

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Mohamed Khider-BISKRA.

Faculté des sciences exactes et des
sciences de la nature et de la vie



Département d'informatique

Calcul à l'ADN et Bio-informatique

Réalisé par :

Dr. TBERMACINE Ahmed

SOMMAIRE DU CONTENU

Chapitre 1 : Bases biologiques et bio-informatiques

Chapitre2 : Alignement des séquences

Chapitre3 : Alignement multiples

Chapitre 4: Arbres phylogénétiques

CHAPITRE 1: BASES BIOLOGIQUES ET BIO-INFORMATIQUES

INTRODUCTION

- Dans le cadre de leurs travaux, les biologistes ont collecté et interprété un grand nombre de données.
- Ces dernières années, ce nombre de données est augmenté exponentiellement grâce aux **nouvelles technologies expérimentales complexes** qui permettent une collecte de données beaucoup plus rapide que leur interprétation.
- Afin d'aider les biologistes dans la collecte, le stockage et l'analyse de ces données, une nouvelle discipline a vu le jour, appelée **Bioinformatique**.

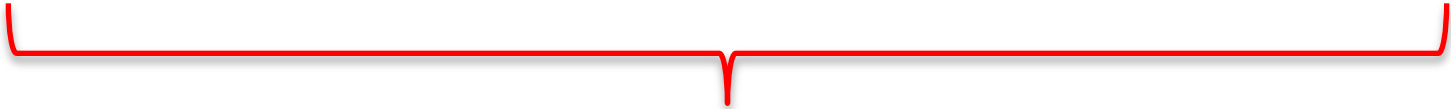
LA BIO-INFORMATIQUE. C'EST QUOI ?

- Discipline récente qui propose et développe des modèles, des méthodes et des outils afin de **gérer, manipuler** et **analyser** l'information biologique.
- La bio-informatique met en jeu plusieurs champs disciplinaires:

Informatique

Mathématiques

Biologie



Bio-informatique

LA BIO-INFORMATIQUE. POURQUOI

- Collecter et stocker des informations dans des bases de données, accessibles en ligne.
- Fournir des outils de comparaison de séquences (protéiques ou nucléotidiques).
- Fournir des outils de traduction de séquences.

QUELQUES REPÈRES HISTORIQUES...

○ Année 80:

- Début de la micro-informatique
- Création des premières bases de données (**GENBANK**)

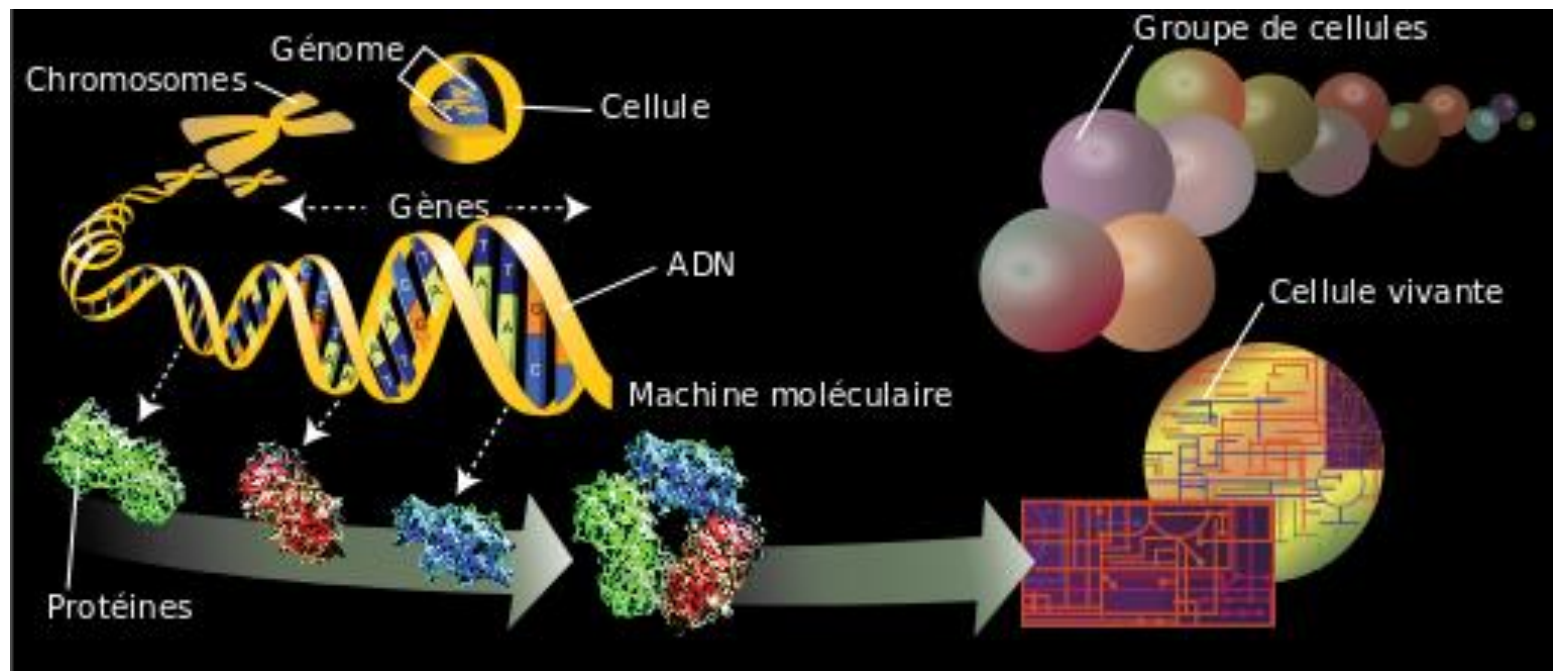
○ Année 90:

- Développement de l'internet et des réseaux
- Apparition des logiciels de comparaison de séquences (**FASTA, BLAST**).

○ Année 2000:

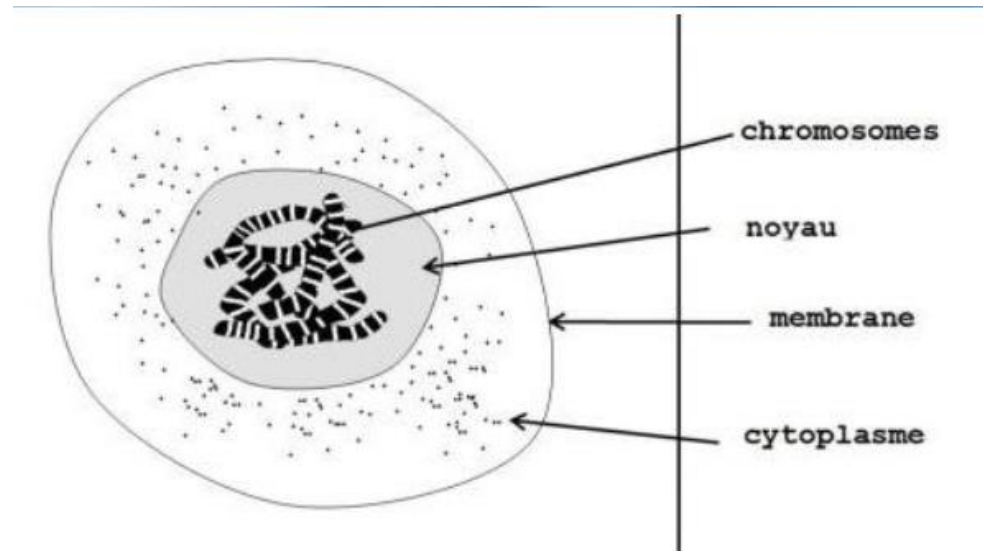
- Consultation libre en ligne des bases de données;
- Mutualisation des données avec les projets de séquençages de génomes

LES FONDEMENTS BIOLOGIQUES DE LA BIO-INFORMATIQUE



LES FONDEMENTS BIOLOGIQUES DE LA BIO-INFORMATIQUE

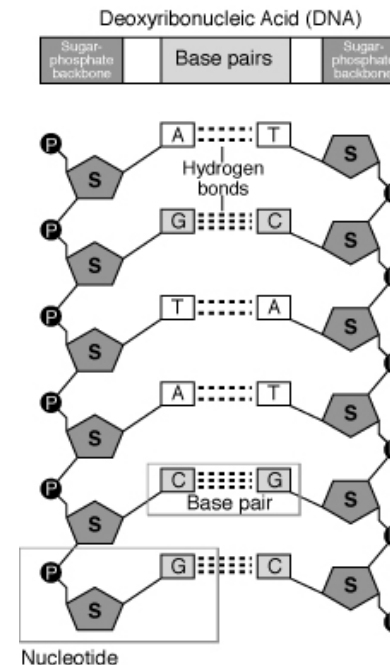
- La **biologie moléculaire** cherche à comprendre les mécanismes cellulaires au niveau moléculaire.
- La **cellule**, unité structurale et fonctionnelle fondamentale, est présente dans tous les êtres vivants connus.



LES FONDEMENTS BIOLOGIQUES DE LA BIO-INFORMATIQUE

○ Acide désoxyribonucléique (ADN)

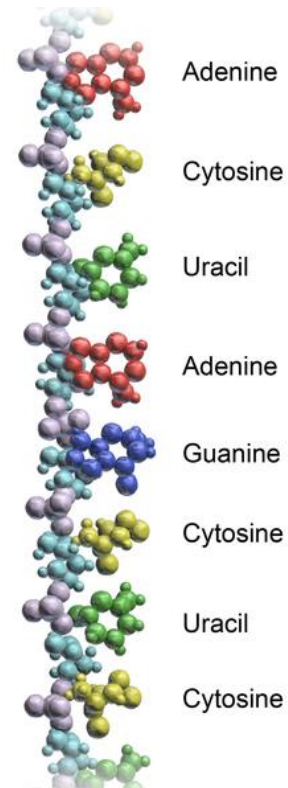
- L'ADN contient le code génétique qui renferme toute l'information héréditaire d'un organisme.
- Cette molécule est l'unité de base des gènes.



LES FONDEMENTS BIOLOGIQUES DE LA BIO-INFORMATIQUE

○ Acide Ribonucléique (ARN)

- L'ARN ressemble nommément à l'ADN mais il y a des différences telles que:
- l'ARN est généralement simple brin.
- La Thymine (T) de l'ADN est remplacé par *l'uracile (U)* dans l'ARN.



LES FONDEMENTS BIOLOGIQUES DE LA BIO-INFORMATIQUE

- **Les chromosomes et les gènes**
- Un **chromosome**, en forme de X, est constitué de deux brins d'ADN reliés au centre. Ces structures, présentes dans les cellules, renferment les gènes d'un organisme.
- Un **gène**, situé à une position précise sur un chromosome, est un segment d'ADN porteur d'informations génétiques pour la production de molécules spécifiques ou la détermination de caractères particuliers.

LES FONDEMENTS BIOLOGIQUES DE LA BIO-INFORMATIQUE

○ Le contrôle de l'expression des gènes

- Réplication:
- Mutation:
- Insertion:
- Délétion:
- Croisement :
- Transfert horizontal:

LES FONDEMENTS BIOLOGIQUES DE LA BIO-INFORMATIQUE

○ **Protéines**

- Les protéines sont les macromolécules les plus importantes.
- Elles sont responsables presque de toutes les réactions biochimiques qui ont lieu à l'intérieur de la cellule.
- Les protéines sont de sortes différentes et avec une variété de fonctionnalités.

○

LES FONDEMENTS BIOLOGIQUES DE LA BIO-INFORMATIQUE

- **Protéines**

- Les protéines se composent de chaîne des acides aminés
- Chaque acide aminé a une structure constante. Il y a **20 acides aminés**.
- **Exemple d'une protéine de l'insuline:**
5'FVNQHLCGSHLVEALYLVCGERGFFYTPKA3'

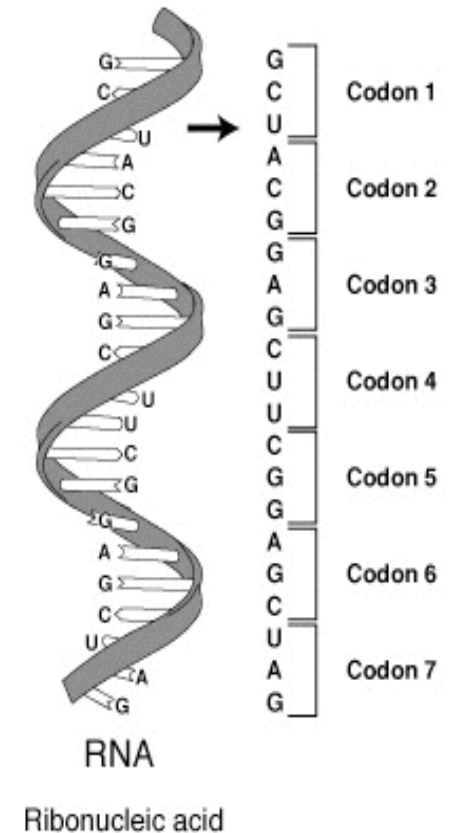
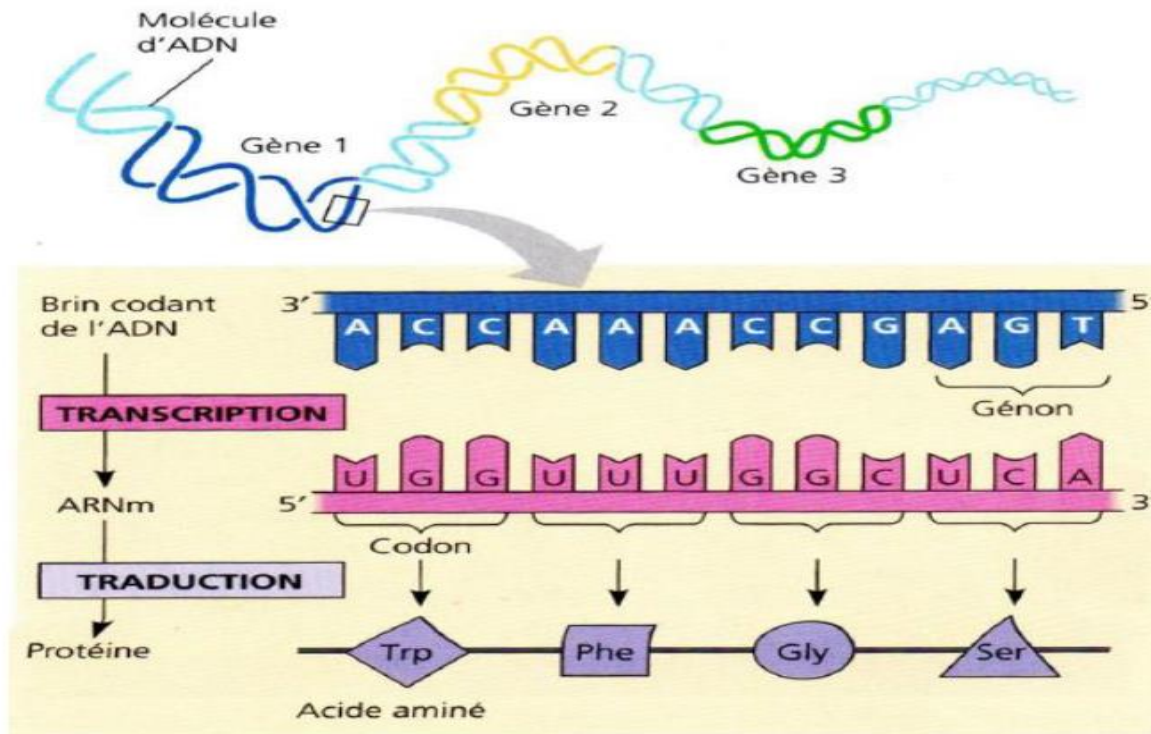
LES FONDEMENTS BIOLOGIQUES DE LA BIO-INFORMATIQUE

◦ Code génétique

	U		C		A		G		
U	UUU	phénylalanine	UCU	sérine	UAU	tyrosine	UGU	cystéine	U
	UUC		UCC		UAC		UGC		C
	UUA	leucine	UCA		UAA	stop	UGA	stop	A
	UUG		UCG		UAG		UGG	tryptophane	G
C	CUU	leucine	CCU	proline	CAU	histidine	CGU	arginine	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	glutamine	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	isoleucine	ACU	thréonine	AAU	asparagine	AGU	sérine	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	lysine	AGA	arginine	A
	AUG	méthionine	ACG		AAG		AGG		G
G	GUU	valine	GCU	alanine	GAU	acide aspartique	GGU	glycine	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	acide glutamique	GGA		A
	GUG		GCG		GAG		GGG		G

LES FONDEMENTS BIOLOGIQUES DE LA BIO-INFORMATIQUE

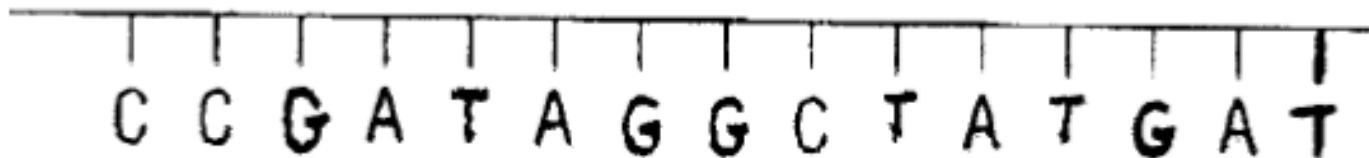
○ Synthèse de protéine



LES FONDEMENTS BIOLOGIQUES DE LA BIO-INFORMATIQUE

○ **Exercice:** La figure présente le schéma d'un brin appartient à une molécule d'ADN

- 1) Justifier cette affirmation.
- 2) Compléter cette molécule en expliquant votre démarche.
- 3) Écrire le brin d'ARN messenger qui correspond à cette molécule d'ADN.
- 4) Établir la séquence des acides aminés en utilisant le codage génétique.



LES FONDEMENTS BIOLOGIQUES DE LA BIO-INFORMATIQUE

- **Solution**

- Le brin schématisé dans la figure ci-dessus appartient à une molécule d'ADN car ce brin possède des bases azotées T.
- Démarche : par complémentarité de base, il faut positionner un A devant un T et un G devant un C.
- ARNm : GGCUAUCCGAUACUA.
- Protéine : Gly-Tyr-Pro-Ile-Leu.

LES BANQUES DE DONNÉES BIOLOGIQUES

- **Problème:**
- organiser l'énorme masse d'information et de la rendre disponible à l'ensemble de la communauté des chercheurs.
- **Solution: Banque de données biologiques**
 - 1- Banques généralistes
 - 2- Banques spécialisées

LES BANQUES DE DONNÉES BIOLOGIQUES

- **Banques généralistes**
- Collecte des données la plus exhaustive possible et offrant un ensemble hétérogène d'informations.
- Les Banques généralistes les plus populaires sont:

Banques ADN	Banques protéines
Genbank «USA »	PIR « USA »
EMBL « Europe »	SwissProt « Suisse »
DDBJ « Japon »	

LES BANQUES DE DONNÉES BIOLOGIQUES

- **Banques spécialisées**
- elles regroupent des données plus homogènes établies autour d'une thématique ou d'une méthode spécifique de production des données
- Chacun est susceptible de créer sa banque spécialisées
- SGD - Génome des Saccharomyces
- MGI - Génome de la souris
- TAXONOMY – Taxonomie
-

LES BANQUES DE DONNÉES BIOLOGIQUES

◦ Format

Certains sont simples comportant seulement la séquence et son nom. C'est le cas du format **FASTA**:

```
>Cytochrome [Homo sapiens]
MDPFVVLVLCLSCLLLLSIWRQSSGRGKLPPGPTPLPVIGNILQIDIKDVSKSLTNLS
KIYGPVFTLYFGLERMVVLHGYEVVKEALIDLGEFSGRGHFPLAERANRGFGIVFS
NGKRWKEIRRFSLMTLRNFGMGKRSIEDRVQEEARCLVEELRKTASPCDPTFILG
CAPCNVICSIIFQKRFDYKDQQFLNLMEKLNENIRIVSTPWIQICNNFPTIIDYFPGT
HNKLLKNLAFMESDILEKVKEHQESMDINNPRDFIDCFLIKMEKEKQNNQQSEFTIE
NLVITAADLLGAGTETTSTTLRYALLLLKHPEVTAKVQEEIERVIGRNRSPCMQDR
GHMPYTDVVHEVQRYIDLPTSLPHAVTCDVKFRNYLIPKGTTLTSLTSVLHDNK
EFPNPEMFDPRHFLDEGGNFKKSNYFMPFSAGKRICVGEGLARMELFLFLTFILQN
FNLKSLIDPKDLDTTPVVNGFASVPPFYQLCFIPV
```

LES BANQUES DE DONNÉES BIOLOGIQUES

◦ Format

D'autres sont plus compliqués et possèdent plus d'informations. C'est le cas du format GenBank:

Diagram illustrating the GenBank format structure with annotations:

- Définition et numéro d'accèsion** → LOCUS AAB59426 490 aa
- Organisme et taxonomie** → SOURCE Homo sapiens (human)
- Référence bibliographique** → REFERENCE 1 (residues 1 to 490)
- Caractéristiques de la séquence: Domaines, Motifs, RBS, CDS...** → FEATURES
- Séquence** → ORIGIN
- Descripteur de fin de fichier** → //

Annotations on the right side:

- linear → **Date de soumission** (PRI 31-DEC-1994)
- ↑
- Type de molécule: linéaire ou circulaire**

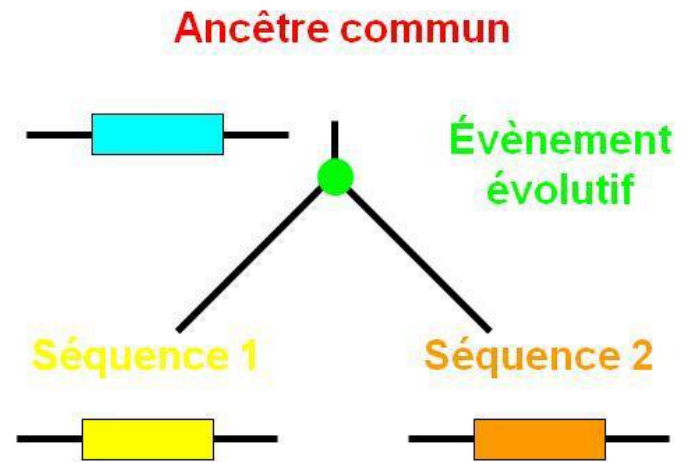
```
LOCUS       AAB59426               490 aa
DEFINITION  cytochrome.
ACCESSION   AAB59426
VERSION     AAB59426.1  GI:181344
DBSOURCE    locus HUMCYPC219 accession M61854.1
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1 (residues 1 to 490)
  AUTHORS   Romkes,M., Faletto,M.E., Blaisdell,J.A., Raucy,J.L. and
            Goldstein,J.A.
  TITLE     Cloning and expression of complementary DNAs for multiple members
            of the human cytochrome P450IIC subfamily
  JOURNAL   Biochemistry 30 (13), 3247-3255 (1991)
  PUBMED   2009263
COMMENT     Method: conceptual translation.
FEATURES             Location/Qualifiers
     source            1..490
                       /organism="Homo sapiens"
                       /db_xref="taxon:9606"
     Protein           1..490
                       /product="cytochrome"
     Region            30..487
                       /region_name="p450"
                       /note="Cytochrome P450; ptam00067"
                       /db_xref="CDD:40168"
     Region            30..490
                       /region_name="CypX"
                       /note="Cytochrome P450 [Secondary metabolites
                       biosynthesis, transport, and catabolism]; COG2124"
                       /db_xref="CDD:32307"
     CDS               1..490
                       /gene="CYP2C19"
                       /coded_by="M61854.1:6..1478"
ORIGIN
1 mdpfvvivic lscililsvi rqsgrgklip poptplpvig nilqididkv sksitnlski
61 ygvvftlyig lsmvvlbgv evvkeslidl geefsgqghf plaerantqf givfngkrv
121 keirrfelmt lrfwgqkris iedrvqear olveelrktk aspcdptfil qpapcnvics
181 iifqkrldyk dqgflnlmek lnenirivst pwiqicnnfp tiidyfpgth akllknlafm
241 esdilekvke hqesmdinnp rdfidoflik wekekqnqqs eftienlvit aadilgagte
301 ttattlryal llllkhpvt akvqeierv igrnrspcmq drghmpytda vvhevqvyid
361 liptslphav tcdvkrfnyl ipkgttlits ltsvlhdnke fpnpemfdpr hfideggntk
421 ksnymfmsa gkracvgegl armelfilft filqnfiks lidpkdiddt pvvngfasvp
481 pfyqlcflpv
//
```

HOMOLOGIE

- **L'homologie** c'est la relation entre les deux caractères qui sont descendus, le plus souvent avec une **divergence**, d'un caractère ancestral commun.
- **La divergence** de 2 séquence d'un ancêtre commun peut résulter d'une:
- **duplication**: dédoublement d'un segment d'ADN menant à la formation de plusieurs copies dans le même génome
- **spéciation**: formation d'espèces séparées « distinctes » à partir d'une espèce unique.

HOMOLOGIE

- Deux séquences sont dites homologues si elles possèdent un ancêtre commun.



HOMOLOGIE

- **ATTENTION: Homologie != similarité**
- L'homologie n'est pas quantifiable
Deux séquences ne sont pas à 50% ou à 75% homologue
Elles sont homologues ou elles ne le sont pas
→ Raisonnement binaire
- La similarité est quantifiable
Deux séquences peuvent être similaires à 50% ou 75%

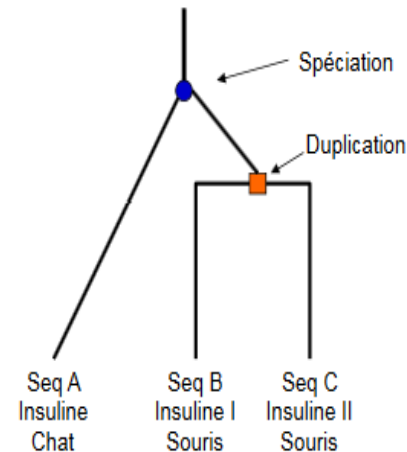
HOMOLOGIE

- Les catégories de l'homologie

- Orthologie:** deux gènes homologues sont dit *orthologue* si leur divergence est due à la spéciation.
- Paralogie:** deux gènes homologues sont dit *paralogues* et si leur divergence est due à la duplication.

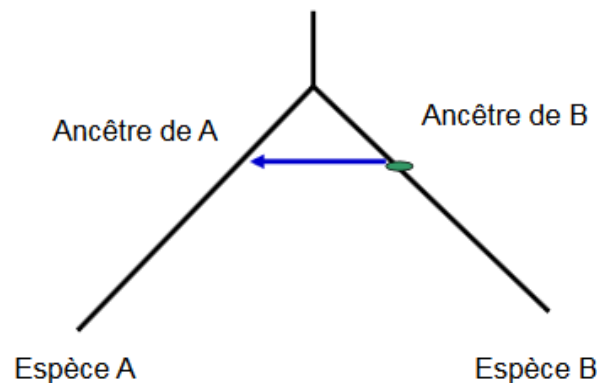
- Exemple

B et C sont paralogues.
A et C sont orthologues.
A et B sont orthologues.



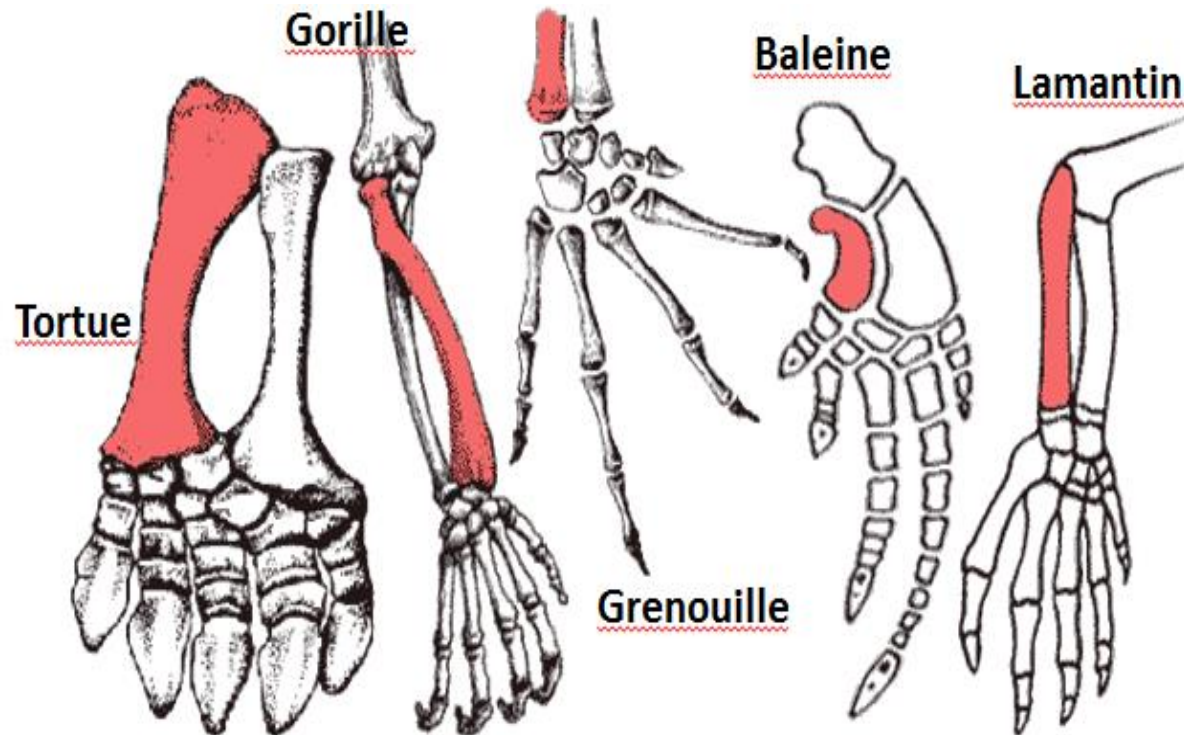
HOMOLOGIE

- Les catégories de l'homologie
- **Xénologie**: deux gènes homologues sont dits xénologues si leur divergence implique le transfert interspécifique du matériel génétique (horizontal).



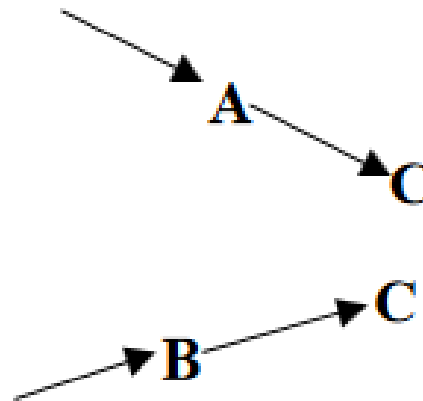
HOMOLOGIE

❖ Exemple: Squelette des membres de divers vertébrés



ANALOGIE

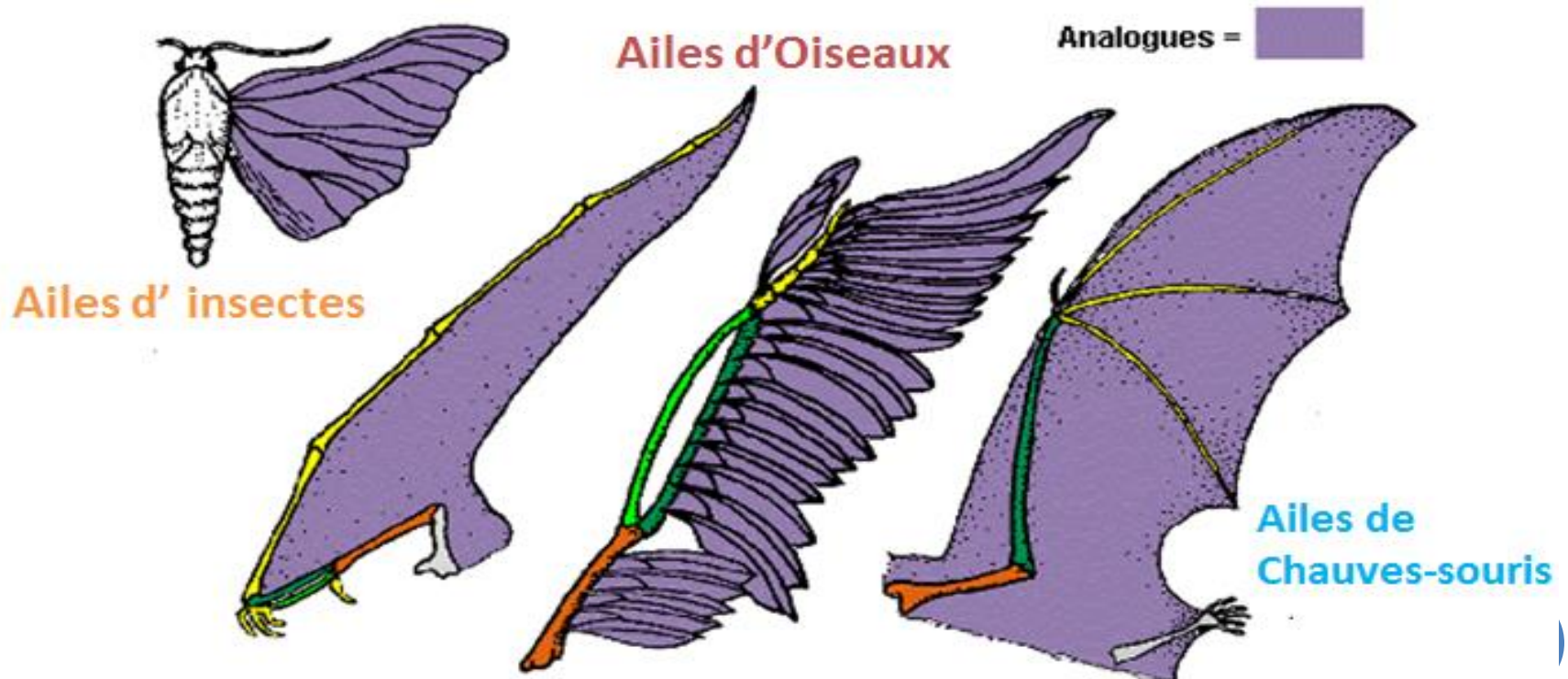
- L'analogie est la relation entre des caractères développés de manière convergente à partir d'ancêtres différents, contrairement à l'homologie.
- Elle évalue les similitudes fonctionnelles ou externes entre les organes d'espèces différentes, même si leurs structures et origines diffèrent, mais ils exécutent la même fonction.



ANALOGIE

- Exemple

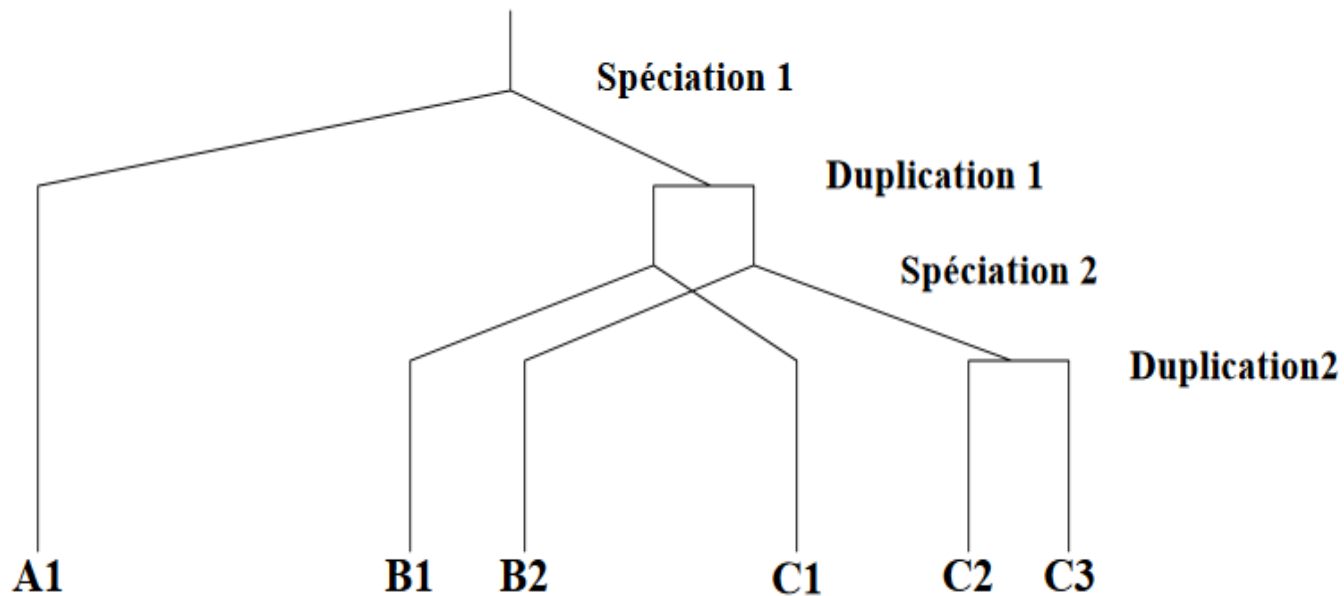
- Les Similitudes fonctionnelles entre les ailes des insectes, des oiseaux et celles des chauves-souris :



HOMOLOGIE

○ Exercice

Qualifiez la relation entre chaque paire de gènes dans le schéma suivant



HOMOLOGIE

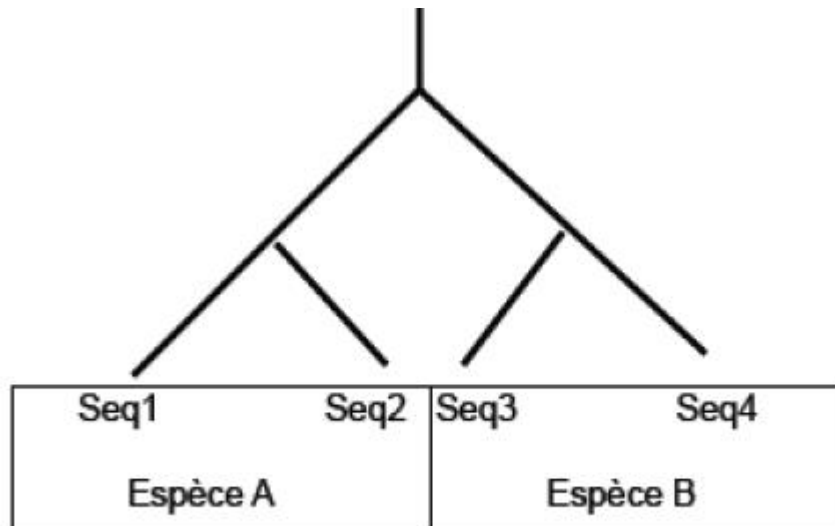
◦ Solution

- A1 est orthologue à B1, B2, C1, C2 et C3
- B1 est orthologue à C1 et paralogue à B2, C2 et C3
- B2 est orthologue à C2 et C3 et paralogue à B1, C1
- C1 est orthologue à B1 et paralogue à B2, C2, et C3
- C2 et C3 sont paralogues, ils sont orthologues à B2 et paralogues à B1 et C1

HOMOLOGIE

○ Exercice

Marquez les spéciations et les duplications sur l'arbre et déterminez le type d'homologie entre



HOMOLOGIE

○ Solution

- ▣ Seq1 et Seq2 paralogues
- ▣ Seq1 et Seq3 orthologues
- ▣ Seq1 et Seq4 orthologues
- ▣ Seq2 et Seq3 orthologues
- ▣ Seq2 et Seq4 orthologues
- ▣ Seq3 et Seq4 paralogues

