements globaux (Needleman-Wunsch) versus locaux (Smith-Waterman)

- Alignement global

- +Approprié, par exemple, pour les protéines homologues qui sont conservées sur toute leur longueur.
- L'alignement final inclut obligatoirement les deux séquences complètes.

```
LQGPSKTGKGS-SRSWDN
|----| - || |---|--| -
LN-ITKAGKGAIMRLGDA
```

- Algorithme: **Needleman-Wunsch** (1970).
- Outil web EMBOSS : **needle** ([nucleic acids](#) (nucleic acids or [proteins](#))).

- Alignement local

- Approprié, par exemple, pour les protéines qui partagent un domaine commun, restreint à un segment de chaque séquence.

```
LQGPSSKTGKGS-SSRIWDN
      | - || |
LN-ITKKAGKGAIMRLGDA
```

- L'alignement final est restreint aux segments conservés.
- KTGKG
- | - || |
- KAGKG
- Algorithme: **Smith-Waterman** (1981).
- Outil Web EMBOSS : **water** ([nucleic acids](#)(nucleic acids or [proteins](#)))

Alignement de séquences – Gènes S de SARS-CoV-2 et RaTG13c

```
# Aligned_sequences: 2
# 1: Human_SARS-CoV-2_BetaCoV/Wuhan/IPBCAMS-WH-01/2019
# 2: Bat_RaTG13
#
# Length: 3822
# Identity:      3549/3822 (92.9%)
# Similarity:    NA/3822 (NA%)
# Gaps:          12/3822 (0.3%)
# Score: 5435.624
```

		Identités		Substitution		Identités	
Human_SARS-CoV-2	1	ATGTTTGTTTTTCTTGTTTTATTGCCACTAGTCTCTAGTCAGTGTGTAA				50	
Bat_RaTG13	21545	ATGTTTGTTTTTCTTGTTTTATTGCCACTAGTCTCTAGTCAGTGTGTAA				21594	
...							
Human_SARS-CoV-2	2001	TGCAGG	TATATGCGCTAGTTATCAGACTCAGACTAATTC	TCC	TC	CGGCGGG	2050
Bat_RaTG13	23545	TGCAGGA	AATATGCGCCAGTTATCAGACTCAA	ACTAATTC	-----		23583
...							
Human_SARS-CoV-2	2051	CACG	TAGTGTAGCTAGTCAATCCATCAT	TG	CCT	TACACTATGTCACTTGGT	2100
Bat_RaTG13	23584	-ACG	TAGTGTGGCCAGTCAATCTATTA	T	TG	CCTTACACTATGTCACTTGGT	23632

Note

- “Indel” signifie “Insertion ou délétion”
- Sur base de ce seul résultat, on ne peut pas déterminer si la différence observée provient d’une insertion chez un ancêtre de SARS-CoV-2, ou d’une délétion chez un ancêtre de RaTG13

Alignement d'une paire de séquences protéiques

- Protéines metL et thrA d'E.coli
- Algorithme : Needleman-Wunsch.
- Barres verticales « | »
 - **Identité**: les deux résidus alignés sont identiques.
- Doubles points « : »
 - **Substitution « conservative »**
 - Les deux résidus alignés sont différents mais *similaires* (la paire de résidus a un score positif dans la matrice de substitution utilisée (ici, BLOSUM62). Voir plus loin pour comprendre ces matrices.
- Points « . »
 - **Substitution non-conservative**
 - Cette paire de résidus (distincts) a un score négatif dans la matrice de substitution.
- Espace: « »
 - **Gap**: les résidus d'une des deux séquences ne correspondent à aucun résidu sur l'autre.
 - Le gap peut provenir soit d'une délétion, soit d'une insertion, on parle donc d'*indel*, pour désigner l'événement évolutif d'où provient ce gap.

```
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
# Length: 867
# Identity:      254/867 (29.3%)
# Similarity:    423/867 (48.8%)
# Gaps:          104/867 (12.0%)
# Score: 929.0
```

```
metL      1 MSVIAQAGAKGRQLHKFVGGSSSLADVCKYLRVAGIMAEYSQPDDM-MVVSA      49
           .:. .| || | : | : | : . : . : | | | . | : . . . : . : | |
thrA      1 MRVLKFGGTSVANAERFLRVADILESNARQQVATVLSA      39

metL     50 AGSTTNQLINWLK-----LSQTDRLSAHQVQOTLRRYQCDLISG      88
           . . . . | | . | : . . : . : | : . : | . : | : : |
thrA     40 PAKITNHLVAMIEKTIISGQDALPNISDAERIFA-----ELLTG      77

metL     89 LLPAAEEADSL--ISAFV-SDLERLAALLDSGIN-----DAVYAEVVGHG      129
           | . | : . . . | . . | | . . . . . : | . | : | : . | . : . . |
thrA     78 LAAQPGFPLAQLKTFVDQEFQAQIKHVL-HGISLLGQCPDSINAALICRG      126

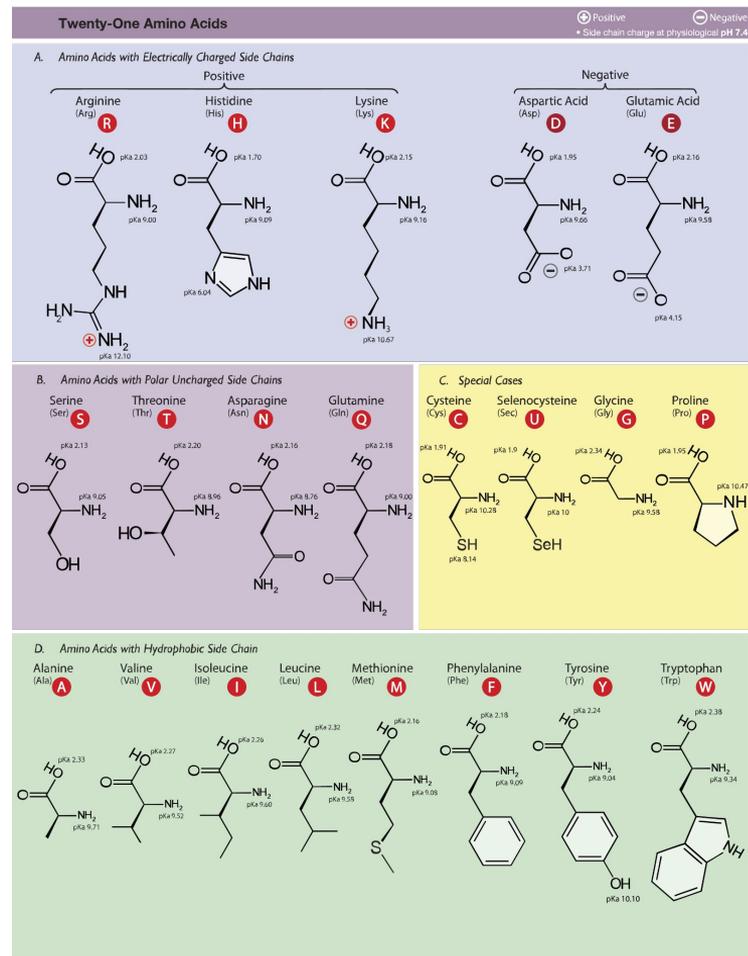
metL    130 EVWSARLMSAVLNQQGLPAAWLD-AREFLRAERAAQPQVD--EGLSYPLL      176
           | . | . : | : | | . : | . . . . : | . . . | . . . . . : | | | . . . . .
thrA    127 EKMSIAIMAGVLEARGHNVTVIDPVEKLLAVGHYLESTVDIAESTRRIAA      176

metL    177 QQLLVQHHPGKRLVVTGFISRNNAGETVLLGRNGSDYSATQIGALAGVSRV      226
           . : . . . | . : . . . | | . : . | . | | . | : | | | | | | | | . . . . | . . . . .
thrA    177 SRIPADH---MVL MAGFTAGNEKGELVVLGRNGSDYSAAVLAACL RADCC      223

metL    227 TIWSDVAGVYSADPRKVKDA CLLPLRLDEASELARLAAPVLHARTLQPV      276
           . | | : | | . | | | : . | | | : | . | | . | | . . . . | | . | | : . . | . | | : . : |
thrA    224 EIWTDVDGVYTC DPRQVPDARLLKSMSYQEAMELSYFGAKV LHPRTITPI      273
```


Rappel – Nomenclature et composition des acides aminés

Amino Acid	Abbrev	1-lettre	Codon(s)
Alanine	Ala	A	GCA, GCC, GCG, GCT
Arginine	Arg	R	CGA, CGC, CGG, CGT, AGA, AGG
Aspartic acid	Asp	D	GAC, GAT
Asparagine	Asn	N	AAC, AAT
Cysteine	Cys	C	TGC, TGT
Glutamic acid	Glu	E	GAA, GAG
Glutamine	Gln	Q	CAA, CAG
Glycine	Gly	G	GGA, GGC, GGG, GGT
Histidine	His	H	CAC, CAT
Isoleucine	Ile	I	ATA, ATC, ATT
Leucine	Leu	L	CTA, CTC, CTG, CTT, TTA, TTG
Lysine	Lys	K	AAA, AAG
Methionine	Met	M	ATG
Phenylalanine	Phe	F	TTC, TTT
Proline	Pro	P	CCA, CCC, CCG, CCT
Serine	Ser	S	TCA, TCC, TCG, TCT, AGC, AGT
Threonine	Thr	T	ACT, ACC, ACG, ACT
Tryptophan	Trp	W	TGG
Tyrosine	Tyr	Y	TAC, TAT
Valine	Val	V	GTA, GTC, GTG, GTT
STOP	-	-	TAG, TAA, TGA



Exemple de résultat de BLAST

Requête peptidique vs DB de peptides

- La ligne entre les séquences “Query” et “Sbjct” indique les correspondances entre acides aminés.
- **Identités**
- **Substitutions “conservatives”**: paires de résidus distincts mais dont la substitution est généralement moins délétère que pour d’autres paires de résidus.
- **Substitutions non conservatives**
- **Positives: identités + substitutions conservatives.**
- **Gaps: espaces (symboles -)** insérés dans une séquence afin d’optimiser l’alignement des fragments avoisinants.

Note: les modalités de représentation des identités et substitutions conservatives ou non conservatives peuvent varier d’un outil logiciel à l’autre.

```
>gi|16127996|ref|NP_414543.1| bifunctional: aspartokinase I
(N-terminal); homoserine dehydrogenase I (C-terminal)
[Escherichia coli K12]
Length = 820
```

```
Score = 344 bits (882), Expect = 2e-95
```

```
Identities = 247/821 (30%) Positives = 410/821 (49%) Gaps = 44/821 (5%)
```

```
Query: 16  KFGGSSLADVKCYLRVAGI MAEYSQPDDMM-VVSAAGSTTNQLINWLKLSQTDRLSAHQV 74
KFGG+S+A+ + +LRVA I+ ++ + V+SA TN L+ ++ + + + + +
Sbjct: 5  KFGGTSVANAERFLRVADILESMARQGQVATVLSAPAKITNHLVAMIEKTISGQDALPNI 64

Query: 75  QQTLRRYQCDLISGLLPAREADSL--ISAFVSDLERLAALLDSGIN-----DAVYAEVV 126
R + +L++GL A+ L + FV + GI+ D++ A ++
Sbjct: 65  SDAERIF-AELLTGLAAACPGFP LAQLKTFVDQEFQAQIKHVLHGISLLGQCPDSINAALI 123

Query: 127 GHGEVWSARLMSAVLNQOGLPAAWLDAREFLRAER---AAQPQVDEGLSYPLLQQLLVQH 183
GE S +M+ VL +G +D E L A + + E ++ H
Sbjct: 124 CRGEKMSIAIMAGVLEARGHNVTVIDPVEKLLAVGHYLESTVDIAESTRRIAASRIPADH 183

Query: 184 PGKRLVVTGFI SRNNAGETVLLGRNGSDYSATQIGALAGVSRVTIWSDVAGVYSADPRKV 243
+++ GF + N GE V+LGRNGSDYSA + A IW+DV GVY+ DPR+V
Sbjct: 184 ---MVL MAGFTAGNEKGELVVLGRNGSDYSAAVLAAACLRADCCEIWTDVDGVYTC DPRQV 240

Query: 244 KDA CLLPLRLDEASELARLAAPVLHARTLQPVSGSEIDLQLRCSYTPDQ-----GSTRI 298
DA LL + EA EL+ A VLH RT+ P++ +I ++ + P G++R
Sbjct: 241 PDARLLKSMSYQEAMELSYFGAKVLPRTITPIAQFQIPCLIKNTGNPQAPGTLIGASRD 300

Query: 299 ERVLASGTGARIVTSHDDVCLIEFQVPASQDFKLAHKEIDQILKRAQVRPLAVGVHNDRQ 358
E L + +++ +++ + P + + + RA++ + + +
Sbjct: 301 EDELP---VKGISNLNNMAMFSVSGPGMKGMVGMARVFAAMSRRARISVVLITQSSEY 356

Query: 359 LLQFCYTSEVADSALKILDEA-----GLPGELRLRQGLALVAMVGVTRNPLHCHRF 411
```

Exemple de résultat de BLAST

Requête peptidique vs DB de peptides

Exemple de résultat de recherche par similarité de séquences.

- Requête (**query**): metaA
- Protéine identifiée dans la base de données: (**subject**): thrA.

Le premier critère d'évaluation d'un résultat de BLAST:

- La **e-valeur (expect)** indique le nombre de faux-positifs attendus au hasard, si l'on plaçait le seuil au niveau du score observé (**344 bits** dans ce cas-ci).
- **Plus la e-valeur est faible, plus le résultat est statistiquement significatif.** Dans le cas présent, il est très significatif (**Expect = 2e-95**)
- **Si la e-valeur est ≥ 1 , le résultat n'est pas significatif** (on s'attendrait à trouver un alignement « aussi bon » avec des séquences aléatoires.

```
>gi|16127996|ref|NP_414543.1| bifunctional: aspartokinase I
(N-terminal); homoserine dehydrogenase I (C-terminal)
[Escherichia coli K12]
Length = 820
```

```
Score = 344 bits (882), Expect = 2e-95
```

```
Identities = 247/821 (30%), Positives = 410/821 (49%), Gaps = 44/821 (5%)
```

```
Query: 16  KFGGSSLADVKCYLRVAGIMA EYSQPDDMM-VVSAAGSTTNQLINWLKLSQTDRLSAHQV 74
KFGG+S+A+ + +LRVA I+ ++ + V+SA TN L+ ++ + + + + +
Sbjct: 5  KFGGTSVANAERFLRVADILESNA RQGQVATVLSAPAKITNHLVAMIEKTISGQDALPNI 64

Query: 75  QQTLRRYQCDLISGLLP AEEADSL--ISAFVSDLERLAALLDSGIN-----DAVYAEVV 126
R + +L++GL A+ L + FV + GI+ D++ A ++
Sbjct: 65  SDAERIF-AELLTGLAAAQPGFPLAQLKTFVDQEFQAQIKHVLHGISLLGQCPDSINAALI 123

Query: 127 GHGEVWSARLMSAVLNQQGLPAAWLDAREFLRAER---AAQPQVDEGLSYPLLQQLLVQH 183
GE S +M+ VL +G +D E L A + + E ++ H
Sbjct: 124 CRGEKMSIAIMAGVLEARGHNVTVIDPVEKLLAVGHYLESTVDIAESTRRIAASRIPADH 183

Query: 184 PGKRLVVTGFI SRNNAGETVLLGRNGSDYSATQIGALAGVSRVTIWSDVAGVYSADPRKV 243
+++ GF + N GE V+LGRNGSDYSA + A IW+DV GVY+ DPR+V
Sbjct: 184 ---MVL MAGFTAGNEKGELVVLGRNGSDYSAAVLAACL RADCC EIWTDVDGVYTC DPRQV 240

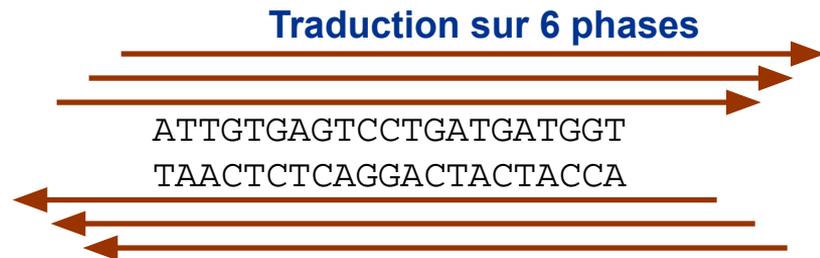
Query: 244 KDA CLLPLRLDEASELARLAAPVLHARTLQPVSGSEIDLQLRCSYTPDQ-----GSTRI 298
DA LL + EA EL+ A VLH RT+ P++ +I ++ + P G++R
Sbjct: 241 PDARLLKSMSYQEAMELSYFGAKV LHPRTITPIAQFQIPCLIKNTGNPQAPGTLIGASRD 300

Query: 299 ERVLASGTGARIVTSHDDVCLIEFQVPASQDFKLAHKEIDQILKRAQVRPLAVGVHND RQ 358
E L + +++ +++ + P + + + RA++ + + +
Sbjct: 301 EDELP----VKGISNLNNMAMFSVSGPGMKGMVGM AARVFAAMSRARISVVLITQSSEY 356

Query: 359 LLQFCYTSEVADSALKILDEA-----GLPGELRLRQGLALVAMVVGAGVTRNPLHCHRF 411
```

Traduction d'une séquence nucléique sur les 6 phases de lecture

- Si l'on dispose d'une séquence nucléique, on peut facilement déduire la séquence de la protéine qui pourrait être produite par sa traduction, sur chacun des 6 brins.
- Si cette séquence n'est pas codante, on s'attend à trouver des codons stop (noté *****) assez fréquemment (3 codons sur 64).
- Les similarités entre une séquence traduite à partir d'ADN et des protéines connues constituent des indices pour la localisation de régions codantes, et pour la fonction potentielle des nouvelles séquences.



Résultat

F1	I	V	S	P	D	D	G
F2	L	*	V	L	M	M	V
F3	C	E	S	*	*	W	X
1	ATTG	TGAGTCCT	TGATGA	TGGT	21		
	----	:-----	-----	:-----	-		
1	TAA	CACTCAGGACTACTACCA	21				
F6	X	T	L	G	S	S	P
F5	X	Q	S	D	Q	H	H
F4	N	H	T	R	I	I	T

Modalités de BLAST

Le logiciel [BLAST](#) présente 5 modalités différentes en fonction du type des séquences (peptidique ou nucléotidique) de requête et de la base de données.

Pour les comparaisons entre séquences nucléotidiques et peptidiques, la séquence nucléotidique est traduite dans les 6 phases de lecture (3 par brin), et on lance ensuite une recherche de similarité “protéine *versus* protéine”.

Translation in 6-frames



Séquence requête	Base de données	Outil	Exemples d'applications
peptidique	peptidique	blastp	En partant d'une protéine de fonction connue, collecter les protéines similaires dans la base de données Uniprot afin de constituer la famille de protéine supposées homologues.
nucléique	nucléique	blastn	Comparer les séquences d'ARNm aux séquences génomiques. Aligner un ARN d'interférence (ARNi) sur un génome pour détecter ses cibles potentielles.
nucléique (traduite dans les 6 cadres)	peptidique	blastx	Après avoir séquencé un morceau d'ADN, chercher des fragments potentiellement codants (susceptibles de produire un polypeptide similaire à des protéines connues) dans cette séquence même si on ne connaît pas la position des gènes.
peptidique	nucléique (traduite dans les 6 cadres)	tblastn	Identifier une région génomique susceptible de coder pour un homologue d'une protéine d'intérêt. Identifier dans un génome les pseudo-gènes (gènes défectifs, qui peuvent contenir un ou plusieurs codons stop) correspondant à une protéine d'intérêt.
nucléique (traduite dans les 6 cadres)	nucléique (traduite dans les 6 cadres)	tblastx	A partir d'une séquence d'ADN, identifier des segments de régions codantes ayant une contrepartie dans un génome ou une base de données de référence

Exemple de recherche de similarité : blastp (protéine vs DB de protéines)

BLAST permet de chercher, dans une base de données, toutes les séquences similaires à une séquence d'intérêt ("query", requête). L'analyse peut se faire avec des séquences nucléiques (blastn) ou peptidiques (blastp).

La figure ci-contre affiche le début de la liste des résultats, triés par significativité statistique de la similarité de séquence (les séquences les plus similaires viennent en premier).

La significativité est estimée par la E-valeur et par score de l'alignements (bit score).

Note: une E-valeur de 0 signifie en théorie qu'il n'existe aucune probabilité d'obtenir un aussi bon alignement par hasard. Cependant, en pratique cette valeur signifie que la E-valeur est inférieure à la limite de précision numérique de BLAST (cette limite est de l'ordre de $1e-180$).

NIH National Library of Medicine
National Center for Biotechnology Information

BLAST® » blastp suite » results for RID-EZVGUHB3016

Job Title: P26367:RecName: Full=Paired box protein Pax-6;...
RID: EZVGUHB3016
Program: BLAST
Database: nr_clustered(experimental)
Query ID: P26367.2

Filter Results
Percent Identity: [] to []
E value: [] to []
Query Coverage: [] to []

Cluster Composition	Cluster Ancestor	Representative Sequence	Max Score	Total Score	Query Cover	E value	Per. Ident.	Acc. Len.	Accession
B72 member(s), 63 organism(s)	jawed vertebrates	paired box protein Pax-6 isoform a [Homo sapiens]	1361	1361	100%	0.0	00.00%	422	NP_000271.1
B4 member(s), 4 organism(s)	placentalia	paired box gene 6 isoform CRA_d [Mus musculus]	1351	1351	100%	0.0	96.79%	499	EDL27248.1
B3 member(s), 3 organism(s)	apes	paired box protein Pax-6 isoform e [Homo sapiens]	1347	1347	99%	0.0	00.00%	503	NP_001355539.1
B5 member(s), 5 organism(s)	jawed vertebrates	PREDICTED: paired box protein Pax-6 isoform X1 [Mandillus...]	1347	1347	99%	0.0	00.00%	482	XP_011820972.1
B7 member(s), 6 organism(s)	amniotes	paired box protein Pax-6 isoform X2 [Phacochoerus africanus]	1347	1347	99%	0.0	00.00%	484	XP_047632014.1
B15 member(s), 210 organism(s)	bony vertebrates	paired box protein Pax-6 isoform j [Homo sapiens]	1347	1347	100%	0.0	94.41%	447	NP_001355549.1
B1 member(s), 1 organism(s)	prairie vole	Paired box protein Pax-6 [Microtus ochrogaster]	1343	1343	99%	0.0	99.76%	530	KAH0516854.1
B2 member(s), 2 organism(s)	amniotes	paired box protein Pax-6 isoform X2 [Heterocephalus glaber]	1340	1340	99%	0.0	99.52%	473	XP_021119622.1
B9 member(s), 9 organism(s)	placentalia	paired box protein Pax-6 isoform X1 [Pan caniscus]	1337	1337	99%	0.0	96.78%	504	XP_034786232.1
B44 member(s), 33 organism(s)	amniotes	paired box protein Pax-6 isoform X1 [Phylloscopus discolor]	1337	1337	99%	0.0	96.76%	496	XP_028371676.1
B1 member(s), 1 organism(s)	Brandt's bat	Paired box protein Pax-6 [Myotis brandtii]	1337	1497	99%	0.0	96.76%	564	EPQ11850.1
B18 member(s), 17 organism(s)	jawed vertebrates	paired box protein Pax-6 isoform X4 [Protopterus annectans]	1337	1337	100%	0.0	98.58%	421	XP_043815455.1
B1 member(s), 1 organism(s)	Red-fronted leucorhina	LOW QUALITY PROTEIN: paired box protein Pax-6 [Phogonulu...]	1337	1337	99%	0.0	99.28%	512	XP_064019897.1
B1 member(s), 1 organism(s)	western spadefoot toad	paired box Pax-6 isoform X2 [Pelobates cultripes]	1333	1333	100%	0.0	98.10%	542	CAH2326109.1
B5 member(s), 4 organism(s)	jawed vertebrates	paired box protein Pax-6 isoform X2 [Fukomyia damarensis]	1332	1332	100%	0.0	98.10%	422	XP_019065548.1
B7 member(s), 7 organism(s)	jawed vertebrates	paired box protein Pax-6 isoform X1 [Heterocephalus glaber]	1331	1331	99%	0.0	96.30%	487	XP_021119621.1
B1 member(s), 1 organism(s)	American alligator	paired box protein Pax-6 [Alligator mississippiensis]	1330	1330	99%	0.0	96.30%	492	XP_059577325.1
B1 member(s), 1 organism(s)	Southern elephant seal	hypothetical protein GH733_007345 [Mirounga leonina]	1330	1330	99%	0.0	82.98%	538	KAF3821971.1
B1 member(s), 1 organism(s)	red-backed fairy wren	paired box protein Pax-6 [Malurus melanocephalus]	1327	1327	99%	0.0	96.06%	616	XP_057230966.1
B1 member(s), 1 organism(s)	white-crowned sparrow	LOW QUALITY PROTEIN: paired box protein Pax-6 [Zonotrichi...]	1327	1327	99%	0.0	96.06%	570	XP_064571693.1
B3 member(s), 1 organism(s)	Bourke's parrot	paired box protein Pax-6 isoform X1 [Neoseosephus bourkii]	1327	1327	99%	0.0	96.06%	520	XP_061212598.1
B4 member(s), 3 organism(s)	amniotes	paired box protein Pax-6 isoform X2 [Equus caballus]	1318	1318	99%	0.0	98.09%	494	XP_023502322.1
B1 member(s), 1 organism(s)	African savanna elephant	paired box protein Pax-6 [Loxodonta africana]	1313	1313	99%	0.0	98.16%	654	XP_064145512.1
B1 member(s), 1 organism(s)	eastern diamondback rattl...	paired box protein Pax-6 [Crotalus adamanteus]	1311	1311	99%	0.0	95.14%	815	KAK0409866.1

Assignment de fonction par similarité de séquences - intérêt et limitations

- L'assignation de fonction par similarité est la méthode principale d'annotation des génomes nouvellement séquencés.
- Elle est notamment mise à contribution pour l'annotation automatique des protéines dans la section Unreviewed (TrEMBL) d'Uniprot, qui contient l'énorme majorité des séquences connues (245 millions de séquences protéiques au 22 septembre 2024).
- Il s'agit cependant d'une inférence très approximative. On ne peut pas définir un seuil d'identité ou de similarité qui permettrait d'inférer sans ambiguïté que deux protéines ont la même fonction.
 - Exemple: quelques changements d'acides aminés dans le site actif d'une enzyme peuvent suffire à changer sa spécificité de substrat ou de produit, même si le taux global d'identité ou de similarité reste très élevé.
- On essaie donc de combiner cette première phase des annotations par des informations complémentaires fournies par d'autres approches.

Status

 Reviewed (Swiss-Prot)
(571,864)

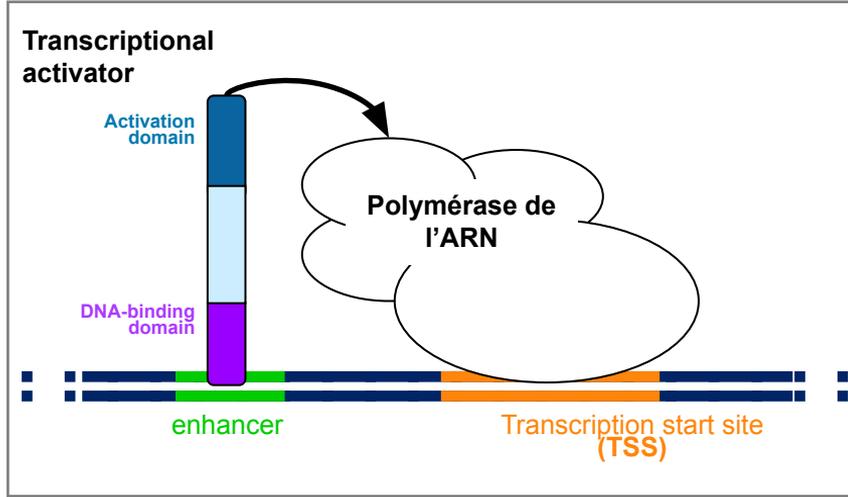
 Unreviewed (TrEMBL)
(245,324,902)

Un élément structurant des génomes: la régulation

Pour pouvoir comprendre la structure des gènes et l'organisation des génomes, il est nécessaire de connaître quelques éléments concernant la régulation génétique.

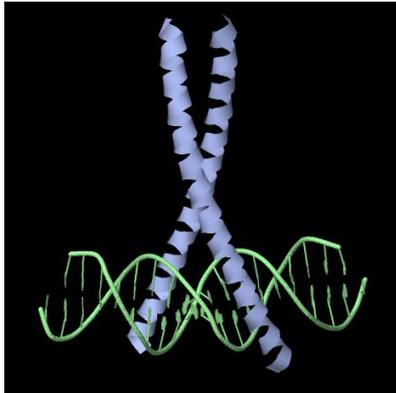
Nous résumons ci-après les notions de base indispensables, sachant que ces concepts seront développés dans vos cours de génétique et de biologie moléculaire.

Activation de la transcription



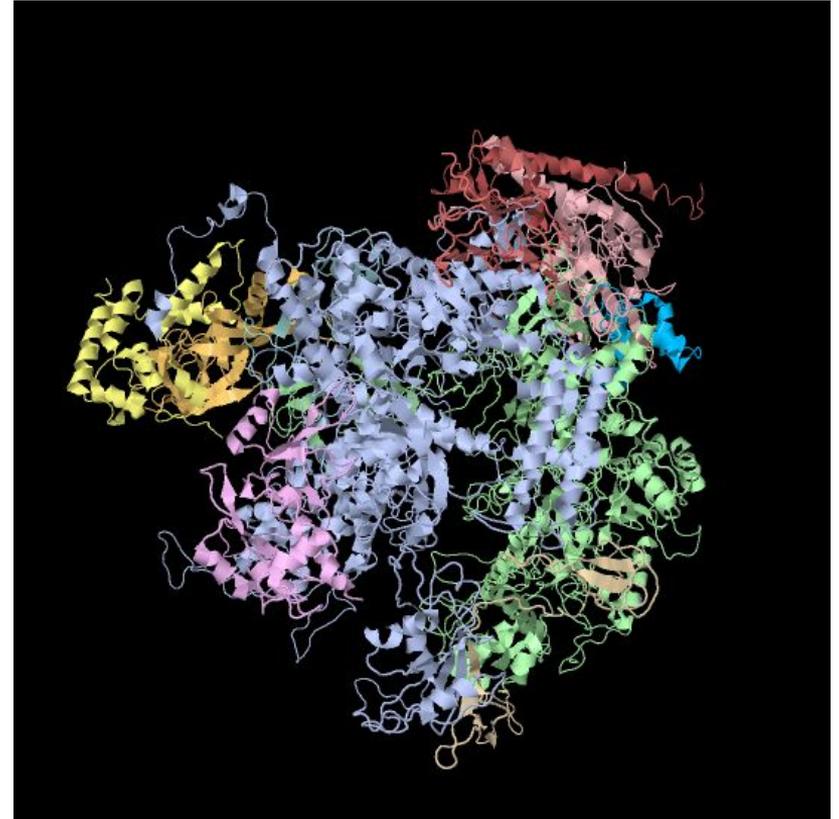
Gcn4p from *Saccharomyces cerevisiae*

PDB 2DGC <http://www.rcsb.org/pdb/explore.do?structureId=2DGC>



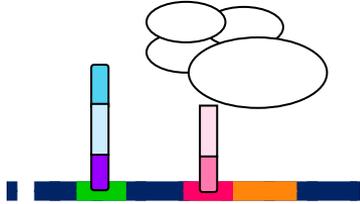
RNA polymerase II from *Schizosaccharomyces pombe*. (PDB 3H0G)

<http://www.rcsb.org/pdb/explore.do?structureId=3H0G>

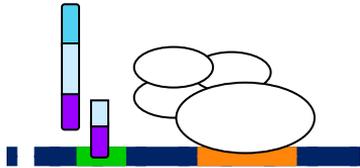


Répression transcriptionnelle

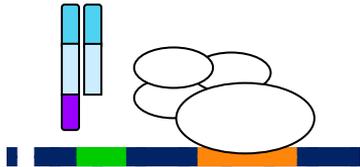
- The concept of transcriptional repression encompasses a variety of molecular mechanisms.



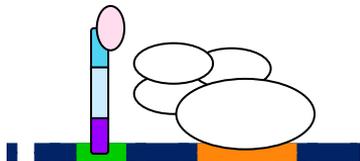
Promoter occupancy: prevent RNA polymerase from accessing DNA (e.g. many bacterial repressors)



Cis-regulatory element occupancy: competition for factor binding site (e.g. yeast GATA factors)



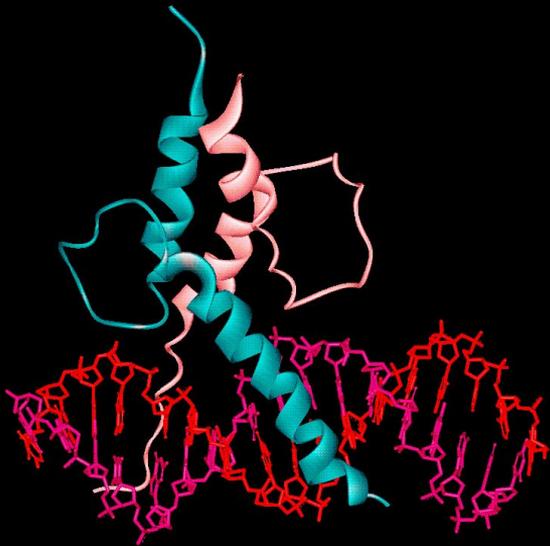
Titration of the activator: repressor forms dimer with activator, which prevents its binding to TFBS (e.g. Drosophila Helix-loop-helix)



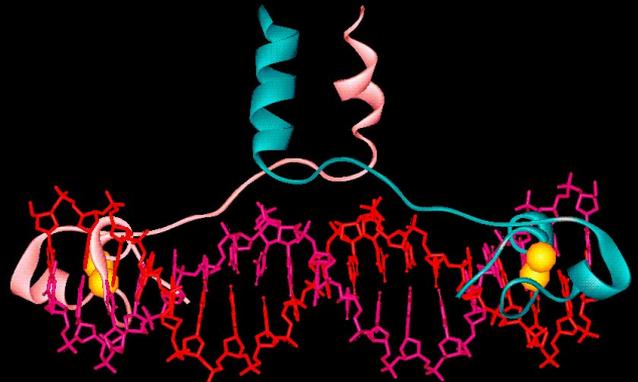
Allosteric regulation: repressor binds to activator, which alters activator conformation and prevents it from interacting with RNA-polymerase (e.g. yeast Gal80p)

Interfaces facteurs transcriptionnels / ADN

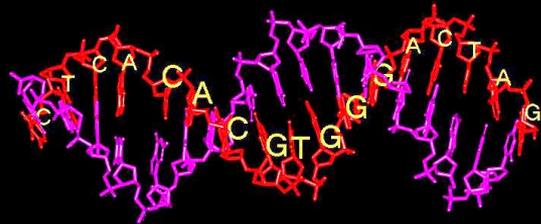
Pho4p (yeast)



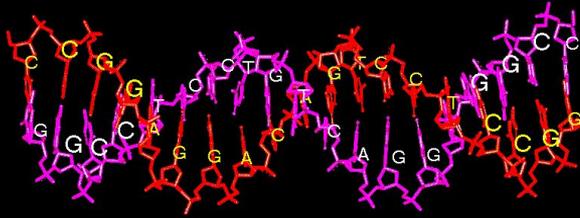
Gal4p (yeast)



Pho4p DNA binding site (oligonucleotide)



Gal4p DNA binding site (dyad)



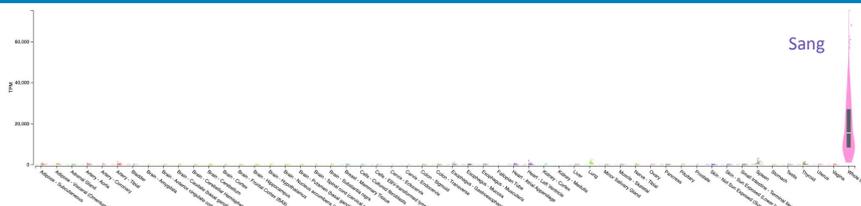
Dis-moi dans quels tissus tu t'exprimes, je te dirai qui tu es

Le projet GTEX (Adult Genotype Expression)

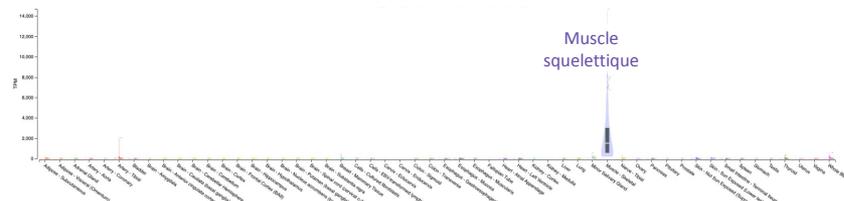
- Collecte d'échantillons de 54 tissus chez 1000 individus
- Extraction de l'ARN
- Séquençage et quantification dans chaque tissu (RNA-seq)

Exemples ci-contre: profils tissulaires d'expression pour quelques gènes illustratifs

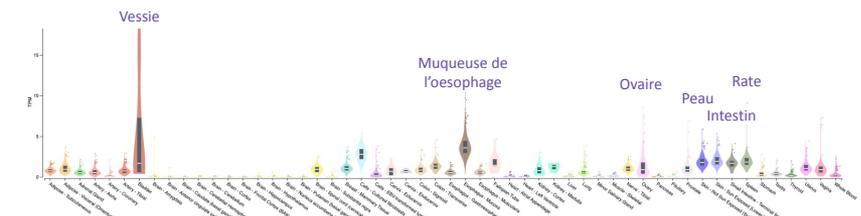
- L'hémoglobine s'exprime uniquement dans le sang
- La myoglobine s'exprime dans les muscles squelettiques
- Les gènes HoxA1 et HoxB9, impliqués dans la différenciation entre segments lors du développement, sont exprimés dans des tissus différents.



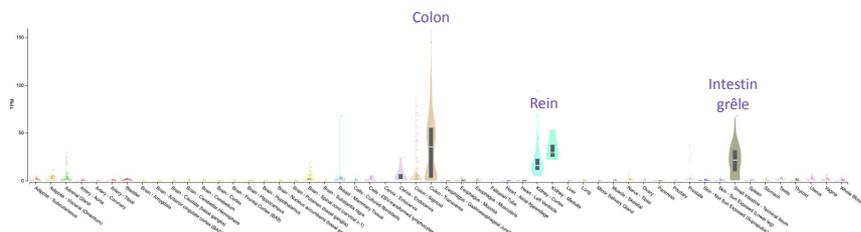
Gène HBA (chaîne alpha de l'hémoglobine)



Gène MYH1 (myoglobine)



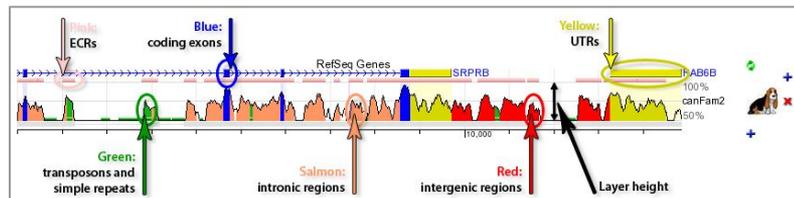
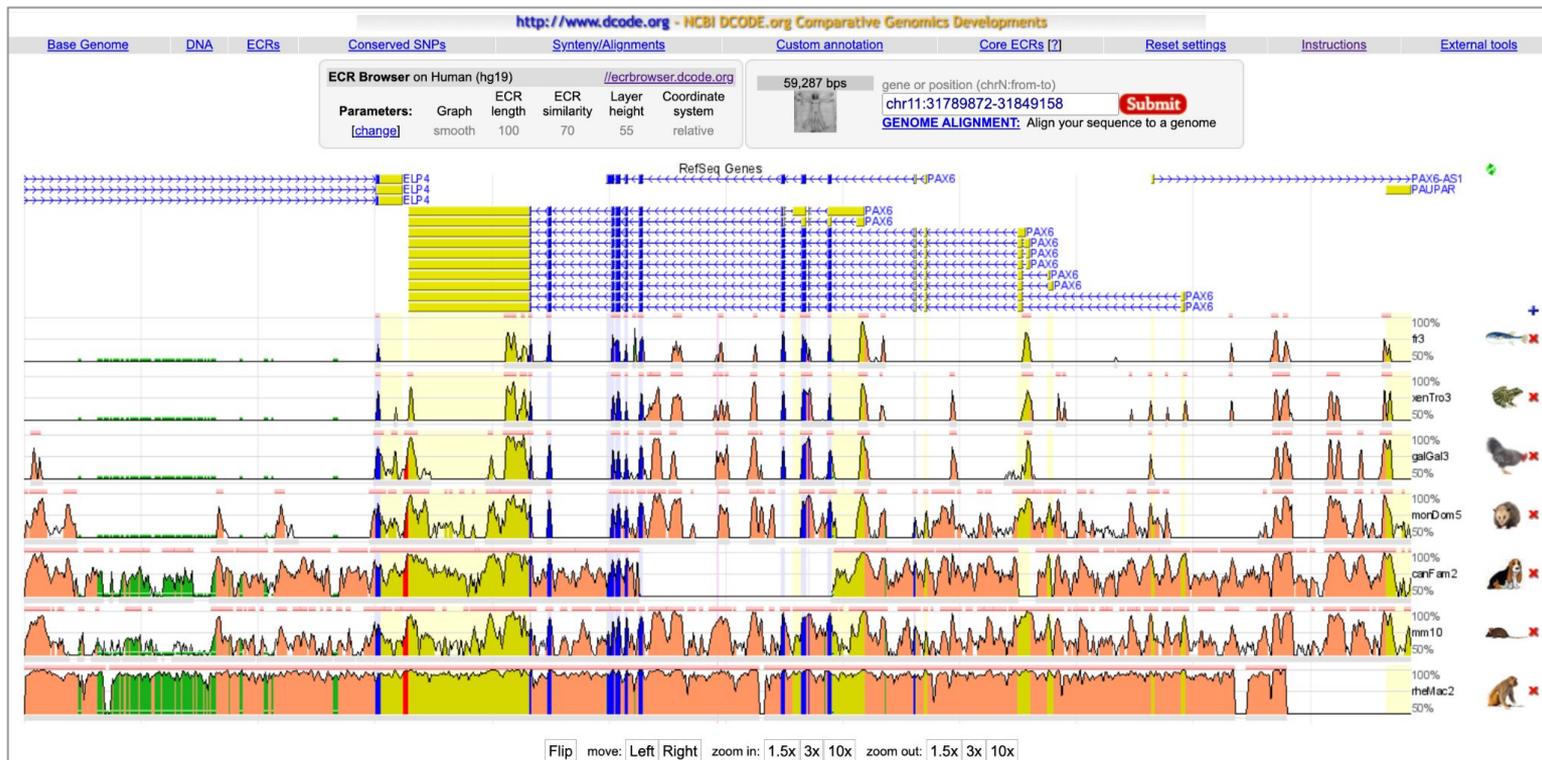
HOXA1 (gène de spécification segmentaire)



HOXB9 (gène de spécification segmentaire)

Génomique comparative

Génomique comparative : ECR genome browser (eukaryotic conserved regions)



Une vue sur le gène PAX6 – UCSC Genome Browser

Exemple “brut”: vue sur le gène PAX6 au UCSC Genome Browser. Un peu indigeste, nous le présenterons progressivement



https://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr11%3A31750005%2D31850342&hgssid=2346774682_HANILICODtNRhZ5qhQaHwllcu7OR

Coupable par association

Des génomes aux transcriptomes

Chez tous les êtres vivants l'expression des gènes fait l'objet d'un contrôle moléculaire à différents niveaux: transcription, maturation de l'ARN, traduction, post-traduction.

Une indication importante concernant la fonction des gènes est de savoir dans quelles conditions ils sont exprimés.

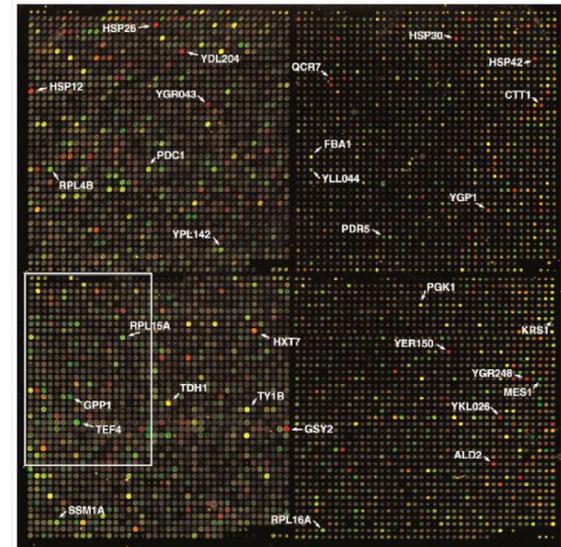
- Microbes: substrats disponibles, conditions environnementales, ...
- Multicellulaires: spécificité tissulaire, stades du développement, réponse aux conditions internes et externe de l'organisme

La transcriptomique consiste à mesurer simultanément l'expression de *tous* les gènes d'un échantillon prélevé sur un organisme dans des conditions particulières.

- 1997: premières approches de transcriptomiques par biopuces
- 2007: transcriptomique par séquençage massivement parallèle (RNA-seq)

La première biopuce transcriptomique (de Risi et al., 1997). Chacun des 6000 points lumineux correspond à un transcrite (ARN) de la levure du boulanger, *Saccharomyces cerevisiae*.

- L'intensité lumineuse est proportionnelle au niveau d'expression
- La couleur indique le sens de la régulation
 - Rouge: gènes sur-exprimés par rapport à l'échantillon témoin
 - Vert: gènes sous-exprimés
 - Jaune: gènes fortement exprimés dans les deux échantillons.



Des protéomes aux interactomes

Une protéine n'agit généralement pas seule: les protéines interagissent

- De façon stable, en formant des complexes multimériques (plusieurs polypeptides)
- De façon transitoire, en établissant des liaisons temporaires qui modifient leur niveau d'activité

Au début des années 2000, plusieurs méthodes sont mises au point pour déterminer l'**interactome**, c'est-à-dire l'ensemble des interactions entre protéines d'un système biologique (organisme, tissu, échantillon).

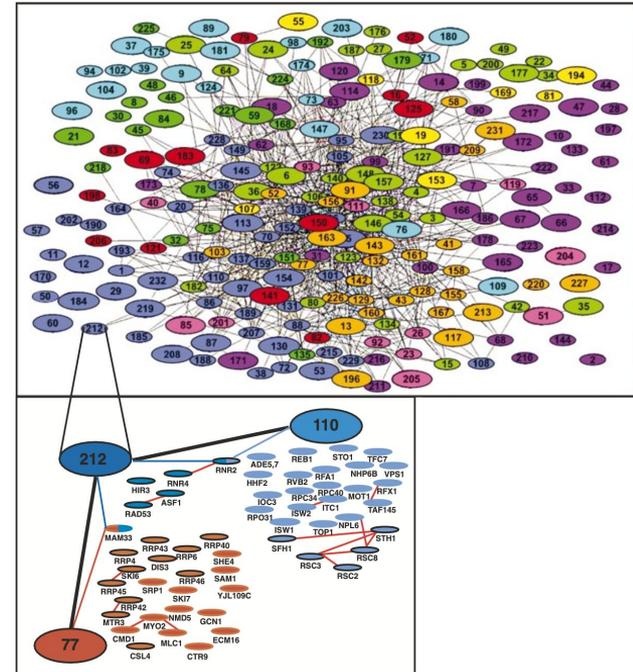


Figure 4 The protein complex network, and grouping of connected complexes. Links were established between complexes sharing at least one protein. For clarity, proteins found in more than nine complexes were omitted. The graphs were generated automatically by a relaxation algorithm that finds a local minimum in the distribution of nodes by minimizing the distance of connected nodes and maximizing distance of unconnected nodes. In the upper panel, cellular roles of the individual complexes (ascribed in Supplementary Information Table S3) are colour coded: red, cell cycle; dark green, signalling; dark blue, transcription, DNA maintenance, chromatin structure; pink, protein and RNA transport; orange, RNA metabolism; light green, protein synthesis and turnover; brown, cell polarity and structure; violet, intermediate and energy metabolism; light blue, membrane biogenesis and traffic. The lower panel is an example of a complex (yeast TAP-C112) linked to two other complexes (yeast TAP-C77 and TAP-C110) by shared components. It illustrates the connection between the protein and complex levels of organization. Red lines indicate physical interactions as listed in YPD²².

- Le principe de culpabilité par association (*guilt by association*) en annotation fonctionnelle : si l'on ignore la fonction d'un gène ou d'une protéine, mais qu'on constate qu'elle est fréquemment associée à des gènes ou protéines de fonction connue, on suppose qu'ils peuvent participer à une même fonction.
- Les critères d'association peuvent être multiples
 - Interactions physiques entre protéines détectées dans les interactomes
 - Corrélation de présence / absence d'homologues dans les génomes / protéomes de différents organismes (profils phylogénétiques)
 - Corrélation entre profils transcriptomiques
 - Procaryotes: inclusion dans le même opéron
 - ...
- La dénomination est ironique, car ce principe est bien entendu invalide en matière juridique : on ne peut pas condamner quelqu'un pour la seule raison qu'il a fréquenté des personnes qui ont commis un délit.

La Gene Ontology – Définir et structurer les termes d'annotation des gènes et de leurs produits

Gene Ontology (GO)

En 2000, Ashburner et collègues proposent à tous les projets de génomique d'adopter une "ontologie" pour annoter les fonctions des gènes (et des protéines qu'elles produisent).

Ils illustrent le concept avec trois organismes modèles.

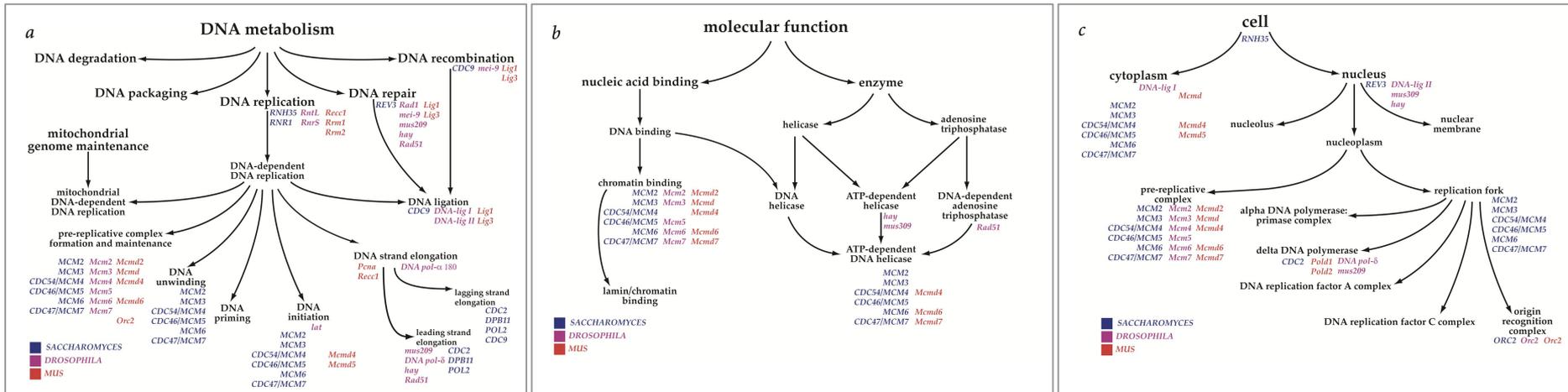
- *Saccharomyces cerevisiae* (levure du boulanger)
- *Drosophila melanogaster* (mouche à vinaigre)
- *Mus musculus* (souris)

La Gene Ontology initiale définit 3 niveaux d'annotation

- Processus biologique (figure de gauche)
- Fonction moléculaire (milieu)
- Composante cellulaire (droite)

Principes de l'ontologie

- **Vocabulaire contrôlé** : on définit une liste des **termes standards**, pour éviter les ambiguïtés liées à des formulations différentes des mêmes concepts (ex: cytoplasme = cytosol)
- **Vocabulaire structuré** : les relations hiérarchiques (flèches) sont établies entre ces termes.
- Les **relations ascendantes ou descendantes** peuvent être **multiples** (chaque noeud du graphe peut avoir plusieurs "parents" et plusieurs "enfants").
- Chaque gène est annoté en l'attachant à un ou plusieurs termes de l'ontologie (exemples colorés en bleu, rose ou rouge sur le graphe)



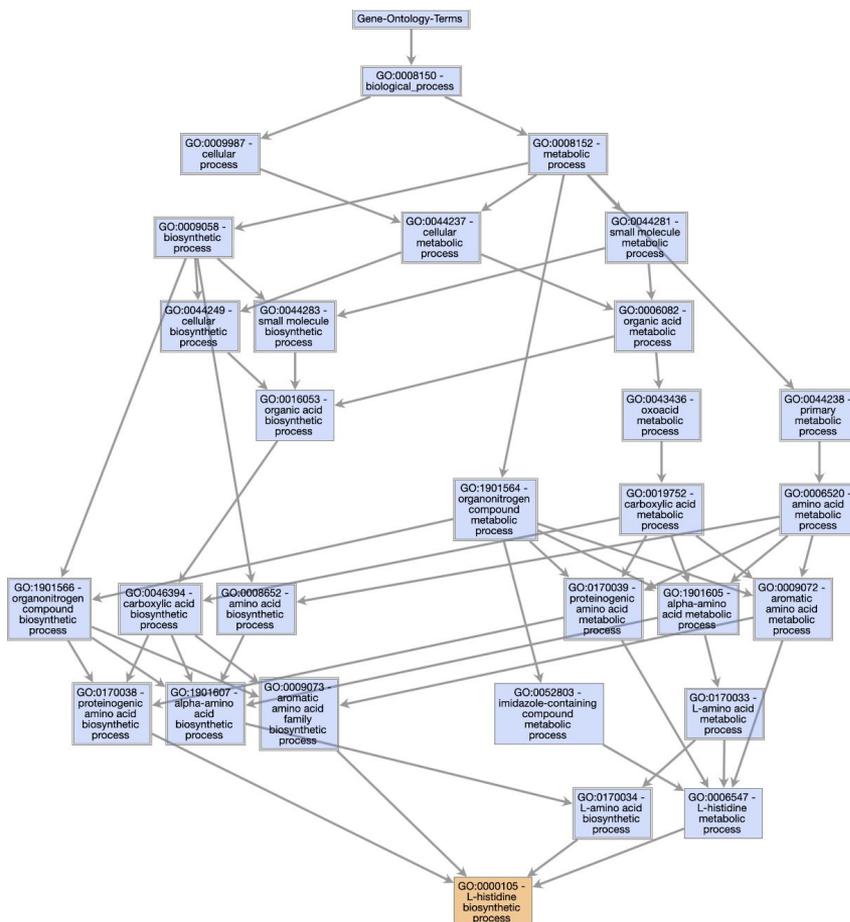
Exemple: diagramme GO du processus “biosynthèse de la L-histidine”

La voie métabolique de biosynthèse de l’histidine est rattachée à plusieurs processus parents :

- Métabolisme de la L-histidine
- Biosynthèse des acides aminés lévogyres
- Biosynthèse des acides aminés aromatiques
- Biosynthèse des acides aminés protéogéniques (impliqués dans la composition des protéines)

Ces classes ontologiques ont à leur tour des classes parentes, avec certains entrecroisements.

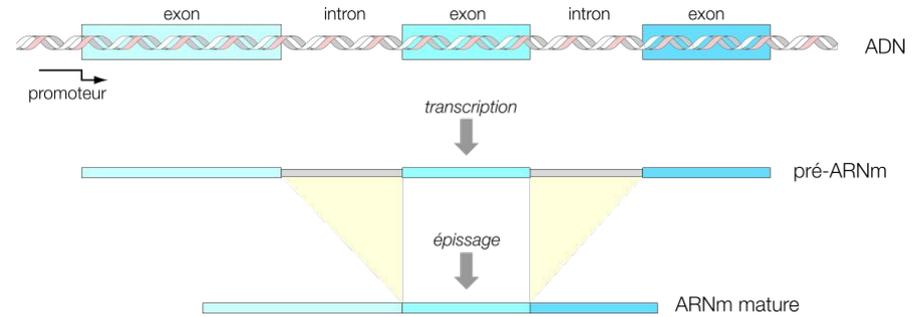
Cette structuration paraît complexe au premier abord, mais permet d’annoter chaque gène / protéine à un niveau plus ou moins détaillé de l’arborescence des termes de l’ontologie.



Information complémentaire

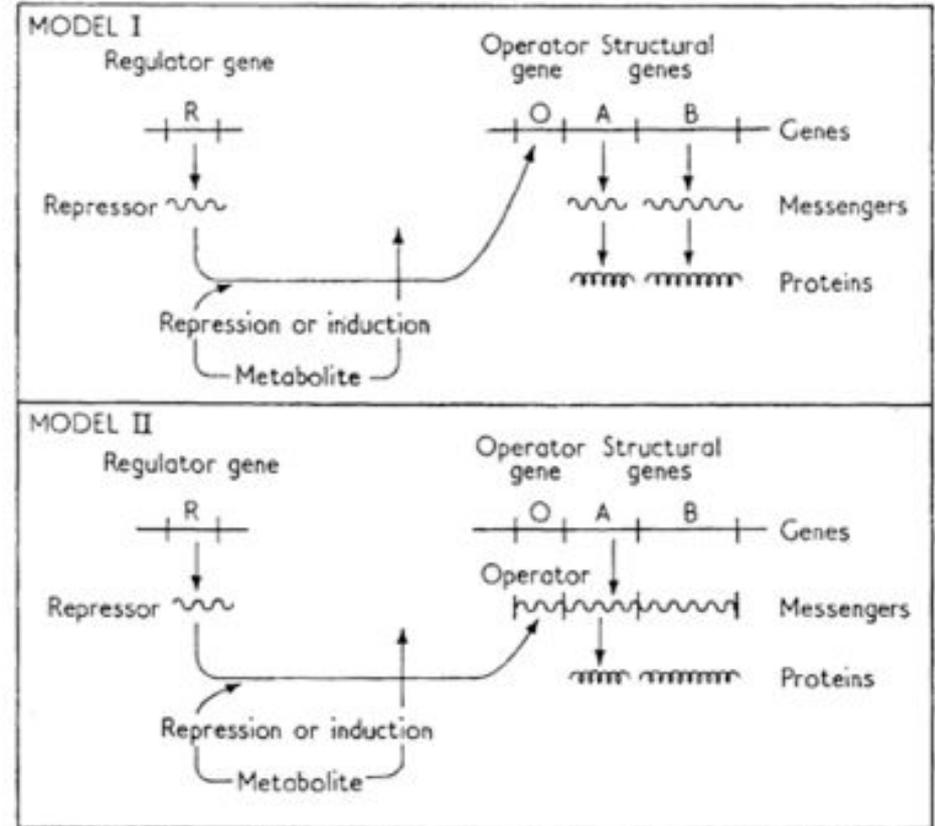
L'épissage

- Haut: ADN
- Milieu: pré-ARN = transcrit primaire
 - Principale composante de la fraction nucléaire de l'ARN (extraite du noyau cellulaire)
- Bas: résultat de l'épissage: les exons sont
 - Principale composante de la fraction cytoplasmique de l'ARN (extraite du cytoplasme)
- Exons: parties de l'ADN qui se retrouvent dans l'ARN mature
- Introns: parties de l'ADN qui sont excisées entre ARN primaire et ARN mature
- **Attention:** les exons *ne correspondent pas* aux parties codantes des gènes
 - Il existe des ARN non-codants (ex: ARN de transfert, ribosomiques, ...)
 - Le concept d'**ARN messenger** ne concerne donc que les gènes codant pour des protéines
 - Même pour les gènes codants, l'ARN messenger inclut des parties non traduites à ses extrémités 3' et 5' (*UTR: untranslated regions*)



Structuration des gènes bactériens - La découverte de l'opéron

- Depuis les années 40, Jacques Monod entreprend de comprendre les mécanismes de régulation métabolique chez la bactérie *Escherichia coli*
- 1960: François Jacob and Jacques Monod proposent deux modèles alternatifs pour la régulation de l'opéron Lac
 - au niveau de la transcription
 - au niveau de l'ARN
- Le modèle de base sous-jacent à ces deux modèles est le contrôle négatif (répression) de l'expression des gènes.
- Dans les deux cas, ils soulignent l'importance des boucles de rétroaction



Jacob, F. and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* 3, 318-56.

- Jacob, F., Perrin, D., Sanchez, C. and Monod, J. (1960). [Operon: a group of genes with the expression coordinated by an operator.]. *C R Hebd Seances Acad Sci* 250, 1727-9.
- Jacob, F. and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* 3, 318-56.
- Jacob, F. (1997). L'opéron, 25 ans après. *C. R. Acad. Sci. PAris* 320, 199-206.