COURS DE BIOINFORMATIQUE

Dr. Berghiche Amine

Introduction à l'analyse des séquences et à la modélisation moléculaire

SUPPORT PÉDAGOGIQUE EN:

BIO-INFORMATIQUE

Avant Propos

Bio-informatique est la fusion de la biologie, de l'informatique et de la technologie dans une seule discipline de recherche qui analyse et interprète les données biologiques, en utilisant des programmes et des méthodes informatiques, pour créer de nouvelles connaissances biologiques.

Les défis de la biologie appliquée, en particulier dans le domaine de la biotechnologie, à l'interface avec l'informatique et les mathématiques appliquées et les méthodes informatiques de programmation et d'exploration de bases de données, vont générer de nouvelles connaissances et apporter des réponses aux attentes des chercheurs dans les domaines fondamentaux et appliqués de la biologie.

La bio-informatique permet d'analyser des données informatisées sur des organismes utiles ou pathogènes (microbiens, parasites, végétaux ou animaux) ; de concevoir in silico de nouveaux médicaments et produits chimiques ; d'étudier et de prévoir l'adaptation aux conditions environnementales biotiques et abiotiques (sécheresse, salinité, résistance aux pesticides/antibiotiques, interactions hôte-parasite, etc.)

Le présent support pédagogique est spécifiquement conçu pour accompagner les étudiants de deuxième année vétérinaire dans l'acquisition des connaissances fondamentales et appliquées relatives à la Bio-informatique, à la Biologie Moléculaire et à la Génétique.

Dans un monde où la médecine vétérinaire, la surveillance épidémiologique et la biotechnologie animale sont intrinsèquement liées à l'analyse du vivant à l'échelle moléculaire, la maîtrise des outils d'analyse des séquences est devenue indispensable. Que ce soit pour comprendre la résistance aux antibiotiques des pathogènes microbiens, pour analyser le génome d'une espèce menacée, ou pour appliquer les principes de la pharmacogénomique aux animaux de rente ou de compagnie, la bio-informatique se positionne comme la discipline pivot de la biologie moderne.

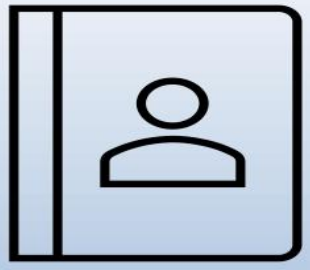
Ce document est conçu comme un pont entre la théorie biologique (le flux de l'information génétique, les mutations) et la pratique in silico (l'algorithmique d'alignement, la phylogénie). Il couvre un large éventail de modules, allant des fondements de l'ADN et des protéines, aux

technologies d'édition du génome (CRISPR/Cas9), en passant par les méthodes clés d'analyse de données (BLAST, arbres phylogénétiques).

Afin d'assurer une assimilation complète des concepts, chaque grande partie est complétée par des exercices d'application avec solutions et des figures thématiques, visant à transformer la compréhension théorique en compétences analytiques concrètes. Un lexique technique et des références ciblées sont également inclus pour encourager l'autonomie et l'approfondissement.

Nous encourageons vivement les futurs praticiens à s'approprier ces concepts, car ils sont la clé pour décrypter le langage génétique et pour innover dans les domaines du diagnostic, de la prévention et du traitement des maladies animales.

Dr BERGHICHE Amine
Maitre de conference A



Fiche Contact

Fiche-contact

Cours Bio-informatique

Faculté des sciences agronomiques et vétérinaires

Cours 1h30 par semaine

Lieu et Horaire de *Cours* :

Année Universitaire: 2024/2025

Modalité de suivi (calendrier du tutorat): Tout les Marid13h a 16h au niveau de la bibliothèque de l'institut.

Fiche-contact

Enseignant de la matière : Dr Berghiche Amine,

Contacts : amine_berghiche@yahoo.com / a.berghiche@univ-soukahras.dz ·

Coefficient: 2 · Crédits : /

Volume horaire global: 45h·

Volume horaire de travail requis/semaine: 1h.30 ·

Modalité d'évaluation: Examen final 60 % et évaluation continue 40 %·

Sommaire

Préambule : L'Ère de l'Information Biologique

I. L'Évolution des Sciences du Vivant

Le début du XXIe siècle marque une révolution dans les sciences biologiques, caractérisée par une transition d'une approche réductrice (étude de molécules ou de gènes isolés) à une **approche systémique et globale**. Cette transformation est principalement due à l'explosion des données génétiques et protéiques générées par les technologies de séquençage à haut débit. L'ère actuelle est celle de l'**Information Biologique**, où la puissance de calcul est aussi essentielle que la paillasse du laboratoire.

Discipline	**Définition Clé**	**Approche**
Biologie Moléculaire	L'étude des molécules (ADN, ARN, Protéines) qui orchestrent la vie et le **flux de l'information génétique** (le Dogme Central).	**Wet Lab** (*in vitro* et *in vivo*)
Génomique	Science qui vise à la cartographie, au séquençage et à l'étude fonctionnelle des **génomes** dans leur entièreté.	**Wet Lab & Dry Lab**
Bio-informatique	L'utilisation des outils et méthodes informatiques pour la gestion, l'analyse et l'interprétation des données biologiques.	**Dry Lab** (*in silico*)

II. Les Piliers du Flux de l'Information Génétique

La compréhension de l'information biologique repose sur le concept central de la **Biologie Moléculaire** : le flux de l'information génétique, connu sous le nom de **Dogme Central**.

1. **ADN (Acide Désoxyribonucléique) : Le Support**
    - ✓ L'ADN est la molécule support de l'hérédité. Il est structuré en une **double hélice** et sa fonction est de **stocker l'information génétique** (les gènes) de manière stable. L'information est codée par la séquence des quatre bases azotées (Adénine, Cytosine, Guanine, Thymine).
    - ✓ Le processus de **Réplication** assure la transmission fidèle de cet ADN de génération en génération cellulaire.

2. **ARN (Acide Ribonucléique) : L'Intermédiaire**
    - ✓ L'ARN est l'intermédiaire du message génétique.
    - ✓ La **Transcription** est l'étape où un segment d'ADN (un gène) est copié en ARN messager (ARNm).
3. **Protéines : Les Effecteurs**
    - ✓ Les protéines sont les molécules fonctionnelles qui exercent la majorité des fonctions cellulaires (enzymes, transporteurs, structures, etc.).
    - ✓ La **Traduction** est l'étape où le message de l'ARNm (lu par les ribosomes) est décodé en séquence d'acides aminés, formant la protéine finale. Les règles de ce décodage sont données par le **Code Génétique** universel.

III. L'Avènement de la Bio-informatique

Historiquement, la Biologie Moléculaire a été limitée par la difficulté à manipuler et analyser les longues séquences d'ADN et le grand nombre de protéines. L'émergence des disciplines **"Omics"** (Génomique et Protéomique) a rendu l'informatique indispensable.

1. Rôle et Nécessité

La **Bio-informatique** est définie comme l'approche « **in silico** » de la biologie traditionnelle. Elle utilise les outils mathématiques, statistiques et informatiques pour :

- **Stockage et Gestion :** Organiser les téraoctets de données de séquences dans des bases de données structurées (ex : GenBank, SwissProt).
- **Traitement et Analyse :** Développer et appliquer des **algorithmes** (comme BLAST ou Needleman-Wunsch) pour identifier des similarités, prédire des fonctions, et analyser les relations évolutives (phylogénie).
- **Interprétation :** Transformer des alignements complexes et des profils d'expression génique en **connaissances biologiques exploitables**.

2. La Convergence des Disciplines

La **Protéomique** (étude du Protéome, l'ensemble des protéines) génère une quantité effarante de données expérimentales (ex : profils de l'électrophorèse bidimensionnelle) que seule la Bio-informatique peut classer, interpréter et stocker efficacement.

La **Génomique** dépend également de la Bio-informatique pour **assembler** les fragments d'ADN séquencés (le "puzzle génétique") et pour identifier les **SNP** (Polymorphismes d'un seul nucléotide), essentiels à la **Pharmacogénomique** (médecine personnalisée).

PARTIE I

FONDEMENTS DE LA BIOLOGIE MOLÉCULAIRE ET DU FLUX DE L'INFORMATION GÉNÉTIQUE

🧬 PARTIE I : FONDEMENTS DE LA BIOLOGIE MOLÉCULAIRE ET DU FLUX DE L'INFORMATION GÉNÉTIQUE

Chapitre 1 : L'ADN, Support de l'Hérédité

1.1. Structure et Composition des Acides Nucléiques

L'**ADN** (Acide Désoxyribonucléique) constitue le **support universel de l'information génétique** chez la plupart des organismes. Sa fonction première est d'assurer le stockage, la transmission et l'expression de ce patrimoine héréditaire. L'ADN est une macromolécule dont l'unité de base est le **nucléotide**.

1.1.1. Le Nucléotide : Structure Chimique Fondamentale

Chaque nucléotide est une entité complexe et polarisée, composée de trois éléments clés liés de manière covalente :

1. **Un groupement phosphate** (ou acide phosphorique) : Identique pour tous les nucléotides de l'ADN et de l'ARN, il confère une charge négative à la molécule et forme la structure externe (ou **squelette**) du brin.
2. **Un sucre à cinq atomes de carbone (pentose)** :
    - Le **désoxyribose** dans le cas de l'ADN (car il a perdu un atome d'oxygène sur le carbone 2').
    - Le **ribose** dans le cas de l'ARN.
3. **Une base azotée variable** : C'est la partie variable et informationnelle de la molécule.
    - **Bases Puriques (double cycle)** : Adénine (A) et Guanine (G).
    - **Bases Pyrimidiques (cycle simple)** : Cytosine (C) et **Thymine (T)** pour l'ADN ; Cytosine (C) et **Uracile (U)** pour l'ARN.

1.1.2. La Double Hélice et la Complémentarité des Bases

L'ADN se présente sous la forme d'une **double hélice**. Deux brins simples, orientés dans des directions opposées (**antiparallèles**), s'enroulent l'un autour de l'autre autour d'un axe central.

- **Squelette sucre-phosphate :** Il est régulier, hydrophile, et se situe à l'extérieur de la structure.
- **Bases azotées :** Elles sont hydrophobes et s'empilent au centre de l'hélice, formant les "barreaux de l'échelle".

La cohésion des deux brins est assurée par la **règle de complémentarité des bases** :

- L'**Adénine (A)** s'apparie toujours avec la **Thymine (T)**, via deux liaisons hydrogène.
- La **Guanine (G)** s'apparie toujours avec la **Cytosine (C)**, via trois liaisons hydrogène.

Cette structure est essentielle pour l'encodage de l'information et le processus de réplication.

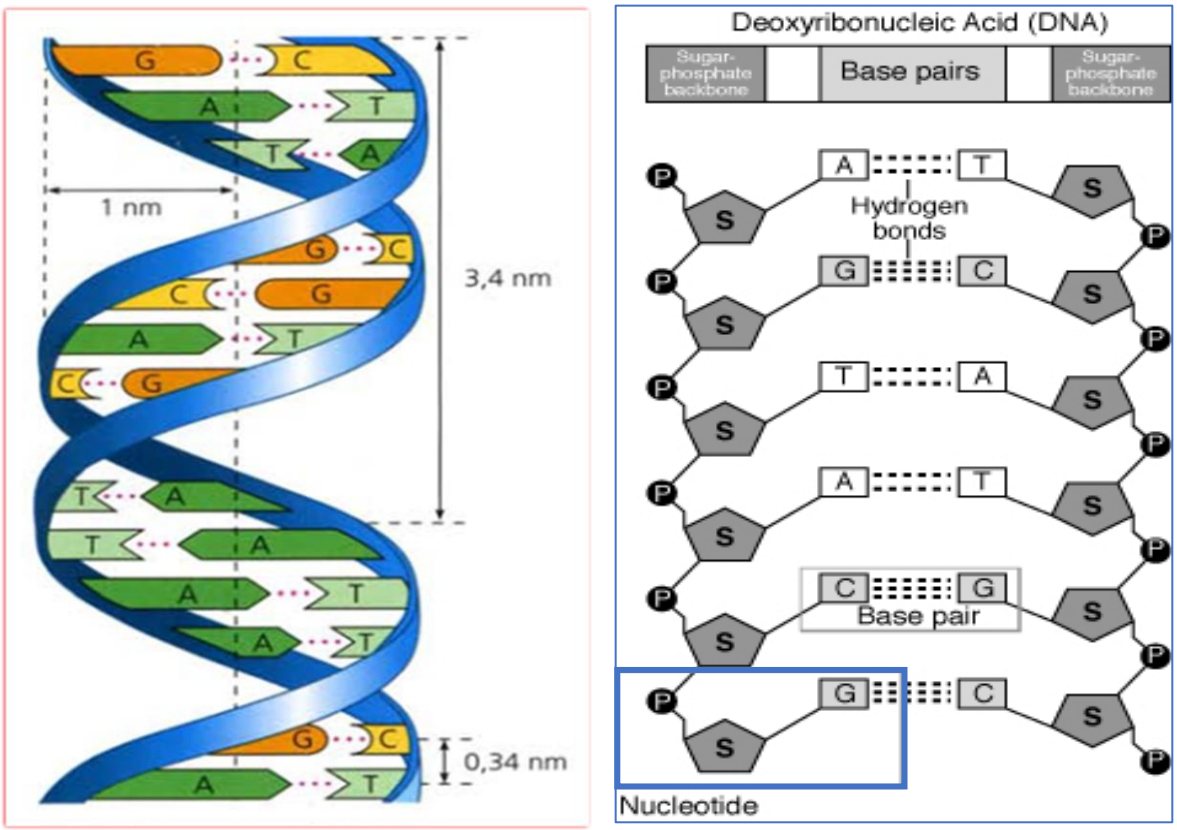


Schéma 1.1 : Représentation détaillée du nucléotide et de la double hélice d'ADN, incluant les liaisons hydrogène (A-T et C-G)

1.2. Organisation du Matériel Génétique

1.2.1. Le Gène, Locus et Unité d'Hérédité

Le **gène** est la fraction fonctionnelle de l'ADN. Il est défini comme l'unité d'hérédité contrôlant un caractère particulier et correspondant à un segment d'ADN (ou d'ARN pour certains virus). Fondamentalement, un gène est **toute région de l'ADN qui produit une molécule d'ARN fonctionnelle**.

- Chaque gène est localisé à un emplacement bien précis sur un chromosome, appelé le **locus**.

1.2.2. Le Génome et le Chromosome

- **Le Chromosome :** Chez les eucaryotes, l'ADN est structuré en complexes appelés chromosomes. Le noyau cellulaire est assimilé à la « **bibliothèque** » qui renferme le patrimoine héréditaire, et le chromosome est un « **livre** » de cette bibliothèque. L'être humain possède 23 paires de chromosomes (22 paires autosomales et une paire de chromosomes sexuels - XX ou XY).
- **Le Génome :** C'est l'ensemble complet du matériel génétique (ADN) d'un individu ou d'une espèce. Le **Génome Humain** est composé de plus de 25 000 gènes.

1.3. La Réplication de l'ADN (Synthèse d'ADN)

La réplication est le processus essentiel qui permet à l'ADN d'être reproduit à l'identique, garantissant ainsi la transmission fidèle de l'information génétique lors de la division cellulaire.

1.3.1. Principe de la Réplication Semi-Conservatrice

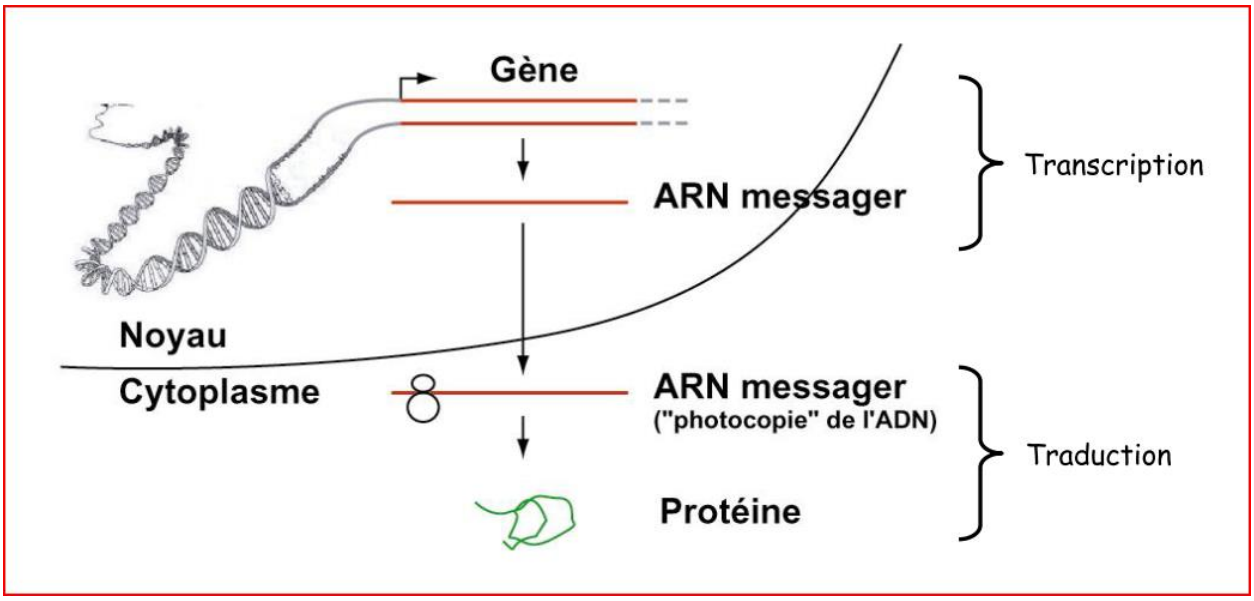
La réplication est dite **semi-conservatrice**. Ce terme signifie que chaque nouvelle molécule d'ADN synthétisée est composée d'un **brin parental (ancien)** et d'un **brin néoformé (nouveau)**.

- Le mécanisme commence par la **dénaturation** locale de l'ADN : la molécule s'ouvre comme une « fermeture éclair » par rupture des liaisons hydrogène faibles entre les bases appariées.
- Chaque brin parental sert ensuite de **matrice** pour la synthèse de son brin complémentaire, grâce à l'incorporation de nucléotides libres selon la règle A-T/C-G.

1.3.2. Le Mécanisme Moléculaire : Le Réplisome

La synthèse d'ADN est catalysée par un complexe enzymatique appelé **réplisome**. Les enzymes clés sont :

- **Hélicase :** Dénature l'ADN en rompant les liaisons hydrogène et en déroulant la double hélice.
- **Primase :** Synthétise une petite séquence d'ARN, appelée **amorce**, qui est indispensable pour que l'ADN polymérase puisse démarrer la synthèse.
- **ADN Polymérase :** L'enzyme principale. Elle catalyse l'élongation du nouveau brin dans le sens **5'-3'**. Elle possède également des fonctions de relecture et de correction des erreurs.
- **Ligase :** Soude les fragments d'ADN sur le brin retardé.



📝 **Exercices d'Application avec Solutions**

Exercice 1: Complémentarité et Réplication de l'ADN

Contexte : Vous analysez une séquence d'ADN double brin d'un virus pathogène. L'un des brins est le suivant (lu de 5' vers 3'):

5' → ATG CAC GGT TAA →3'

Questions :

1. Déterminez la séquence du **brin complémentaire** (en indiquant son orientation 3'→5').
2. Expliquez brièvement le principe de la réplication de l'ADN en utilisant le terme **semi-conservateur**.

Solution

1. Séquence du brin complémentaire :

Selon la règle de complémentarité (A s'apparie avec T, C avec G) et le sens antiparallèle (5'→ 3' en face de 3' →5'):

- **Brin complémentaire :**

3'→TAC GTG CCA ATT →5'

Principe de la réplication semi-conservatrice :

- La réplication de l'ADN est dite **semi-conservatrice** car chaque nouvelle molécule d'ADN double brin formée est constituée d'un **brin parental (ancien)** et d'un **brin nouvellement synthétisé** (néoformé). Le brin parental sert de matrice pour la synthèse du nouveau brin.

Chapitre 2 : Expression des Gènes (Du Gène à la Protéine)

L'expression génétique est le processus en deux étapes (Transcription puis Traduction) par lequel l'information encodée dans l'ADN est utilisée pour synthétiser des protéines fonctionnelles. C'est le cœur du **Dogme Central de la Biologie Moléculaire**.

2.1. La Transcription (ADN vers ARNm)

La **transcription** est la synthèse d'une molécule d'ARN messager (ARNm) à partir d'un gène matriciel sur l'ADN.

2.1.1. Processus et Localisation

- **Localisation :** Se déroule dans le **noyau** chez les eucaryotes, mais dans le cytoplasme chez les procaryotes.
- **L'ARN Polymérase :** C'est le complexe enzymatique responsable de la synthèse de l'ARN. Elle reconnaît la séquence promotrice du gène, ouvre l'hélice et assemble les ribonucléotides selon la matrice d'ADN.

2.1.2. Maturation de l'ARNm (chez les Eucaryotes)

Chez les eucaryotes, l'ARN messager primaire (pré-ARNm) subit un traitement intensif avant d'être exporté vers le cytoplasme :

- **Ajout d'une coiffe 5' et d'une queue poly-A 3' :** Ces modifications protègent l'ARNm de la dégradation et sont cruciales pour l'initiation de la traduction.
- **Épissage (Splicing) :** Les séquences non codantes (**introns**) sont excisées, et les séquences codantes (**exons**) sont ligaturées ensemble pour former l'ARNm mature et fonctionnel.

2.2. La Traduction (ARNm vers Protéine)

La **traduction** est le déchiffrement de la séquence de nucléotides de l'ARNm en séquence d'acides aminés (protéine). Elle a lieu dans le **cytoplasme**.

2.2.1. Le Code Génétique et le Codon

- **Codon :** L'unité d'information élémentaire est un **triplet de nucléotides** (A, C, U, ou G) de l'ARNm, appelé codon.
- **Universalité et Redondance :** Le code génétique est quasiment **universel** (commun à tous les êtres vivants) et **redondant** (ou dégénéré), car plusieurs codons peuvent coder pour le même acide aminé.

- **Codons-Stop :** Sur les 64 codons possibles, 61 désignent un acide aminé, et 3 (**UAA, UGA, UAG**) provoquent l'arrêt de la synthèse.

2.2.2. Les Acteurs Clés de la Synthèse Protéique

- **Ribosomes :** Complexes ribonucléoprotéiques (composés d'ARN ribosomal et de protéines) agissant comme l'usine d'assemblage. Ils se dissocient en une petite sous-unité (40S) et une grande sous-unité (60S) après la terminaison.
- **ARN de Transfert (ARNt) :** Molécules adaptatrices qui reconnaissent un codon spécifique de l'ARNm et y acheminent l'acide aminé correspondant vers le ribosome.

2.2.3. Phases de la Traduction

1. **Initiation :** L'ARNm se lie à la petite sous-unité du ribosome. Le processus démarre au **codon-initiateur AUG**, qui code pour la méthionine.
2. **Élongation :** Le ribosome se déplace le long de l'ARNm (**translocation**), et les ARNt successifs apportent les acides aminés qui sont liés entre eux par des **liaisons peptidiques**, formant la **chaîne polypeptidique**.
3. **Terminaison :** Le ribosome atteint un codon-stop. L'ensemble se dissocie, libérant la protéine complète et l'ARNm, qui peut ensuite être utilisé pour de nouvelles synthèses (**polysome**).

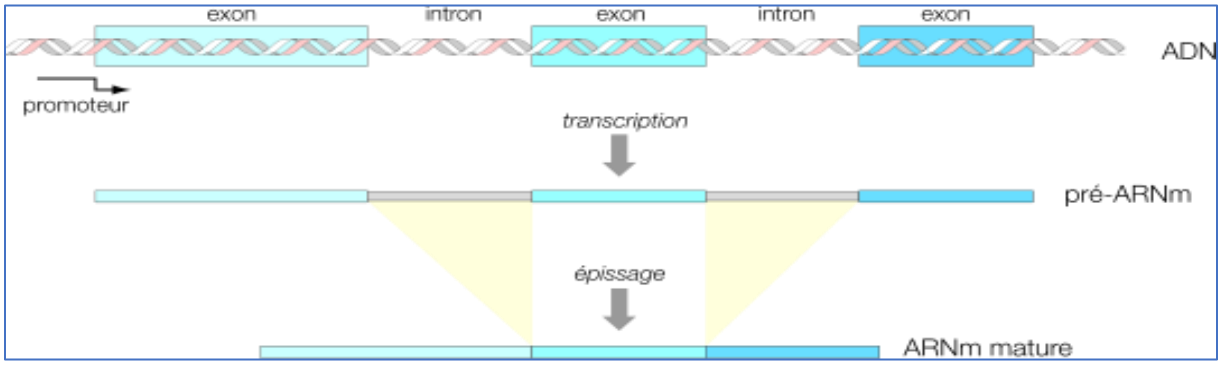


Schéma 2.2 : Représentation schématique du Dogme Central (ADN → ARN → Protéine).

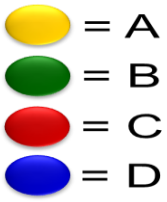
2.2.4. Code génétique : désigne le système de correspondance mis en jeu lors de la transformation de l'information génétique des gènes en protéines, au cours du processus de traduction. Les ribosomes traduisent ainsi, en suivant ce code, l'enchaînement des bases nucléotidiques de l'ARN en une séquence d'acides aminés appelée séquence peptidique. Le code génétique est un code de

longueur fixe, lu trois nucléotides par trois nucléotides, un triplet de nucléotides ou codon correspondant à un acide aminé donné.

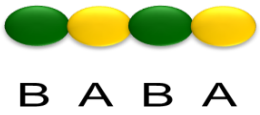
Pour comprendre ce code, utilisons une analogie :

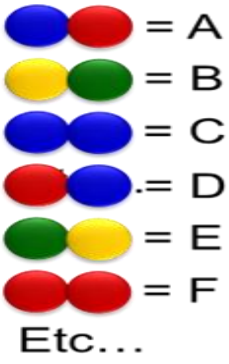
Supposons que vous voulez écrire un message secret en utilisant des billes de couleur que vous enfilez sur un fil. Vous devez donc faire correspondre des billes à des lettres de l'alphabet. Supposons aussi que vous disposez de seulement quatre couleurs de billes:

rouge, jaune, vert et **bleu** : Seulement 4 couleurs constituent 4 nucléotides de l'ADN.

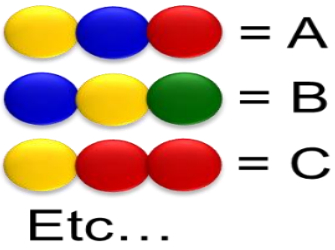


Notre code ne permettrait que de désigner **4** lettres sur les 26 de l'alphabet

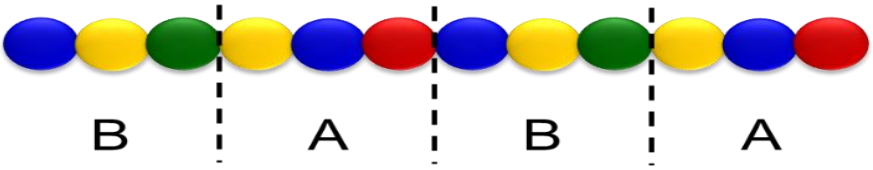




C'est déjà mieux, on a maintenant **16** lettres, **combinaisons possibles** (4^2).
Malheureusement, ce n'est pas assez

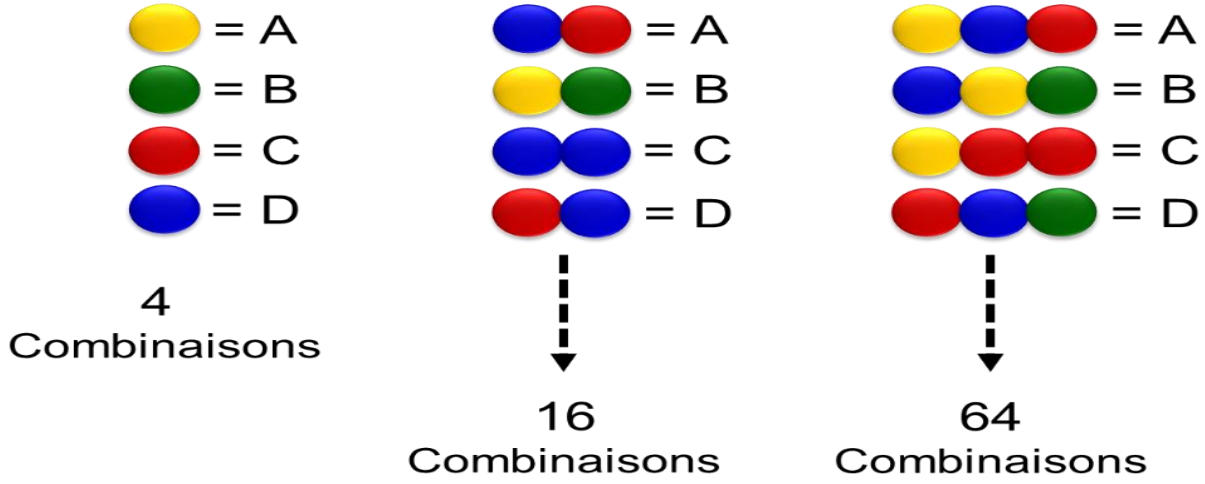


on peut former **64** **combinaisons différentes** (4^3), c'est plus que ce qui est nécessaire pour coder 26 lettres



Remplaçons les billes par des nucléotides, **A**, **T**, **C** et **G**.

On pourrait imaginer un code où chaque groupe de trois nucléotides correspondrait non pas à une lettre de l'alphabet mais à un des **20** acides aminés formant les protéines. On pourrait ainsi former des messages correspondant à la recette d'une protéine.



Convenons, par exemple, que les trois nucléotides A-A-A représentent l'acide aminé phénylalanine, G-A-C représentent la leucine, T-C-T l'arginine et ainsi de suite.

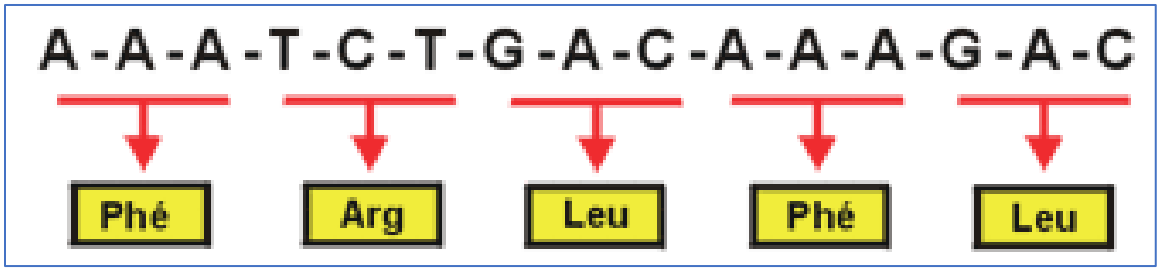
A-A-A = Phé

G-A-C = Leu

T-C-T = Arg

Note bien que La protéine formée des cinq acides aminés : Phé-Arg-Leu-Phé-Leu

pourrait être représentée (codée) par la chaîne formée des nucléotides suivante :



	U	C	A	G	
U	UUU Phe (F) UUC " UUA Leu (L) UUG "	UCU Ser (S) UCC " UCA " UCG "	UAU Tyr (Y) UAC UAA Ter UAG Ter	UGU Cys (C) UGC UGA Ter UGG Trp (W)	U C A G
C	CUU Leu (L) CUC " CUA " CUG "	CCU Pro (P) CCC " CCA " CCG "	CAU His (H) CAC " CAA Gln (Q) CAG "	CGU Arg (R) CGC " CGA " CGG "	U C A G
A	AUU Ile (I) AUC " AUA " AUG Met (M)	ACU Thr (T) ACC " ACA " ACG "	AAU Asn (N) AAC " AAA Lys (K) AAG "	AGU Ser (S) AGC " AGA Arg (R) AGG "	U C A G
G	GUU Val (V) GUC " GUA " GUG "	GCU Ala (A) GCC " GCA " GCG "	GAU Asp (D) GAC " GAA Glu (E) GAG "	GGU Gly (G) GGC " GGA " GGG "	U C A G

Schéma 2.2 : Représentation schématique du Code Génétique

📝 Exercices d'Application avec Solutions

Exercice: Transcription et Traduction (Le Dogme Central)

Voici la séquence d'un brin d'ADN matrice (brin transcrit) d'un gène.

Le codon initiateur sur l'ARNm sera AUG.

3' →TAC GCT ACT GCC →5'

Questions :

1. Déterminez la séquence de l'**ARNm** mature (en précisant son sens).
2. Déterminez la séquence des **trois premiers acides aminés** de la protéine correspondante.

Code Génétique (rappel) : AUG = Méthionine (Start) ; GCU = Alanine ; CGA = Arginine ; UGA = Stop.

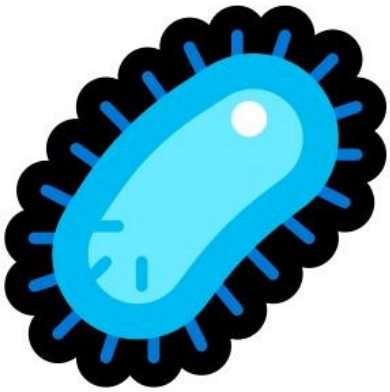
Solution

1. Séquence de l'ARNm :

L'ARNm est complémentaire et antiparallèle à l'ADN matrice (T → A, A →U, G → C, C →G) :

    - **ARNm :** 5' →AUG CGA UGA CGG →3'

2. **Séquence d'acides aminés :**
    - Codon 1 (AUG) : **Méthionine** (Start)
    - Codon 2 (CGA) : **Arginine**
    - Codon 3 (UGA) : **STOP**
    - **Séquence :** Méthionine - Arginine - STOP.



PARTIE II
ANOMALIES, VARIATIONS ET LES "OMICS"

🖇️ PARTIE II : ANOMALIES, VARIATIONS ET LES "OMICS"

Chapitre 3 : Les Mutations Génétiques et leurs Conséquences

Une **mutation** est une modification permanente de la séquence de l'information génétique (ADN ou ARN) dans le génome. Ces altérations sont la principale source de **diversité génétique** et sont le moteur essentiel de l'évolution des espèces.

3.1. Classification et Types de Mutations

Les mutations sont classées selon la nature et l'étendue de l'altération qu'elles provoquent.

3.1.1. Mutations selon leur Localisation

- **Mutations Autosomiques :** Elles touchent les chromosomes non-sexuels (autosomes).
- **Mutations Sexuelles :** Elles affectent les chromosomes sexuels (X ou Y).

3.1.2. Mutations Ponctuelles (Substitution)

Une mutation ponctuelle affecte **un seul nucléotide**. La substitution (remplacement d'une base par une autre) est la forme la plus courante. Les conséquences sur la protéine codée varient considérablement, en fonction de la modification du codon résultant :

Type de Mutation Ponctuelle	**Description et Effet Moléculaire**	**Conséquence Fonctionnelle**
Faux-sens	Changement d'un nucléotide qui entraîne la modification de l'acide aminé codé.	Peut avoir une répercussion sur la structure spatiale et la fonction de la protéine, ou être neutre si le nouvel acide aminé est chimiquement similaire.
Non-sens	Changement qui convertit un codon codant un acide aminé en un **codon-stop** (UAA, UGA, UAG).	Arrêt prématuré de la traduction. La protéine est tronquée et souvent non fonctionnelle.
Silencieuse	Changement de nucléotide qui, grâce à la **redondance du code génétique**, ne modifie pas l'acide aminé codé.	Aucune conséquence sur la protéine ou sur l'organisme.
Neutre	Changement de nucléotide qui n'affecte pas la capacité de reproduction (valeur sélective).	Ne modifie pas la capacité à se reproduire (ex : mutations des groupes sanguins).

3.1.3. Insertions, Délétions et Décalage du Cadre de Lecture

Les insertions (ajout d'une ou plusieurs paires de bases) et les délétions (retrait d'une ou plusieurs paires de bases) ont des conséquences particulièrement graves si elles ne sont pas un multiple de trois. Dans ce cas, elles provoquent un **décalage du cadre de lecture (*frameshift*)**, modifiant tous les codons en aval de la mutation et aboutissant presque toujours à une protéine non fonctionnelle et à terminaison prématurée.

3.2. Conséquences Biologiques et Évolution

Les mutations sont considérées comme le **matériau brut de l'évolution**.

- **Sélection Naturelle :** Les mutations **délétères** (défavorables) sont éliminées. Les mutations **avantageuses** (rares) tendent à s'accumuler.

- **Dérive Génétique :** Les mutations **neutres** n'influencent pas la valeur sélective et peuvent se fixer ou disparaître au hasard par le jeu de la dérive génétique.

📝 Exercices d'Application avec Solutions

Exercice: Classification des Mutations

Une séquence d'ARNm normale et deux séquences mutées sont données.

Séquence	Codons	Acides Aminés
Normal	**5'-GAA GAG UUA-3'**	Glutamate - Glutamate - Leucine
Muté A	**5'-GAA GAU UUA-3'**	Glutamate - Aspartate - Leucine
Muté B	**5'-GAA G UUU A...-3'**	Glutamate - *décalage du cadre*

Questions :

1. Quelle est la nature du changement dans la Séquence Mutée A et quel type de mutation en résulte ?
2. Quelle est la nature du changement dans la Séquence Mutée B et quel est son effet principal sur la protéine ?

Solution

1. **Séquence Mutée A :**

✓ **Nature du changement :** Substitution du nucléotide G par U dans le deuxième codon (GAG → GAU).

✓ **Type de mutation : Mutation Faux-Sens** (*Missense*), car elle entraîne le remplacement d'un acide aminé (Glutamate) par un autre (Aspartate).

2. **Séquence Mutée B :**

✓ **Nature du changement :** Délétion d'un nucléotide (G) dans le deuxième codon.

- ✓ **Effet principal :** La délétion d'un nucléotide (non-multiple de trois) provoque un **décalage du cadre de lecture** (*Frameshift*), altérant tous les codons et donc tous les acides aminés en aval du point de mutation.

Chapitre 4 : Génomique, Protéomique et Applications

Ces disciplines, regroupées sous le terme d'« **Omics** », permettent l'étude globale et systématique de grandes classes de molécules biologiques (ADN, ARN, protéines).

4.1. Définitions et Concepts Fondamentaux

4.1.1. La Génomique

La **Génomique** est la science qui s'intéresse à l'étude des génomes dans leur ensemble. Elle couvre un large éventail d'analyses :

- **Cartographie :** Localisation et établissement de l'ordre des gènes sur les chromosomes.
- **Séquençage :** Détermination de la séquence nucléotidique complète du génome.
- Fonction : Étude de la fonction des gènes identifiés.

  Le Projet Génome Humain a été un catalyseur pour cette discipline, révélant la complexité de l'ADN eucaryote (gènes découpés et éparpillés).

4.1.2. La Protéomique

- **Protéeome :** L'ensemble des protéines exprimées par une cellule, un tissu ou un organisme à un moment donné.
- **Caractère dynamique du Protéome :** Contrairement au génome qui est **stable**, le protéome est **dynamique** car il varie constamment en fonction du temps, de l'environnement, du stade de développement, du tissu, et de l'état (sain ou malade).
- **Protéomique :** Domaine de recherche qui vise à l'identification, la quantification et la caractérisation des protéines pour comprendre le fonctionnement de la machinerie cellulaire.

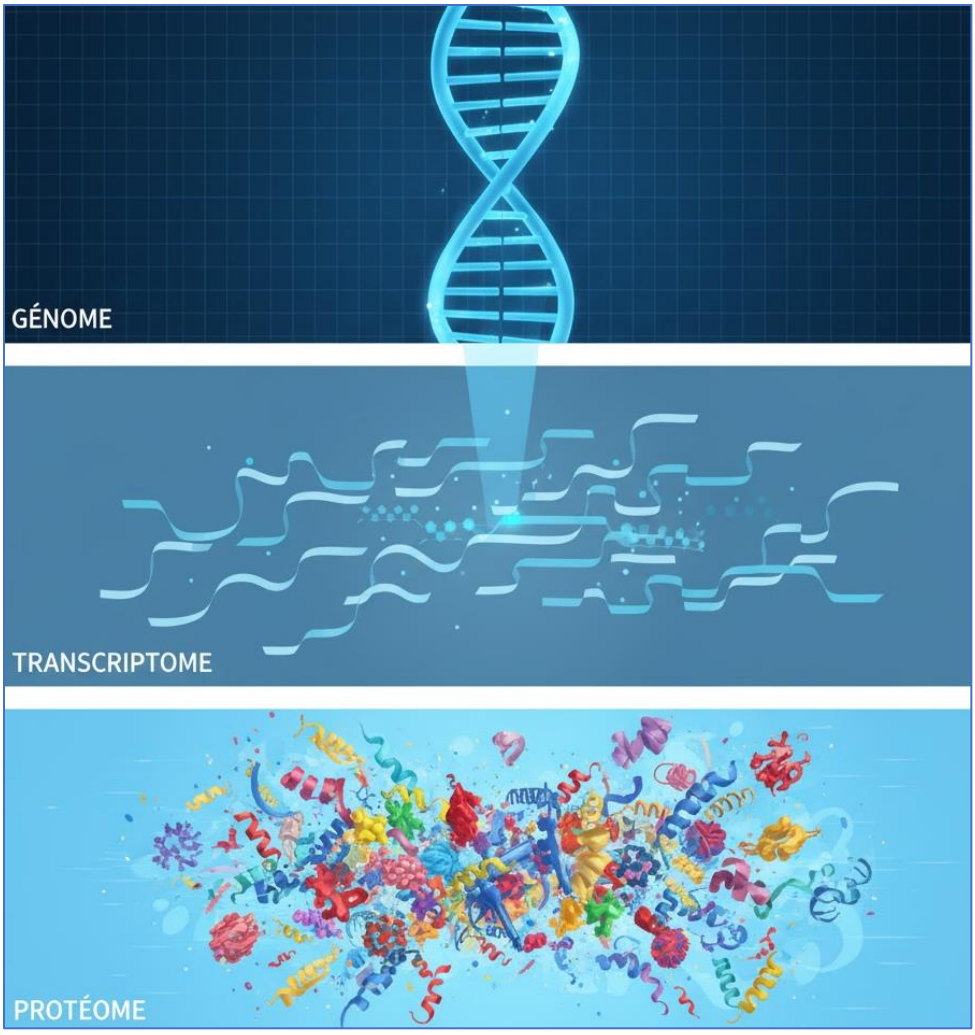


Figure 4.1 : Le Dogme Central et l'Interconnexion "Omics"

4.2. Techniques d'Analyse Protéomique

Les protéines étant les indicateurs de l'état cellulaire, leur analyse simultanée est cruciale.

4.2.1. Électrophorèse Bidimensionnelle (2D-E)

Développée par O'Farell en 1975, c'est la technique de référence pour la séparation massive des protéines. Elle est réalisée en deux étapes orthogonales :

1. **Isoélectrofocalisation (IEF) :** Les protéines sont séparées selon leur **point isoélectrique (pI)**, c'est-à-dire le pH pour lequel leur charge nette est nulle.
2. Électrophorèse sur Gel de Polyacrylamide en présence de SDS (SDS-PAGE) : Dans la deuxième dimension, les protéines sont séparées selon leur masse moléculaire (MW).

   Le résultat est un profil protéique contenant des centaines de "spots" (protéines individuelles), dont la comparaison entre échantillons (ex : cellule saine vs cellule cancéreuse) permet de déceler des protéines associées à un état pathologique.

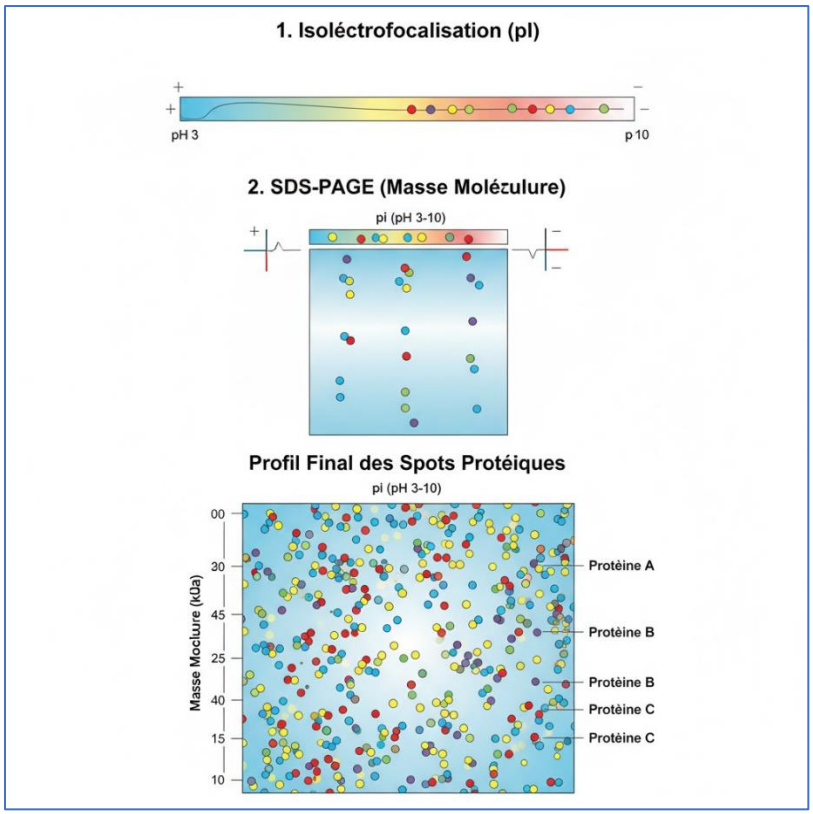


Figure 4.2 : L'Électrophorèse Bidimensionnelle (2D-E)

4.2.2. Autres Méthodes de Séparation et Purification

- **Chromatographie en Phase Liquide (HPLC) :** Technique courante de chimie analytique où la séparation repose sur la charge électrique des molécules dans un mélange, qui sont faites passer sur une phase stationnaire chargée.

- **Chromatographie par Affinité :** Séparation basée sur l'**affinité biologique ou fonctionnelle** de la molécule cible pour un composant spécifique de l'adsorbant (ex : utilisation d'un anticorps spécifique).
- **Ultracentrifugation :** Séparation des protéines ou complexes protéiques en fonction de leur **coefficient de sédimentation** dans un gradient.

4.3. Applications Clés de la Génomique et de la Protéomique

4.3.1. Pharmacogénomique : Vers la Médecine Personnalisée

Cette discipline vise à personnaliser les traitements en tenant compte des **variations génétiques individuelles** pour optimiser l'efficacité et la sécurité des médicaments.

- **Polymorphismes d'un seul Nucléotide (SNP) :** Ces variations correspondent au remplacement d'une seule base dans la séquence d'ADN et peuvent modifier l'activité des enzymes de métabolisation des médicaments.
- **Objectifs :** Développer des tests prédictifs de la réponse d'un malade au traitement et concevoir des médicaments adaptés à des **profils génétiques** spécifiques (médecine de précision).
- **Outils :** La pharmacogénomique repose entièrement sur l'analyse de données massives, nécessitant la **Bio-informatique** et la **Protéomique**.

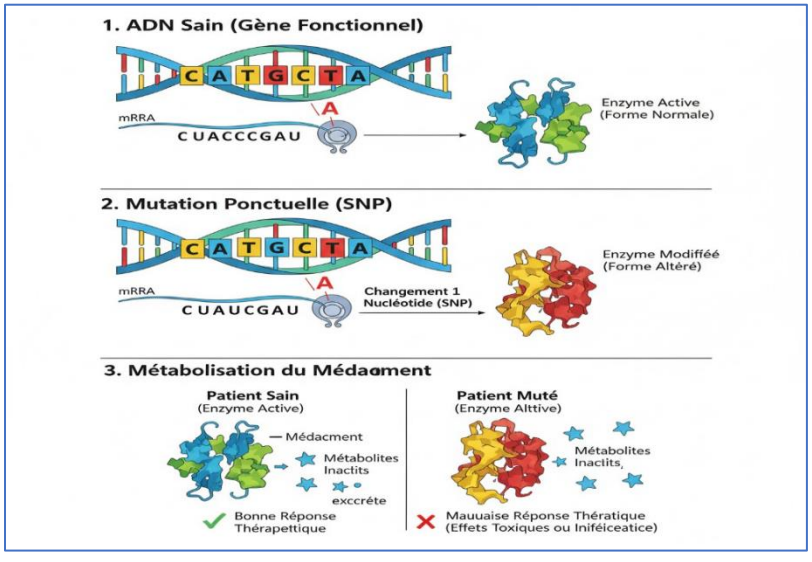


Figure 4.3 : La Pharmacogénomique et le SNP

4.3.2. Toxicogénomique

C'est l'étude des effets des agents toxiques (chimiques, environnementaux) sur l'expression des gènes et leur génome. Elle permet d'identifier des biomarqueurs précoces d'exposition ou de sensibilité aux toxiques.

📝 Exercices d'Application :

Exercice 1 : L'Interprétation du SNP en Pharmacogénomique

L'enzyme Cytochrome P450 2D6 (CYP2D6) métabolise un médicament essentiel.
Un gène sauvage (normal) permet une activité enzymatique élevée (métaboliseur rapide). La présence d'un SNP (variation nucléotidique) sur ce gène conduit à une activité enzymatique réduite (métaboliseur lent).

1. Si un patient est un **métaboliseur rapide** pour ce médicament, et que le médecin lui prescrit la dose standard, quelle est la conséquence probable sur le taux de médicament actif dans son sang ?
2. Si un patient est un **métaboliseur lent** et reçoit la dose standard, quel est le risque, et comment la Pharmacogénomique permet-elle de prévenir ce risque ?

Solution

1. **Métaboliseur Rapide :** L'enzyme est très efficace et dégrade rapidement le médicament. Conséquence : Le taux de médicament actif (thérapeutique) dans le sang risque d'être **trop faible**. Le patient pourrait ne pas bénéficier de l'effet thérapeutique désiré.
2. **Métaboliseur Lent :** L'enzyme met beaucoup de temps à dégrader le médicament. Conséquence : Le médicament s'accumule. Le risque est le **surdosage** et l'apparition d'effets secondaires toxiques.
    - **Prévention par Pharmacogénomique :** En identifiant le SNP (grâce à un test génétique), le médecin peut **ajuster la dose (la réduire)** dès la première prescription, évitant ainsi le risque toxique et personnalisant le traitement.

Exercice 2 :

L'analyse protéomique par Électrophorèse Bidimensionnelle (2D-E) d'un tissu montre deux isoformes d'une protéine (P1 et P2).

Protéine	Point Isoélectrique (pI)	Masse Moléculaire (MW)
P1	7.5	45 kDa
P2	6.8	45 kDa

Questions :

1. Quelle est la différence moléculaire entre P1 et P2 qui permet leur séparation lors de l'isoélectrofocalisation (première dimension de la 2D-E) ?
2. Un SNP est identifié dans le gène codant l'enzyme CYP2C9 chez un chien. Si ce SNP rend l'enzyme **moins active**, expliquez l'impact possible d'une dose standard de médicament métabolisé par cette enzyme.

Solution

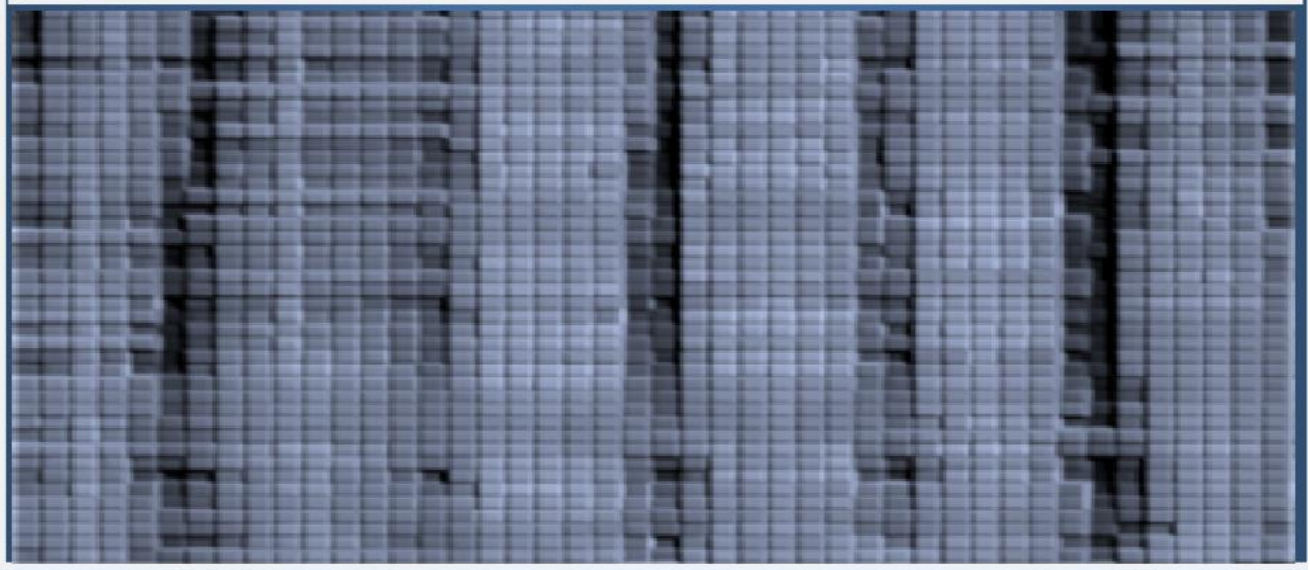
1. **Différence moléculaire :**

La séparation selon le Point Isoélectrique (pI) est basée sur la charge nette de la protéine. La différence de pI entre P1 (7.5) et P2 (6.8) suggère une modification post-traductionnelle (comme une phosphorylation qui ajoute une charge négative) ou un changement d'acide aminé dû à une mutation, modifiant ainsi le profil d'ionisation de la protéine.

2. **Impact pharmacogénomique :**

Si l'enzyme CYP2C9 est moins active, elle dégrade ou transforme plus lentement le médicament. L'impact probable est une accumulation du médicament sous sa forme active dans l'organisme de l'animal. Cela augmente le risque de surdosage et l'apparition d'effets secondaires toxiques ou indésirables.

Partie III
MÉTHODES D'ANALYSE ET OUTILS DE LA BIOLOGIE MOLÉCULAIRE



💻 PARTIE III : MÉTHODES D'ANALYSE ET OUTILS DE LA BIOLOGIE MOLÉCULAIRE

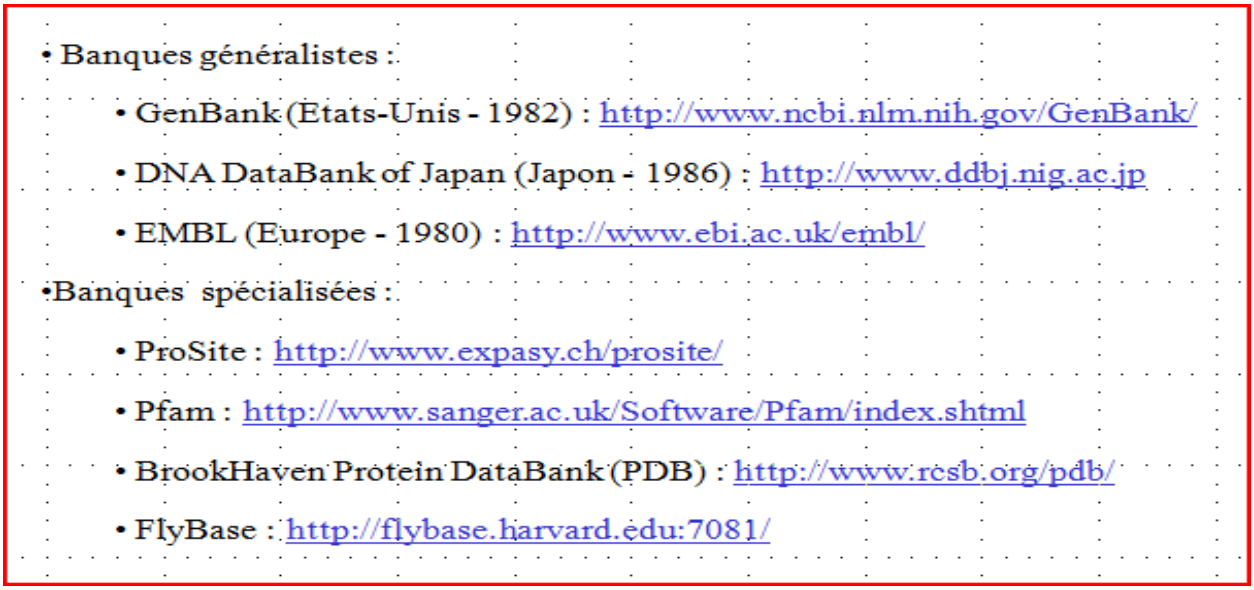
Chapitre 5 : Outils Bio-informatiques pour l'Analyse des Séquences

La **Bio-informatique** est l'approche « **in silico** » de la biologie traditionnelle. Elle consiste à utiliser l'outil informatique pour traiter, stocker, gérer et interpréter les quantités colossales de données générées par les disciplines de la Biologie Moléculaire et les « Omics » (génomique, protéomique, etc.).

5.1. Gestion et Stockage des Données Biologiques

Le rôle initial de la bio-informatique a été d'organiser et de rendre accessibles les séquences biologiques au niveau mondial.

- **Banques de Séquences de Nucléotides :** Les trois grandes bases de données primaires qui collaborent au niveau international (**International Nucleotide Sequence Database Collaboration**) sont **Genbank** (NCBI - USA), **EMBL** (Europe) et **DDBJ** (Japon). Elles collectent toutes les séquences d'ADN et d'ARN soumises par les chercheurs.
- **Banques de Séquences de Protéines : SwissProt** et **TrEMBL** (Europe) sont les références, proposant des séquences de protéines hautement annotées (SwissProt) ou annotées automatiquement (TrEMBL).



5.2. Alignement de Séquences : Détection de l'Homologie

L'analyse comparative des séquences est essentielle pour déduire des fonctions biologiques et des liens évolutifs. L'**alignement de séquences** consiste à faire correspondre des nucléotides ou des acides aminés entre deux ou plusieurs séquences afin d'identifier des régions de similarité.

5.2.1. Les Approches d'Alignement

- **Alignement Global (Méthode Needleman & Wunsch) :** Vise à aligner l'intégralité des deux séquences sur toute leur longueur.
- **Alignement Local (Méthode Smith & Waterman) :** Vise à identifier uniquement les **régions de forte similarité** (sous-séquences) à l'intérieur de séquences plus longues. Ces régions sont souvent les plus pertinentes fonctionnellement.

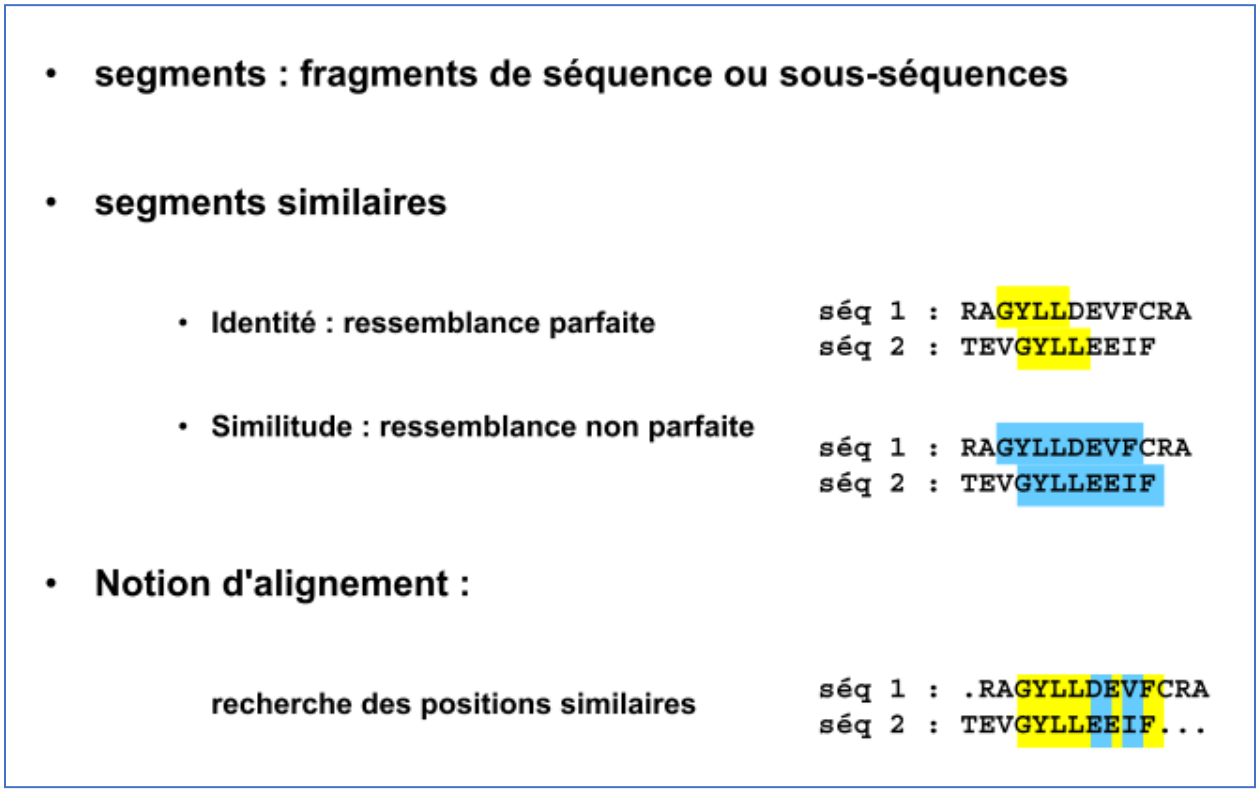


Figure 4.4 : Notion d'Alignment des sequences

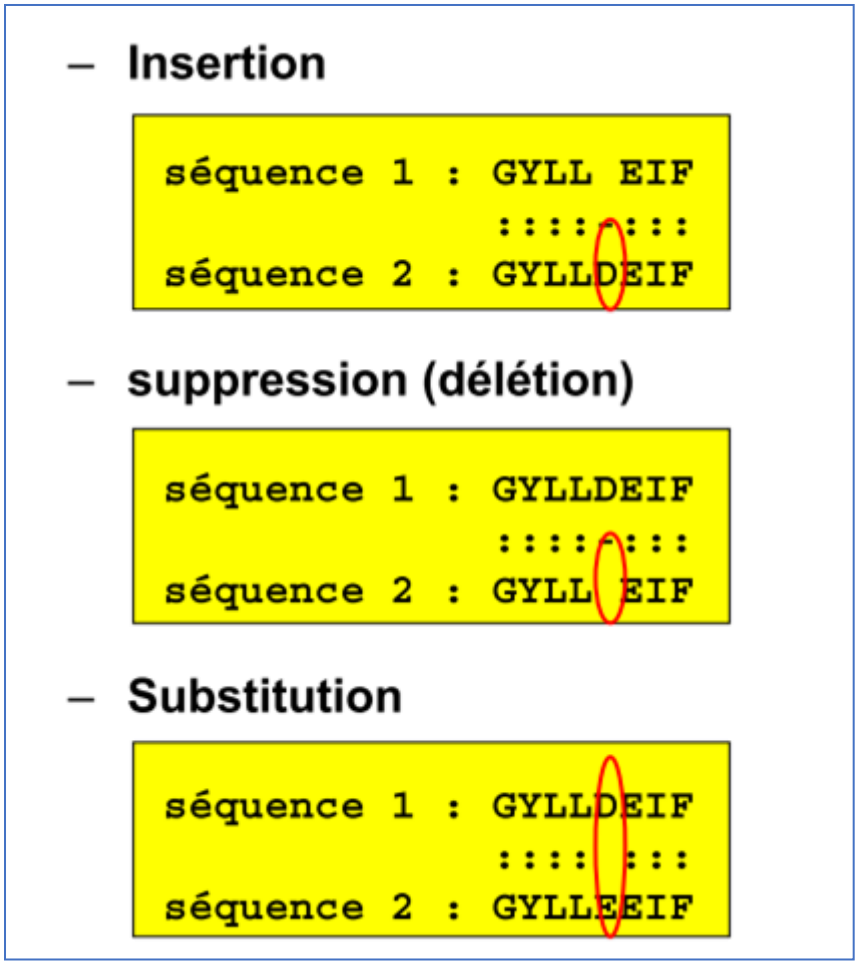


Figure 4.5 : Type d'Alignment des sequences

5.2.2. L'Outil Heuristique BLAST

La méthode d'alignement local (Smith & Waterman) est trop lente pour cribler des banques de données entières, car le temps de calcul croît avec le produit des longueurs des séquences. Pour pallier ce problème, des **méthodes heuristiques** ont été développées.

- **BLAST (Basic Local Alignment Search Tool) :** C'est le programme le plus populaire. Il utilise les principes de l'alignement local de Smith & Waterman, mais avec un algorithme optimisé pour la vitesse.
- **Fonction :** BLAST permet de prendre une séquence requête (nucléique ou protéique) et de rechercher très rapidement les séquences similaires dans une base de données, au prix d'une garantie non absolue de l'optimalité de l'alignement.

5.2.3. Homologie et Alignement Multiple

- **Homologie :** Deux gènes ou protéines sont dits **homologues** s'ils dérivent d'un **ancêtre commun**.
    - **Orthologie :** L'homologie résulte d'un événement de **spéciation** (séparation des espèces). Les gènes orthologues conservent généralement la même fonction.
    - **Paralogie :** L'homologie résulte d'une **duplication** du gène au sein du même génome. Les gènes paralogues peuvent avoir acquis de nouvelles fonctions.
- **Alignement Multiple (Exemple : Clustal W) :** Permet d'aligner **plus de deux séquences** simultanément. Ce type d'alignement est indispensable pour identifier les régions conservées (importantes pour la fonction) et pour construire les **arbres phylogénétiques** (dendrogrammes).

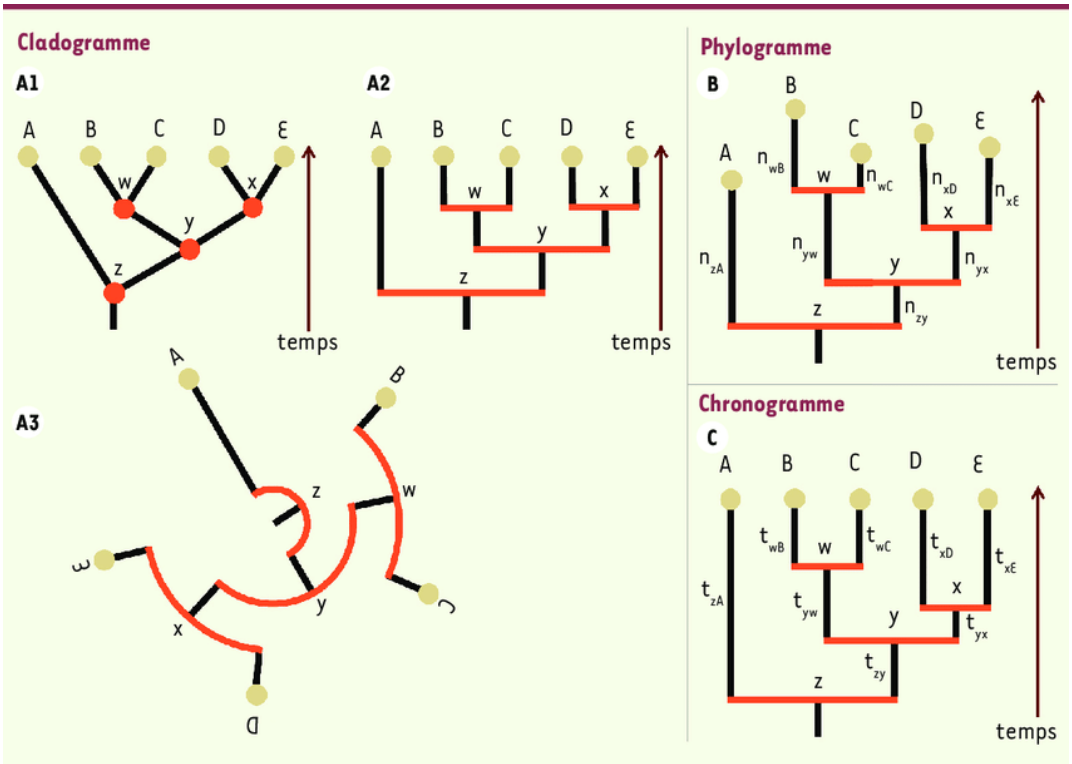


Figure 5 : Types d'arbre phylogenetique

Les arbres phylogénétiques, ou **dendrogrammes**, sont des outils graphiques qui représentent les **relations évolutives** (généalogiques) entre des séquences, des gènes ou des espèces (appelés *taxons*). Ils sont construits à partir de l'alignement multiple des séquences. La morphologie et l'interprétation d'un arbre dépendent de la manière dont les branches sont représentées et interprétées.

I. Classification par Représentation (Format)

Les arbres peuvent être dessinés selon différents formats graphiques, mais les plus courants sont :

Type de Représentation	Description	Forme Caractéristique
Arbre Cladogramme (Rectangulaire)	Le format le plus courant. Toutes les pointes terminales sont alignées verticalement ou horizontalement. L'interprétation est focalisée sur les **relations de branchement** (la topologie), et non sur la longueur des branches.	Branches en angles droits, pointes alignées.
Arbre Phylogramme	Similaire au cladogramme, mais les **longueurs des branches** ont une signification : elles sont proportionnelles à la quantité de changement évolutif (le nombre de mutations) ou au temps écoulé (si l'horloge moléculaire est utilisée).	Branches de longueurs variables, reflétant la distance génétique.
Arbre Radial (Circulaire)	Les taxons sont disposés autour d'un cercle, le nœud racine se trouve au centre. Ce format est souvent utilisé pour visualiser un **grand nombre de taxons** de manière compacte.	Disposition en cercle.

II. Classification par Racinement (Interprétation de la Racine)

Le **racinement** d'un arbre détermine la direction du temps évolutif et identifie l'ancêtre commun le plus récent de tous les taxons inclus.

1. Arbre Raciné (Rooted Tree)

- **Définition :** Possède un **nœud racine** clairement identifié, qui représente l'**ancêtre commun** le plus récent de tous les taxons représentés.
- **Interprétation :** L'arbre indique la **direction de l'évolution** et permet de distinguer les taxons ancestraux des taxons dérivés.
- **Méthode de Racinement :** Le plus souvent, le racinement est effectué en utilisant un groupe externe (**Outgroup**) : un taxon dont on sait, par des preuves indépendantes, qu'il est moins apparenté à tous les autres taxons étudiés.

2. Arbre Non Raciné (Unrooted Tree)

- **Définition :** Ne spécifie pas l'ancêtre commun. Il représente uniquement les **relations de parenté** (la topologie) sans indiquer la direction du temps.
- **Interprétation :** Il montre la manière dont les taxons sont connectés entre eux, mais ne permet pas de dire lequel a divergé en premier.
- **Avantage :** Tous les arbres construits par la plupart des méthodes d'inférence (comme les méthodes de maximum de parcimonie ou de distance) sont initialement non racinés.

Caractéristique	Arbre Raciné	Arbre Non Raciné
Direction	Le temps évolutif est spécifié (du passé à la branche terminale).	La direction du temps évolutif est inconnue.
Point de départ	Possède un **nœud racine** (ancêtre commun).	Ne possède pas de nœud racine.
Utilité	Utilisé pour la datation évolutive et l'analyse de la transmission (*Phylodynamique*).	Utilisé pour évaluer toutes les topologies possibles.

III. Classification par Topologie (Groupement)

Les relations de groupement entre les branches sont définies par la **topologie** de l'arbre.

1. Arbre Monophylétique

Un groupe est dit **monophylétique** s'il comprend un **ancêtre commun et tous ses descendants**. C'est le seul groupement jugé valide en systématique moderne, car il reflète une lignée évolutive naturelle.

2. Arbre Paraphylétique

Un groupe est **paraphylétique** s'il comprend l'ancêtre commun, **mais pas tous ses descendants** (par exemple, les Reptiles sans inclure les Oiseaux).

3. Arbre Polyphylétique

Un groupe est **polyphylétique** s'il regroupe des organismes dont l'ancêtre commun n'est **pas inclus** dans le groupe ou si l'ancêtre commun est lui-même placé dans un groupe différent.

En conclusion, la lecture d'un arbre phylogénétique est une compétence clé en bio-informatique. Il est essentiel de distinguer un phylogramme (où la longueur de la branche compte) d'un cladogramme (où seule la topologie compte), et de s'assurer si l'arbre est bien enraciné pour pouvoir inférer la direction et le temps de l'évolution.

📝 Exercices d'Application :

Exercice 1 : Interprétation des Résultats BLAST

Vous utilisez BLAST pour comparer une séquence d'ADN d'un gène animal contre une base de données. Voici trois résultats d'alignement (Hits) :

Hit	Identité (%)	Score E-Value
Hit A	99%	1 * times 10-^{180}
Hit B	65%	0.05
Hit C	50%	10.0

Questions :

1. Quel Hit représente la preuve la plus forte d'une homologie (relation évolutive par ancêtre commun) ? Justifiez en utilisant le terme E-Value.

2. Le résultat BLAST est basé sur quel type d'alignement (Global ou Local) et quel algorithme est sa base théorique ?

Solution

1. **Preuve d'homologie la plus forte :**

Le Hit A ($E=1 * 10^{-180}$) représente la preuve la plus forte. L'E-Value est le nombre d'alignements avec un score aussi bon ou meilleur que l'on s'attendrait à trouver par pur hasard. Plus l'E-Value est proche de zéro, plus la similarité est significative. 10^{-180} est extrêmement faible.

2. **Type d'alignement et algorithme :**
    - BLAST est une méthode heuristique qui utilise les principes de l'Alignement Local.
    - Sa base théorique est l'algorithme de Smith et Waterman.

Exercice 2 : Interprétation de la Topologie et du Racinement.

Observez l'arbre phylogénétique ci-dessous représentant les relations évolutives entre quatre taxons (A, B, C, D) et un groupe externe (Outgroup - O).

l'ordre de divergence est : O →D →C → (A, B).

Questions :

1. L'arbre présenté est-il raciné ou non raciné ? Justifiez votre réponse.
2. Identifier l'ancêtre commun le plus récent des taxons A et B.
3. Quel taxon est le plus éloigné génétiquement de l'ensemble (A, B, C, D) ?
4. Le groupe formé par les taxons (A, B) et leur ancêtre commun est-il monophylétique, paraphylétique ou polyphylétique ? Justifiez.
5. Si les longueurs des branches sont significatives (Phylogramme), quelle conclusion pouvez-vous tirer sur l'évolution du taxon D par rapport au taxon C ?

Solution

1. Racinement : L'arbre est raciné. La racine est identifiée par l'utilisation du taxon O (Outgroup). L'Outgroup, par définition, est le premier à avoir divergé de l'ensemble des taxons étudiés (le groupe interne ou *ingroup* : A, B, C, D), indiquant la direction temporelle de l'évolution.
2. Ancêtre Commun (A et B) : L'ancêtre commun le plus récent est le nœud immédiatement adjacent et reliant les branches menant à A et B. Il représente le dernier point où les lignées de A et B étaient communes avant de diverger.
3. Taxon le plus éloigné : Le taxon le plus éloigné (ou le premier à diverger) de l'ensemble (A, B, C, D) est le taxon O (Outgroup).
4. Monophylétisme : Le groupe (A, B) est un groupe monophylétique. Un groupe monophylétique inclut un ancêtre commun (leur nœud de jonction) et tous ses descendants. Ici, A et B sont tous les descendants de ce nœud.
5. Conclusion sur l'Évolution (Phylogramme) : Si la longueur des branches est proportionnelle au changement évolutif (nombre de mutations), on peut constater que le taxon D a subi plus de changements (plus de mutations) depuis sa divergence de l'ancêtre commun de (C, D) que le taxon C (branche plus longue pour D que pour C).

Exercice 3 : Distance Génétique et Arbre Non Raciné

Vous avez comparé les séquences d'un gène entre quatre espèces de bactéries (E1, E2, E3, E4). L'alignement multiple a permis de calculer les distances génétiques (nombre de mutations) entre elles, résumées dans le tableau suivant :

Espèce	E1	E2	E3	E4
E1	0			
E2	20	0		
E3	70	70	0	
E4	80	80	10	0

Questions :

1. Quelles sont les deux espèces les plus proches génétiquement ?
2. Quel est le couple d'espèces qui a divergé le plus tôt (qui est le plus éloigné) ?
3. Décrivez la topologie de l'arbre non raciné qui représente au mieux ces distances, en indiquant les deux premières paires qui se rejoignent.
4. Si vous deviez enraciner cet arbre en utilisant la méthode de l'horloge moléculaire, que représente la longueur totale des branches allant de la racine à chaque taxon ?

Solution

1. Espèces les plus proches : Les deux espèces les plus proches sont celles avec la plus petite distance génétique : E3 et E4 (distance de 10).
2. Couple le plus éloigné : Les paires les plus éloignées sont celles qui ont la plus grande distance génétique : (E1, E3) et (E2, E3) et (E1, E4) et (E2, E4) (distance maximale de 80 ou 70), indiquant une divergence très ancienne.
3. Topologie de l'Arbre (Non Raciné) :
    - La première paire à se joindre est celle avec la distance la plus courte : (E3, E4).
    - Ensuite, on cherche le couple le plus proche restant. E1 et E2 ont une distance de 20.
    - L'arbre non raciné montrera une branche joignant E3 à E4, et une autre branche joignant E1 à E2. Ces deux groupes se joignent ensuite à un nœud ancestral commun.

4. Horloge Moléculaire : Si l'on enracine l'arbre avec l'horloge moléculaire (taux de mutation constant), la longueur totale des branches allant de la racine à n'importe quel taxon terminal doit être identique. Cette longueur totale représente le temps évolutif (le temps écoulé) depuis l'ancêtre commun jusqu'aux espèces actuelles, sous l'hypothèse que l'évolution se fait à vitesse constante.

Exercice 4 : Distinguer Cladogramme et Phylogramme

Vous avez deux représentations graphiques (Figure A et Figure B) des relations entre cinq taxons (T1 à T5). La Figure A est un arbre aux branches de longueurs différentes, tandis que la Figure B est un arbre aux pointes alignées et aux branches angulaires.

Questions :

1. Identifier quel arbre est un Cladogramme et quel arbre est un Phylogramme.
2. Quel arbre est le plus utile si l'objectif principal est de déduire la chronologie ou le moment des événements de spéciation ? Justifiez.
3. Les deux arbres (A et B) peuvent-ils avoir la même topologie ? Expliquez.
4. Dans la Figure A (Phylogramme), si la branche menant à T5 est très courte, qu'est-ce que cela signifie en termes de distance génétique ?

Solution

1. Identification :
    - Figure A : Phylogramme (Longueurs des branches variables et significatives).
    - Figure B : Cladogramme (Pointes alignées, seule la relation de branchement importe).
2. Utilité pour la Chronologie : Le Phylogramme est le plus utile. Un phylogramme permet, grâce aux longueurs de ses branches, de représenter la quantité de changement évolutif (ou le temps, s'il est daté par l'horloge moléculaire) entre les taxons. Le Cladogramme ne fournit aucune information sur le temps ou la distance.
3. Même Topologie : Oui, ils peuvent avoir la même topologie. La topologie fait référence à la structure de branchement (qui est plus proche de qui). Un Cladogramme est souvent la

version "simplifiée" d'un Phylogramme, où les informations de distance sont ignorées, mais les relations de parenté restent les mêmes.

4. Interprétation de la Branche Courte (T5) : Une branche courte sur un Phylogramme signifie que le taxon T5 a subi très peu de changements génétiques (très peu de mutations) depuis sa dernière divergence de son ancêtre commun par rapport aux autres taxons. Il est génétiquement très proche de son ancêtre immédiat.

5.3. Algorithmique d'Alignement de Séquences : La Programmation Dynamique

L'alignement de deux séquences repose sur l'identification d'une **similarité maximale** en attribuant des scores (positifs pour les correspondances, négatifs pour les mésappariements et les *gaps*). L'approche de la **Programmation Dynamique** garantit de trouver l'alignement **optimal** en construisant et en remplissant une matrice bidimensionnelle.

5.3.1. Alignement Global : Needleman et Wunsch (1970)

L'algorithme de **Needleman et Wunsch** est le fondement de l'alignement global.

- **Principe :** Il force l'alignement des deux séquences (Séquence A et Séquence B) sur toute leur longueur, du début à la fin.
- **Méthode :** Le score de l'alignement optimal est calculé en parcourant la matrice de similarité. Chaque cellule de la matrice représente le meilleur score possible pour aligner un préfixe de la séquence A avec un préfixe de la séquence B. Le meilleur chemin est déterminé par la **rétro-trace** (ou *traceback*) à partir de la dernière cellule de la matrice (en bas à droite).
- **Application :** Essentiel pour la comparaison de gènes orthologues présumés, c'est-à-dire qui ont conservé une forte similarité sur l'ensemble de leur séquence.

5.3.2. Alignement Local : Smith et Waterman (1981)

L'algorithme de **Smith et Waterman** s'est imposé pour l'alignement local, c'est-à-dire la recherche des régions de **plus forte conservation** (sous-séquences).

- **Distinction majeure :** La matrice est initialisée avec des zéros. Les scores négatifs sont également ramenés à zéro. Cela permet à l'algorithme de **recommencer** l'alignement si le score devient trop bas, ignorant les régions non similaires.
- **Résultat :** Le meilleur alignement correspond à la **valeur maximale** trouvée n'importe où dans la matrice (pas nécessairement la dernière cellule). La rétro-trace s'arrête dès qu'elle rencontre un zéro.
- **Application :** Indispensable pour l'identification des domaines fonctionnels, des motifs protéiques ou pour comparer des séquences génomiques qui ne sont conservées que par endroits.

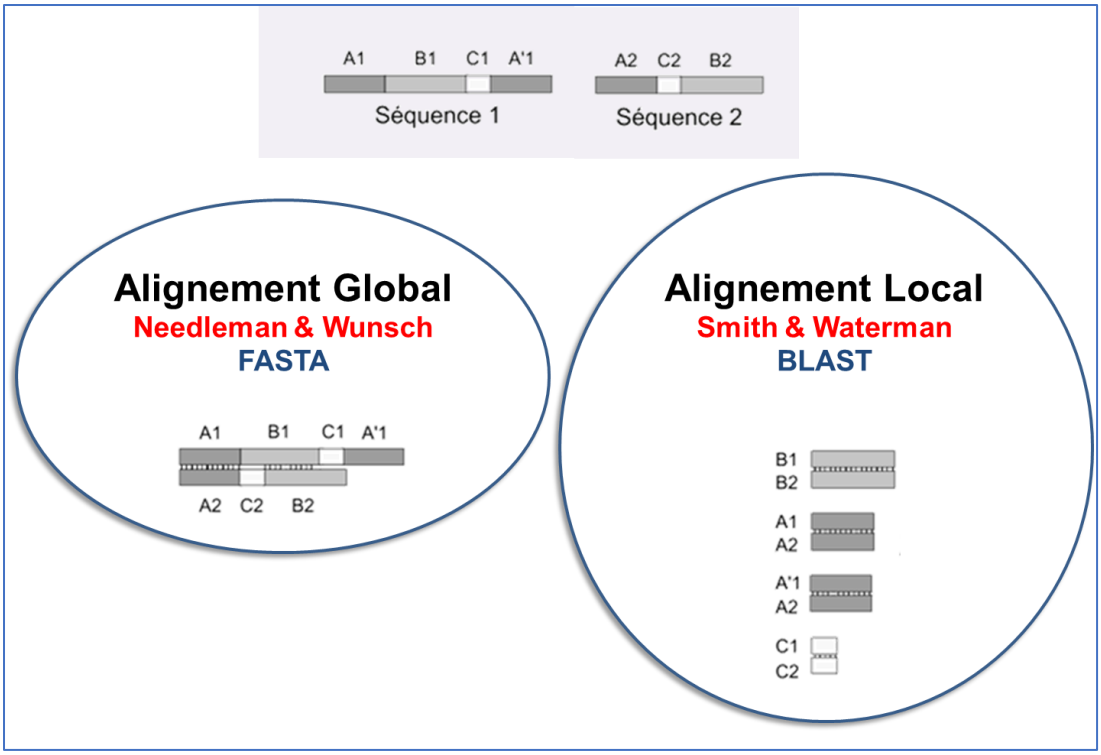


Figure 4.5 : Comparaison entre l'alignment global et local des sequences

5.4. Heuristique de Criblage : La Nécessité de BLAST

La **Programmation Dynamique** est extrêmement **coûteuse en temps de calcul** (complexité en L1 * L2.

Comparer une séquence d'ADN de 1000 bases à une banque de 10 millions de séquences de 1000 bases prendrait un temps inacceptable.

Pour cette raison, des **méthodes heuristiques** ont été développées. Ces méthodes sont des raccourcis efficaces et rapides, mais qui ne garantissent plus l'alignement optimal.

5.4.1. Le Mécanisme de BLAST (Basic Local Alignment Search Tool)

BLAST utilise l'heuristique pour effectuer des alignements locaux ultra-rapides.

1. **Recherche des "Mots" (*Words*) :** BLAST fragmente la séquence requête en petits mots (généralement 3 acides aminés ou 11 nucléotides).
2. **Filtrage :** Il identifie très rapidement (grâce à des tables de hachage) les séquences de la base de données contenant des mots **similaires** (selon une matrice de score et un seuil T) aux mots de la requête.

3. **Extension de l'Alignement :** Seules les correspondances de mots prometteuses sont étendues localement, sans remplir toute la matrice. Si le score de l'extension chute trop bas, le processus s'arrête.

Ce processus permet un gain de temps considérable, permettant le criblage de banques entières en quelques secondes ou minutes.

5.4.2. Évaluation Statistique : Le Score E-Value

Les résultats des alignements BLAST sont accompagnés d'une métrique statistique essentielle pour juger de la pertinence biologique : le **Score E-Value (Expectation Value)**.

- **Définition :** L'E-Value est le nombre d'alignements avec un score aussi bon ou meilleur que celui obtenu qui seraient **attendus d'être trouvés par pur hasard** dans la base de données consultée.
- **Interprétation :** Une petite E-Value (par exemple, 10^{-5}) indique une probabilité très faible que la similarité soit due au hasard. C'est une forte indication d'une **homologie** (relation par ancêtre commun).

📝 Exercices d'Application :

Exercice 1 : Codage et Conséquences des Mutations (Rappel)

Une séquence d'ADN codante (brin transcrit) est la suivante :
3' - TAC GTT ATC GCG ACT - 5'

Table du Code Génétique (Rappel) :

Codon ARNm	Acide Aminé	Codon ARNm	Acide Aminé
AUG	Méthionine (Start)	UAA, UAG, UGA	Stop
GCA, GCC, GCG, GCU	Alanine	UUU, UUC	Phénylalanine
UAU, UAC	Tyrosine	GUG	Valine

Questions :

1. Déterminez la séquence de l'ARNm mature et la séquence d'acides aminés (protéine) correspondante.

2. Si le 7ème nucléotide de l'ADN (A) est muté en G (**3' - TAC GTT GT C GCG ACT - 5'**), quel est le type de mutation résultant et sa conséquence sur la protéine ?
3. Si le 10ème nucléotide (G) est délété (**3' - TAC GTT ATC C G ACT - 5'**), quelle est la conséquence principale ?

Solution

1. **Transcription (ADN $\rightarrow$ ARNm) :** Le brin d'ARNm est complémentaire et antiparallèle au brin transcrit (en remplaçant T par U) :
   - ✓ **ARNm :**
     5' - AUG CAA UAG CGC UGA - 3'
     Traduction (ARNm → Protéine) :
   - ✓ **Protéine :** Met (Start) → Glutamin (STOP)
   - ✓ *Note : La séquence protéique s'arrête prématurément au deuxième codon.*
2. **Mutation Ponctuelle (Substitution A → G) :**
   - ✓ Ancien codon de l'ADN : ATC →Codon ARNm : **UAG** (STOP)
   - ✓ Nouveau codon de l'ADN : GTC → Codon ARNm : **CAG** (Glutamine)
   - ✓ **Type de mutation : Non-sens inverse** (car le codon-stop initial est remplacé par un codon codant un acide aminé).
   - ✓ **Conséquence :** La séquence d'ARNm devient
     5' → AUG CAA CAG CGC UGA → 3'
     Nouvelle protéine : Met (Start)} → Glutamine→ Glutamine→ Arginine (STOP)
     La protéine est plus longue (un acide aminé de plus) et sa fonction pourrait être rétablie ou modifiée.
3. **Mutation par Délétion (du G) :**
   - ✓ Séquence ADN après délétion :
     3' → TAC GTT ATC C G ACT → 5' (le 10ème nucléotide est C)
   - ✓ Lecture de l'ARNm (par codons):
     5' → AUG CAA UAC G C U GA → 3'
     Conséquence : La délétion d'un nucléotide (non-multiple de 3) provoque un **décalage du cadre de lecture (*frameshift*)** à partir de ce point, altérant tous les codons en aval.

Exercice 2 : Algorithmes et Vitesse de Calcul

Vous devez comparer une nouvelle séquence (Séquence Q) d'une protéine de 500 acides aminés à deux banques de données :

- **Banque A :** 10 séquences, chacune de 1000 acides aminés.
- **Banque B :** 10 millions de séquences, chacune de 100 acides aminés.

Questions :

1. Quel algorithme (Needleman-Wunsch ou BLAST) est le plus approprié pour comparer Séquence Q aux 10 séquences de la Banque A ? Justifiez.
2. Quel algorithme est obligatoirement requis pour le criblage de la Banque B ? Justifiez en termes de complexité algorithmique.
3. Si BLAST vous donne une E-Value de $\mathbf{0.01}$ pour un alignement, quelle est l'interprétation biologique et la décision à prendre ?

Solution 2

1. **Banque A (petite taille) :**
   - ✓ **Algorithme approprié : Needleman et Wunsch (Alignement Global)** ou **Smith et Waterman (Alignement Local)**.
   - ✓ **Justification :** Le nombre de comparaisons est très faible (seulement 10). L'approche par Programmation Dynamique est possible et souhaitable car elle garantit l'alignement **optimal** de chaque comparaison.
2. **Banque B (grande taille) :**
   - ✓ **Algorithme requis : BLAST (méthode heuristique)**.
   - ✓ **Justification :** La Programmation Dynamique a une complexité en L1 * L2 \ times N (où N est la taille de la banque). Avec 10^7 comparaisons, le temps de calcul est intolérable. BLAST est nécessaire pour sa rapidité, car il évite de remplir la matrice pour les séquences non pertinentes.
3. **Interprétation de l'E-Value :**
   - ✓ Une E-Value de 0.01 signifie que l'on s'attend à trouver **0,01 alignement** de ce score (ou mieux) par pur hasard.
   - ✓ **Décision :** La valeur est très faible. C'est une indication forte que la similarité est **biologiquement significative** et que l'homologie entre la séquence Q et la séquence

trouvée est hautement probable. La séquence trouvée est un bon candidat pour être un orthologue ou un paralogue.

Chapitre 6 : Techniques d'Amplification et de Séquençage

6.1. La Réaction en Chaîne par Polymérase (PCR)

La **PCR** (Polymerase Chain Reaction ou Amplification en Chaîne par Polymérase) est une technique de **réplication ciblée *in vitro***. Inventée par K. Mullis (Prix Nobel 1993), elle permet d'obtenir un **million de copies** ou plus d'un fragment d'ADN spécifique en quelques heures, à partir d'un échantillon peu abondant.

6.1.1. Composants et Rôle de la Taq Polymérase

- **Matériel nécessaire :** ADN matrice, deux **amorces** (oligonucléotides) spécifiques délimitant la cible, les quatre désoxyribonucléotides (dNTP) en large excès, et l'ADN polymérase.
- **Taq Polymérase :** L'automatisation de la PCR a été rendue possible par l'utilisation de l'ADN polymérase de la bactérie *Thermus aquaticus* (**Taq Polymérase**), qui est **thermorésistante**. Cette résistance à la chaleur lui permet de survivre aux cycles de haute température.
- **Appareillage :** La réaction se déroule dans un **thermocycleur**, un appareil programmable capable d'exposer les tubes à des températures précises et pour des durées déterminées.

6.1.2. Les Trois Étapes du Cycle PCR

Un cycle de PCR, répété généralement 25 à 40 fois, se déroule en trois étapes contrôlées par la température :

1. **Dénaturation (Environ 95°C) :** La chaleur rompt les liaisons hydrogène, entraînant la dissociation complète des deux brins d'ADN double-brin.
2. **Hybridation (*Annealing*, 50-60°C) :** La température est abaissée pour permettre aux deux amorces de s'apparient spécifiquement à la matrice d'ADN cible.
3. **Élongation (Polymérisation, Environ 72°C) :** Température optimale pour la Taq polymérase. Elle synthétise les nouveaux brins complémentaires, à partir des amorces, dans le sens **5' —> 3'**.

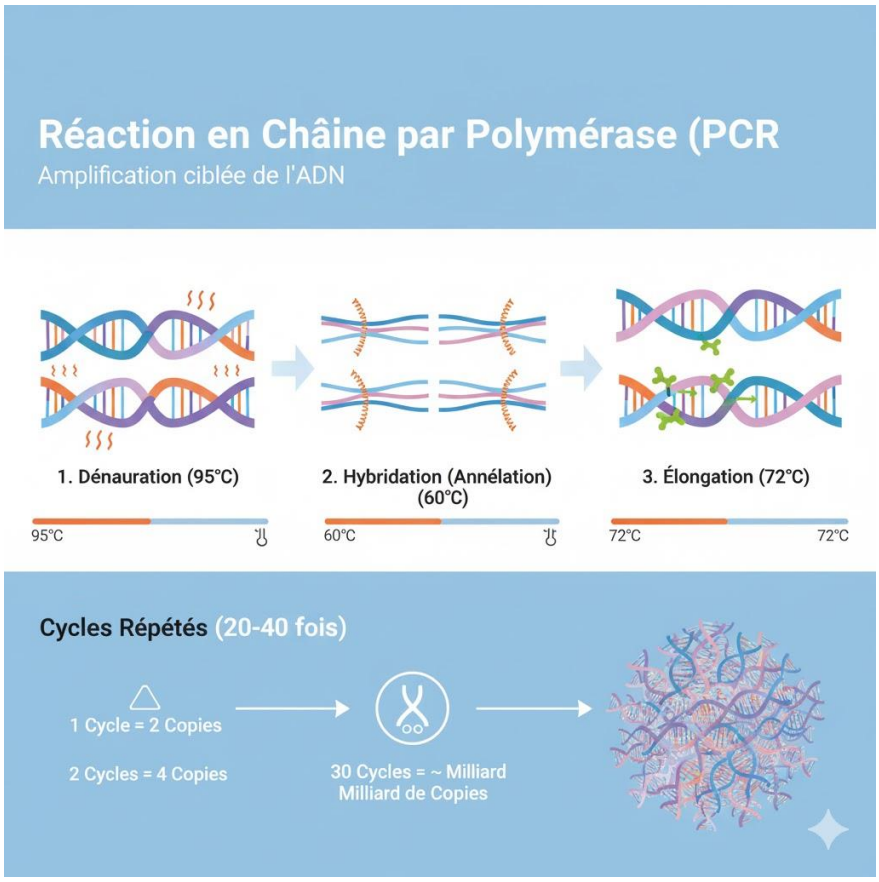


Figure 6 : Illustration du PCR

6.2. Le Séquençage de l'ADN

Le **séquençage de l'ADN** est la technique qui permet de déterminer l'ordre précis des nucléotides (A, C, G, T) composant la molécule d'ADN.

6.2.1. La Méthode de Sanger (Didésoxy)

- **Principe :** Basée sur la réplication *in vitro*, la méthode de Sanger (ou didésoxy) est la méthode classique de référence.
- **Arrêt de Chaîne :** La clé de cette méthode réside dans l'utilisation de **didésoxyribonucléotides (ddNTP)**. Contrairement aux dNTP normaux, les ddNTP n'ont pas de groupement 3'-OH ; lorsqu'ils sont incorporés dans un brin d'ADN en cours de synthèse, ils **provoquent l'arrêt immédiat de l'élongation de la chaîne**.

- **Séquençage :** En mélangeant dNTP et ddNTP marqués différemment, la réaction génère une série de fragments d'ADN de toutes les longueurs possibles.
  Leur séparation par électrophorèse capillaire et leur lecture par un laser permettent de reconstituer la séquence originale.

6.2.2. L'Automatisation et les Défis

- **Séquenceurs Automatiques :** Aujourd'hui, la majorité des séquences sont obtenues sur des séquenceurs automatiques, offrant un gain de temps, un coût moindre et une capacité de lecture de plusieurs centaines à mille nucléotides.
- **Défi d'Assemblage :** La complexité du génome humain réside dans le fait que les gènes ne sont pas inscrits en une seule pièce ; ils sont découpés en fragments éparpillés. L'analyse et l'assemblage de ces données (le « puzzle génétique ») nécessitent une expertise bio-informatique intensive.

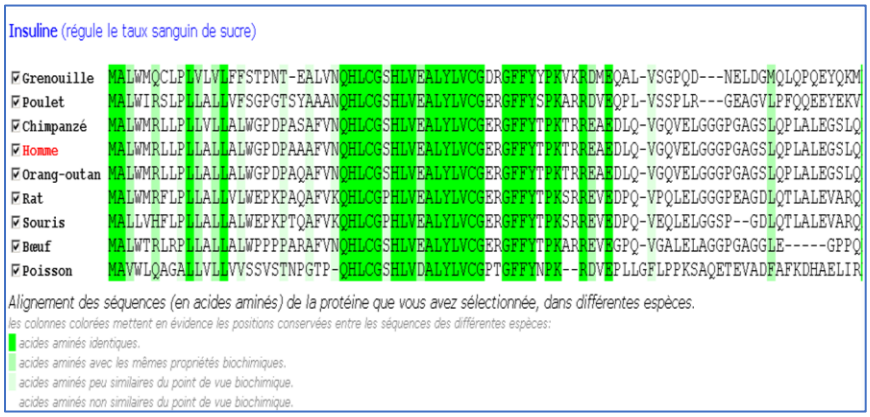


Figure 7 : Exemple d'Alignment de sequence d'insuline

📝 Exercices d'Application :

Exercice : Quantification PCR et Séquençage (ddNTP)

Vous réalisez une PCR pour amplifier un segment d'ADN cible à partir de 5 molécules initiales.

Questions :

1. Combien de molécules d'ADN cibles obtiendrez-vous après **4 cycles** de PCR complets ?
   Formule : $N_{final} = N_{initial} * 2^{cycles}$.

2. Dans le séquençage de l'ADN (méthode de Sanger), quel est le rôle précis des **didésoxyribonucléotides (ddNTP)** ?

Solution

1. **Nombre de molécules après 4 cycles :**
   - ✓ $N_{final} = 5 * 2^{4}$
   - ✓ $N_{final} = 5 * 16$
   - ✓ **Résultat :** Vous obtiendrez **80 molécules** d'ADN cibles après 4 cycles.
2. Rôle des ddNTP :

Les ddNTP sont des nucléotides modifiés qui ne possèdent pas de groupe hydroxyle (-OH) en position 3' de leur sucre.

Lorsqu'un ddNTP est incorporé dans une chaîne d'ADN en cours de synthèse, il bloque (termine) l'élongation de la chaîne, car la liaison phosphodiester avec le nucléotide suivant ne peut se former. Ils sont les terminateurs de la réaction de séquençage.

Chapitre 7 : Phylogénie, Phylogéographie et Phylodynamique (Lien Évolutif)

7.1. La Phylogénie : Reconstruire l'Histoire

La **Phylogénie** permet d'inférer et de représenter graphiquement les **liens évolutifs (généalogiques)** entre les espèces ou les gènes, en se basant sur l'accumulation de mutations au cours du temps. L'arbre phylogénétique est la représentation de ces relations.

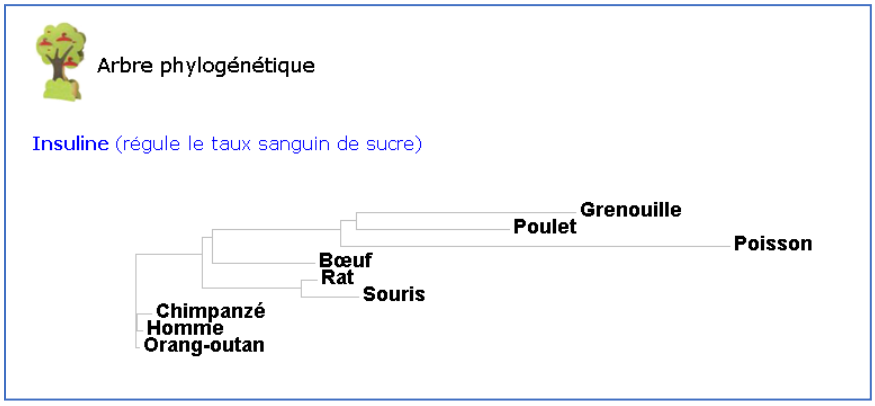


Figure 8 : Exemple d'arbre phylogenetique d'insuline

7.2. Approches Spatio-Temporelles

La **Phylogéographie** et la **Phylodynamique** sont des domaines d'application avancés de la phylogénie.

- **Phylogéographie :** Étude des approches **spatiales** de la distribution des organismes et des gènes, tenant compte des facteurs géographiques et des migrations.
- **Phylodynamique :** Se concentre sur la **dynamique de transmission** des agents pathogènes (selon Grenfell et al., 2004).

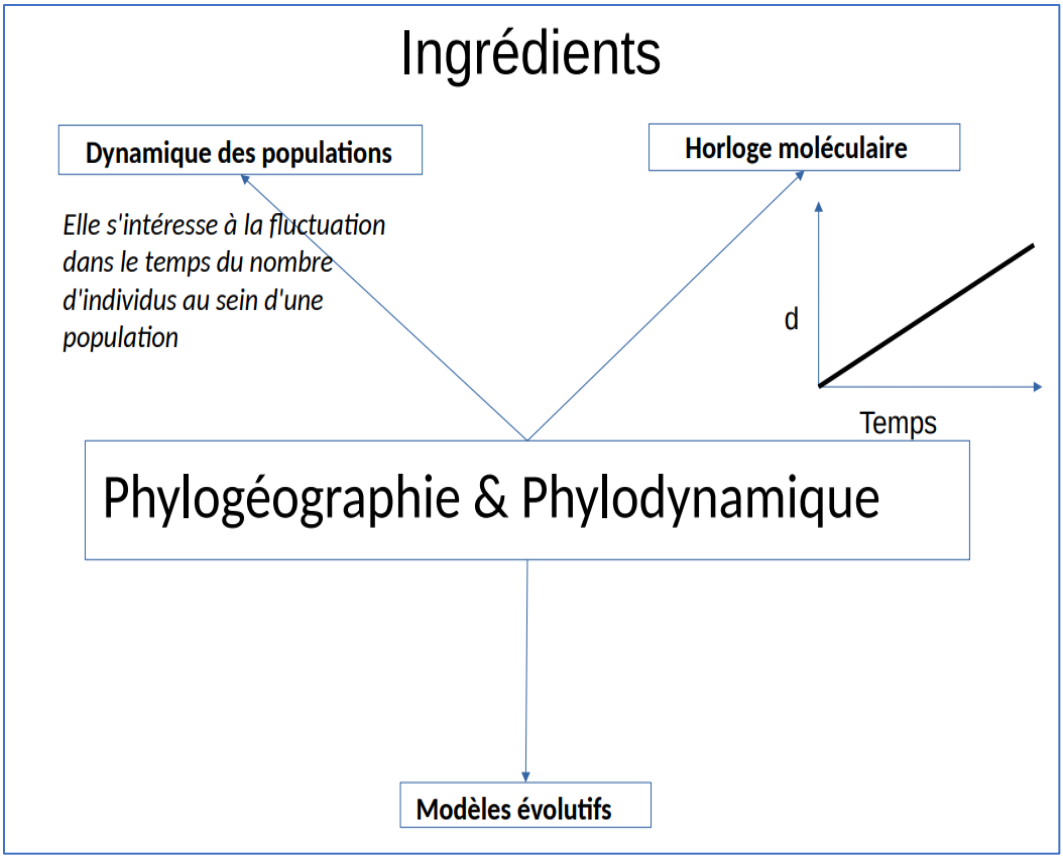


Figure 8 : Comparaison entre la phylogeographique et la phylodynamique

7.3. Les Ingrédients de la Modélisation Évolutive

Ces analyses nécessitent l'intégration de facteurs temporels et probabilistes.

1. **Horloge Moléculaire :** Permet d'intégrer le **Temps** en postulant un taux d'accumulation des mutations constant. Elle sert à **estimer les dates de divergence** entre les séquences.
2. **Modèles Évolutifs :** Modélisation mathématique du changement génétique qui permet, par exemple, d'inférer la séquence génétique d'un **ancêtre hypothétique**.
3. **Dynamique des Populations :** Prise en compte de la fluctuation du nombre d'individus dans le temps.

📝 Exercices d'Application :

Exercice 8.1 : Lecture d'Arbre Phylogénétique

Un arbre phylogénétique non raciné montre les relations entre quatre souches virales (V1, V2, V3, V4). V1 et V2 sont plus proches et divergent d'un ancêtre commun, qui lui-même est plus proche de V3. V4 est le premier à s'être séparé de ce groupe.

Questions :

1. Quel est l'ordre de divergence des souches, de l'ancêtre le plus ancien vers le plus récent ?
2. Si l'on considère le groupe (V1, V2, V3) et leur ancêtre commun immédiat, mais que l'on exclut V4, ce groupe est-il **monophylétique, paraphylétique ou polyphylétique** ? Justifiez.
3. Dans l'étude de l'épidémiologie des virus, quelle est la discipline qui combine phylogénie et dynamique de population ?

Solution

1. **Ordre de divergence :**

L'ordre de divergence, de la lignée la plus ancienne à la plus récente, est : V4 → V3 → (V1, V2).

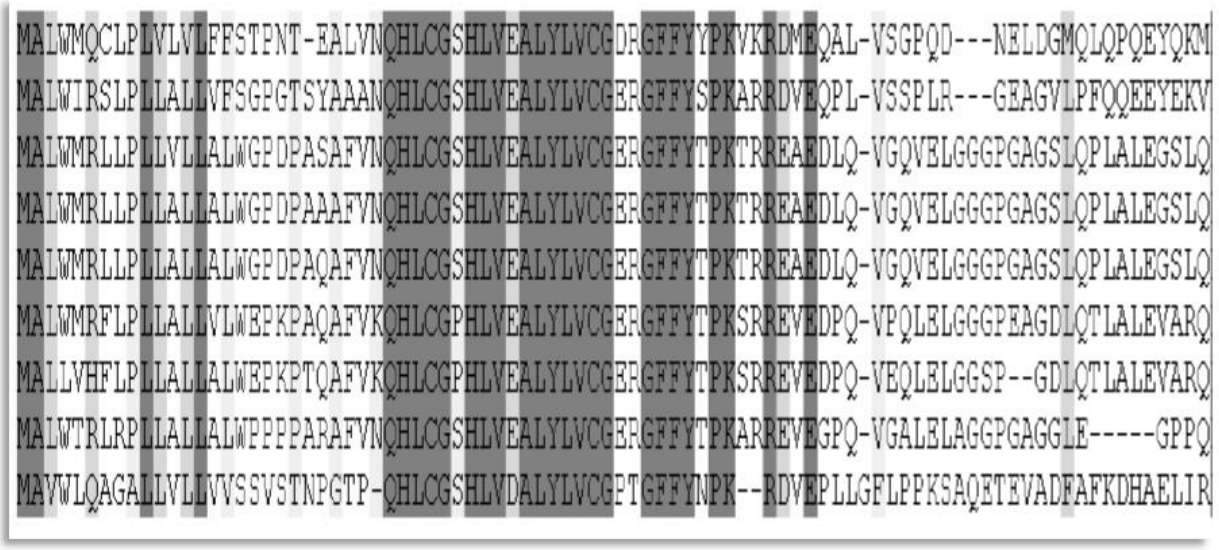
2. **Classification du groupe (V1, V2, V3) :**

Ce groupe est monophylétique. Un groupe monophylétique inclut un ancêtre commun et tous ses descendants. Si l'on considère l'ancêtre commun du groupe V1, V2 et V3, V4 est déjà sorti avant ce point. Le groupe (V1, V2, V3) respecte donc la règle d'inclure leur ancêtre commun et tous les descendants de cet ancêtre.

3. **Discipline :**

La discipline qui combine phylogénie (évolution) et dynamique de population (propagation) est la Phylodynamique.

PARTIE IV
INGÉNIERIE GÉNÉTIQUE ET ÉTHIQUE



⚖️ PARTIE IV : INGÉNIERIE GÉNÉTIQUE ET ÉTHIQUE

Chapitre 8 : Clonage et ses Enjeux

L'**Ingénierie Génétique** est l'ensemble des techniques permettant la manipulation du génome. Le **clonage** est une de ces techniques, désignant un mode de reproduction en laboratoire visant à créer un organisme, un gène ou une entité **génétiquement identique** (le clone) à l'original.

8.1. Définitions et Types de Clonage

- **Le Clone :** Le terme « clone » désigne tout objet ou organisme considéré comme génétiquement identique à un autre, possédant le même ADN.
- **Clonage Naturel :** Il existe naturellement chez les organismes non sexués (bactéries) ou chez les mammifères via la formation des **jumeaux monozygotes** (vrais jumeaux), issus de la séparation précoce de deux cellules filles génétiquement identiques après la fécondation.

8.1.1. Clonage de Gènes (*Clonage Moléculaire*)

C'est la production de copies multiples d'un gène ou d'un fragment d'ADN spécifique. Ce type de clonage utilise des **plasmides** (ADN circulaires bactériens) comme vecteurs et permet la production à grande échelle de protéines thérapeutiques (ex : insuline).

8.1.2. Clonage Cellulaire et Organismique

- **Clonage Reproductif :** Vise à créer un organisme entier, génétiquement identique au donneur de l'ADN. L'exemple le plus célèbre est celui de la brebis Dolly (1996), obtenue par la technique de **Transfert Nucléaire de Cellule Somatique (TNCS)**.
- **Clonage Thérapeutique :** Vise à produire des cellules souches embryonnaires génétiquement compatibles avec un patient, sans intention de donner naissance à un individu. L'objectif est la régénération de tissus ou d'organes (médecine régénératrice).

8.2. Objectifs, Problèmes et Enjeux Éthiques

Le clonage est un sujet de controverse qui fascine et effraie, car il représente à la fois un immense potentiel pour la médecine et des risques de dérives sociétales.

Catégorie	**Objectifs (Progrès Potentiels)**	**Problèmes et Enjeux (Dangers)**
Recherche & Biologie	Meilleure compréhension des maladies génétiquement transmissibles et des cancers.	Vieillissement prématuré, mortalité imprévue, anomalies génétiques et développementales importantes.
Santé & Médecine	**Construction des banques d'organes** (pour des greffes potentielles).	Risques de complications post-natales et échecs importants.
Environnement	Protéger les espèces en voie de disparition.	Augmentation exagérée des populations, problèmes écologiques et perte de diversité génétique.
Philosophie & Éthique	Nouveaux traitements pour des maladies incurables.	La **notion d'âme perd tout son sens**. Risques de dérives politiques, sociales et économiques.

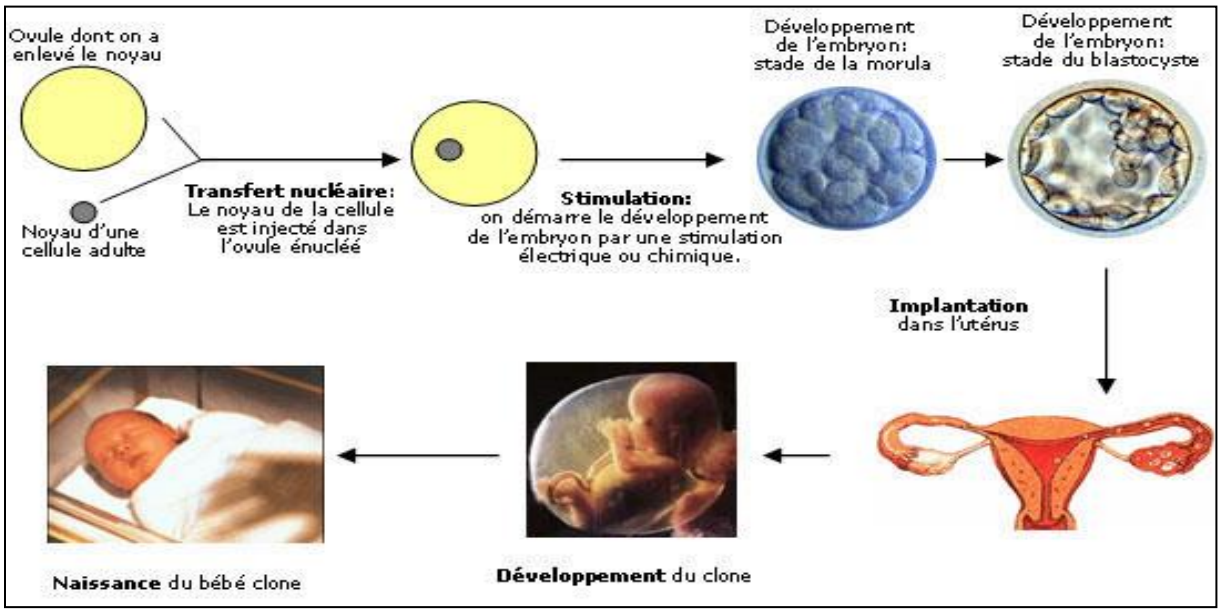


Figure 9 : Les etapes de clonage chez l'humain

📝 Exercices d'Application : Ingénierie et Éthique

Exercice 1 : Terminologie du Clonage

Vous lisez trois études scientifiques sur la manipulation génétique.

- **Étude A :** Utilisation de l'ADN d'un patient pour créer des cellules souches embryonnaires compatibles en vue de remplacer son tissu cardiaque endommagé.
- **Étude B :** Introduction d'un gène fonctionnel dans des cellules rétiniennes à l'aide d'un vecteur viral pour corriger une forme de cécité héréditaire.
- **Étude C :** Transfert du noyau d'une cellule somatique d'une vache primée dans un ovocyte énucléé pour obtenir un veau génétiquement identique à la vache originale.

Questions :

1. Identifier la technique décrite dans l'Étude B.
2. Identifier le type de clonage décrit dans l'Étude C.
3. Identifier le type de clonage décrit dans l'Étude A et justifier pourquoi il est souvent moins controversé que l'Étude C.

Solution

1. **Étude B :** Il s'agit de la **Thérapie Génique**, car elle utilise un vecteur (viral) pour introduire un gène fonctionnel (gène sain) afin de corriger une maladie génétique.
2. **Étude C :** Il s'agit du **Clonage Reproductif** par Transfert Nucléaire de Cellule Somatique (TNCS), car l'objectif est de donner naissance à un organisme entier, génétiquement identique à l'original.
3. **Étude A :** Il s'agit du **Clonage Thérapeutique**. Il est souvent moins controversé que le clonage reproductif (Étude C) car son objectif est purement médical et **ne vise pas à créer un nouvel individu**. Les embryons créés sont utilisés uniquement pour obtenir des cellules souches.

Exercice 2 : Distinction des Types de Clonage

Classifiez les scénarios d'ingénierie génétique suivants.

Scénario	Classification
A	Transférer le noyau d'une cellule de peau d'un animal d'élevage performant dans un ovocyte énucléé pour obtenir un jumeau génétique.
B	Insérer le gène de l'insuline humaine dans un plasmide bactérien pour produire la protéine en grande quantité.
C	Créer des cellules souches embryonnaires génétiquement compatibles avec un patient malade pour remplacer un tissu endommagé.

Questions :

1. Identifier la classification de chaque scénario (A, B et C).
2. Lequel de ces scénarios est le plus controversé éthiquement chez l'humain et pourquoi ?

Solution

1. **Classification des scénarios :**
    - **Scénario A : Clonage Reproductif** (par Transfert Nucléaire de Cellule Somatique - TNCS).
    - **Scénario B : Clonage Moléculaire** (ou Clonage de Gènes).
    - **Scénario C : Clonage Thérapeutique**.
2. Scénario le plus controversé :

Le Scénario A (Clonage Reproductif) est le plus controversé, car son objectif est de créer un organisme entier génétiquement identique à un autre, soulevant des questions éthiques fondamentales sur l'identité, la dignité et les risques sociétaux.

Chapitre 9 : Applications Thérapeutiques de la Génétique et Perspectives

9.1. Thérapie Génique

La **Thérapie Génique** vise à corriger une anomalie génétique en introduisant des gènes sains dans les cellules du patient pour remplacer ou compenser l'activité des gènes défectueux.

- **Vecteurs :** Elle utilise le plus souvent des **virus** (ex : adénovirus, lentivirus) modifiés pour les rendre inoffensifs. Ces virus servent de **navettes** pour transporter le gène thérapeutique jusqu'aux cellules cibles.
- **Applications :** Elle est prometteuse pour le traitement de maladies monogéniques (mucoviscidose, certaines immunodéficiences).

9.2. CRISPR/Cas9 : La Révolution de l'Édition du Génome

La technologie **CRISPR/Cas9** est un outil de pointe pour l'édition génomique. Il permet aux scientifiques d'éditer, avec une grande précision, des séquences d'ADN spécifiques dans le génome.

- **Principe :** Il utilise un ARN guide pour cibler l'endroit précis à modifier et une enzyme (Cas9) pour couper l'ADN à cet endroit. Une fois la coupure effectuée, la machinerie de réparation cellulaire permet d'insérer, de retirer ou de modifier des gènes.
- **Impact :** Outil puissant de la recherche fondamentale et des biotechnologies (création de modèles animaux d'étude), mais aussi la source de débats éthiques intenses sur la modification de la lignée germinale humaine.

9.3. Pharmacogénomique : Médecine de Précision

Comme vu au Chapitre 4, la **Pharmacogénomique** est une application directe de la génomique à la santé, visant à personnaliser les traitements en fonction du profil génétique de l'individu.

- **Base :** L'étude des **Polymorphismes d'un seul Nucléotide (SNP)**, qui peuvent affecter l'activité des enzymes chargées de métaboliser les médicaments (cytochrome P450).
- **Objectif :** Développer des tests pour prédire la réponse d'un patient à un médicament, assurant une plus grande efficacité thérapeutique et évitant des effets secondaires graves.

📝 Exercices d'Application : Ingénierie et Éthique

Exercice : CRISPR/Cas9 et Thérapie Génique

Le système CRISPR/Cas9 est le nouvel outil d'édition du génome.

Questions :

1. Quels sont les **deux composants moléculaires** essentiels du système CRISPR/Cas9, et quel est le rôle de chacun ?
2. Dans la thérapie génique, pourquoi utilise-t-on le plus souvent des **virus** (comme les adénovirus ou les lentivirus) pour transporter le gène thérapeutique dans les cellules du patient ?

Solution

1. **Composants de CRISPR/Cas9 :**
    - **Enzyme Cas9 :** C'est l'enzyme qui agit comme les **"ciseaux moléculaires"**. Son rôle est de **couper** précisément la double hélice d'ADN au site cible.
    - **ARN guide (ARNg) :** Son rôle est de **reconnaître** et de **guider** l'enzyme Cas9 vers la séquence d'ADN spécifique à modifier, par complémentarité de séquence.
2. **Utilisation des vecteurs viraux :**

Les virus sont utilisés comme vecteurs en thérapie génique car ils ont naturellement évolué pour être extrêmement efficaces à infecter des cellules et à insérer leur propre matériel génétique (le gène thérapeutique) dans le noyau de la cellule hôte, permettant ainsi l'expression du gène correcteur.

EXAMEN FINAL

📝 EXAMEN FINAL DU MODULE

Module : Bio-informatique

Destinataires : Étudiants en 2e Année Vétérinaire

Durée : 2 heures

Instructions : Répondez clairement et précisément aux questions. Justifiez vos réponses lorsque cela est demandé.

PARTIE I : Questions de Cours et Définitions (3 points)

Répondez en une ou deux phrases maximum.

Question 1.1 (0,5 pt) : Définissez la Bio-informatique en insistant sur son caractère interdisciplinaire et les domaines qu'elle fusionne.

Question 1.2 (0,5 pt) : Expliquez le principe de la réplication de l'ADN en utilisant le terme semi-conservateur.

Question 1.3 (0,5 pt) : Faites la distinction entre un Phylogramme et un Cladogramme dans le contexte des arbres phylogénétiques.

Question 1.4 (0,5 pt) : Quel est le rôle principal de l'enzyme Taq Polymérase dans la réaction de PCR et pourquoi est-elle spécifiquement choisie ?

Question 1.5 (0,5 pts) : Nommez les deux composants moléculaires essentiels du système d'édition du génome CRISPR/Cas9 et décrivez la fonction de chacun.

Question 1.6 (0,5 pts) : Expliquez la différence entre une mutation faux-sens (missense) et une mutation par décalage du cadre de lecture (frameshift) en termes de conséquences sur la protéine.

PARTIE II : Exercices d'Analyse et Applications (4 points)

Exercice 2.1 : Transcription et Traduction (2 pts)

Le brin transcrit (matrice) d'un gène essentiel chez un parasite est le suivant :

3' →TAC TTA GCA ATG TTT ACT → 5'

- **Rappel du Code :** AUG = Methionine (Start) ; UAA = Stop ; GCU = Alanine ; AAG = Lysine ; CAA = Glutamine ; UAU = Tyrosine.

1. Déterminez la séquence de l'**ARNm** mature (en précisant son sens 5'→3'). **(1 pt)**
2. Déterminez la séquence de la **protéine** correspondante. **(1 pts)**

Exercice 2.2 : Interprétation Bio-informatique (2 pts)

Vous analysez une séquence protéique (votre requête) contre une base de données de protéines virales à l'aide de l'outil BLAST.

1. Expliquez pourquoi l'algorithme de **Smith-Waterman** (Alignement Local) est plus pertinent que **Needleman-Wunsch** (Alignement Global) pour identifier un **domaine fonctionnel conservé** d'un virus à travers des espèces éloignées. **(1 pt)**
2. Un Hit (résultat) d'alignement est obtenu avec un **E-Value de $1 * 10^{-12}$** Que signifie cette valeur et pourquoi représente-t-elle une forte évidence d'homologie ? **(1 pt) :**

PARTIE III : Étude de Cas et Applications Vétérinaires (3 points)

Étude de Cas : Génomique et Pharmacogénomique Vétérinaire

Le gène *CYP2D6* code pour une enzyme (Cytochrome P450) cruciale pour le métabolisme de nombreux médicaments chez le chien (dont certains anesthésiques).
Une analyse génomique sur deux chiens révèle le statut de **Polymorphisme d'un seul Nucléotide (SNP)** suivant :

- **Chien Alpha (Métaboliseur Rapide) :** Possède la version sauvage du gène, ce qui lui confère une activité CYP2D6 **très élevée**.
- **Chien Beta (Métaboliseur Lent) :** Possède un SNP qui rend l'activité CYP2D6 **très faible**.

Questions :

1. Le médicament X est une **pro-drogue** (sa forme inactive est convertie en forme active par l'enzyme CYP2D6). Si les deux chiens reçoivent la même dose standard du médicament X, quel chien risque de ne pas bénéficier de l'effet thérapeutique souhaité ? Justifiez. **(1 pts)**
2. La **Protéomique** permet d'analyser l'état du protéome. Quelle technique de séparation permettrait de mettre en évidence que les enzymes CYP2D6 du Chien Alpha et du Chien Beta pourraient différer par leur **Point Isoélectrique (pI)** ? (Nommez la technique et le principe de séparation). **(1 pts)**
3. Dans le contexte de la surveillance épidémiologique d'un virus canin, si vous reconstruisez un arbre phylogénétique des souches. Comment le concept de l'**Outgroup** vous permet-il de déterminer le point d'origine ou la racine de l'arbre ? **(1 pts)**

PARTIE IV :**Cochez vrai ou faux (5 pts : 0.5 / proposition)**

n	Propositions	Vrai	Faux
1	UniProt archive exclusivement des séquences nucléotidiques		
2	GenBank est utilisé pour stocker des séquences d'ADN		
3	PDB (Protein Data Bank) contient uniquement des séquences d'ARN		
4	Les outils bioinformatiques sont indispensables pour l'annotation des génomes		
5	Un alignement de séquences compare des protéines mais pas l'AND		
6	L'algorithme Needleman-Wunsch est utilisé pour les alignements globaux		
7	FASTA est un format de fichier pour les séquences protéiques uniquement		
8	Les scores d'alignement dépendent des matrices de substitution		
9	Les gaps dans un alignement reflètent toujours des erreurs de séquençage		
10	Le NGS (Next-Generation Sequencing) est une méthode de séquençage haut debit		
11	Le séquençage de Sanger est plus rapide que le NGS		
12	Illumina est une plateforme de séquençage NGS		
13	La PCR est une étape obligatoire avant le séquençage NGS		
14	L'assemblage de novo nécessite un génome de référence		
15	La qualité des reads est évaluée par le score Phred		
16	Les codons STOP codent aussi pour des acides aminés		
17	Le biais d'utilisation des codons montre une préférence pour certains codons synonymes		
18	Les introns sont codants		
19	Le GC skew mesure le déséquilibre entre les bases G et C sur un brin d'AND		
20	Les pseudogènes sont fonctionnels		

PARTIE V : Choisir la ou les bonnes réponses : (5 pts : 0.25 / proposition)

1 La Bioinformatique :

A) Étudier exclusivement les écosystèmes naturels

B) Analyser et interpréter des données biologiques à l'aide d'outils informatiques

C) Développer des médicaments sans utiliser de modèles informatiques

D) Concevoir des robots pour la chirurgie

2 Quelle base de données est utilisée pour stocker des séquences d'ADN ?

A) PDB (Protein Data Bank)

B) GenBank

C) UniProt

D) KEGG

3 Qu'est-ce qu'un alignement de séquences en bioinformatique ?

A) Une méthode pour comparer des séquences d'ADN, d'ARN ou de protéines

B) Une technique de clonage moléculaire

C) Un algorithme de séquençage de nouvelle génération

D) Un protocole de purification de protéines

4 Qu'est-ce que le BLAST ?

A) Un algorithme d'alignement local de séquences

B) Un langage de programmation pour la biologie

C) Une base de données de structures protéiques

D) Un outil de séquençage de l'ADN

5 Quelle plateforme est utilisée pour l'analyse statistique en biologie ?

A) MATLAB

B) R/Bioconductor

C) Excel

D) Python sans librairies spécialisées

6 Qu'est-ce que le séquençage de nouvelle génération (NGS) :

A) Une méthode de séquençage haut débit

B) Un algorithme d'alignement

C) Une technique de clonage moléculaire

D) Un protocole de PCR

7 Quelle méthode est utilisée pour annoter des fonctions de gènes ?

A) Alignement avec des bases de données (ex : GO, KEGG)

B) Microscopie électronique

C) Spectrométrie de masse

D) Chromatographie en phase gazeuse

8 Qu'est-ce que le code génétique ?

A) Un système de règles qui associe chaque codon à un acide aminé

B) Une méthode de séquençage de l'ADN

C) Un algorithme informatique pour analyser les génomes

D) Une technique de clonage moléculaire

9 Combien de nucléotides forment un codon ?

A) 2

B) 3

C) 4

D) 5

10 Combien de codons différents existe-t-il ?

A) 16

B) 20

C) 64

D) 128

11 Quel codon code pour le démarrage de la traduction chez les eucaryotes ?

A) UAA

B) UGA

C) AUG

D) UAG

12 Quels sont les codons STOP (non-sens)

A) AUG, UAA, UGA

B) UAA, UAG, UGA

C) UGG, UAA, AUG

D) AUC, UAG, UGA

13 Le code génétique est-il redondant ?A) Non, chaque codon code pour un seul acide aminé

B) Oui, plusieurs codons peuvent coder le même acide aminé

C) Seulement chez les bactéries

D) Seulement chez les plantes

14 Quel acide aminé est codé par le plus grand nombre de codons ?

A) Leucine (6 codons)

B) Méthionine (1 codon)

C) Tryptophane (1 codon)

D) Histidine (2 codons)

15 Qu'entend-on par "biais d'utilisation des codons" ?

A) Une erreur systématique dans la traduction des ARNm

B) Une préférence pour certains codons synonymes plutôt que d'autres

C) Une mutation fréquente dans les gènes essentiels

D) Un défaut dans la réplication de l'AD

16 Qu'est-ce que le GC skew ?

A) Une méthode de séquençage de l'ADN

B) Mesure de la différence en pourcentage entre les bases G et C sur un brin d'ADN

C) Un outil d'analyse des protéines

D) Une technique de PCR

17 Qu'est-ce qu'un motif biologique en bioinformatique ?

A) Une séquence aléatoire d'ADN

B) Un modèle représentant un modèle conservé dans les séquences biologiques

C) Une erreur de séquençage

D) Un type de mutation ponctuelle

18 Quel est le but principal de l'analyse des motifs en génomique ?

A) Identifier des régions fonctionnelles

B) Calculer les taux de mutation

C) Déterminer la taille du génome

D) Analyser l'expression génique

N°,Propositions,Vrai,Faux

19 L'orthologie décrit

A) Des gènes similaires chez des espèces différentes

B) Des gènes similaires chez de meme espèce

C) Des gènes similaires issus d'un événement de duplication

D) La **spéciation**. La similarité due à la **duplication** au sein d'un même génome est la **paralogie**

20 L'Outgroup est utilisé dans

A) les arbres phylogénétiques pour déterminer la dynamique

B) Les arbres phylogénétiques pour déterminer le sens du temps évolutif

C) Les arbres phylogénétiques pour déterminer le sens du temps non évolutif

D) Les arbres phylogénétiques pour déterminer le contre sens du temps évolutif

Corrigé-type

PARTIE I : Questions de Cours et Définitions

Q.	Réponse Attendue
1.1	La Bio-informatique est l'application des outils informatiques (algorithmes, statistiques) et de la technologie pour analyser, interpréter et gérer les données biologiques (séquences, structures, fonctions).
1.2	La réplication est **semi-conservatrice** car chaque nouvelle molécule d'ADN double brin produite contient **un brin parental** (ancien) et **un brin nouvellement synthétisé**.
1.3	Le **Phylogramme** a des longueurs de branches significatives, représentant la distance génétique ou le temps évolutif. Le **Cladogramme** a des branches de longueur arbitraire et ne représente que la topologie (l'ordre de branchement).
1.4	La **Taq Polymérase** est l'ADN polymérase utilisée en PCR. Elle est choisie pour sa **thermorésistance**, lui permettant de résister aux hautes températures de la phase de dénaturation.
1.5	**Enzyme Cas9** : Coupe la double hélice d'ADN au site précis. **ARN guide (ARNg)** : Reconnaît la séquence cible et dirige Cas9 vers ce site.
1.6	Une mutation **faux-sens** change un seul acide aminé. Une mutation par **décalage du cadre de lecture** (due à une insertion/délétion non-multiple de trois) décale la lecture de tous les codons en aval, altérant gravement toute la séquence protéique.

PARTIE II : Exercices d'Analyse et Applications

Exercice 2.1 : Transcription et Traduction

1. ARNm mature (5'→3') :

   5' - AUG AAU CGU UAC AAA UGA - 3'

2. **Séquence de la protéine :**
   - AUG = Méthionine (Start)
   - AAU = Asparagine (Non codé, mais déduit ou supposé)
   - CGU = Arginine (Non codé, mais déduit ou supposé)
   - UAC = Tyrosine
   - AAA = Lysine
   - UGA = STOP

Séquence : Méthionine - Asparagine - Arginine - Tyrosine - Lysine - STOP

Exercice 2.2 : Interprétation Bio-informatique

1. Pertinence de Smith-Waterman :

   L'alignement local de Smith-Waterman est préférable car il cherche les régions de plus grande similarité locale (les domaines conservés) sans pénaliser les parties divergentes (non homologues) situées avant ou après, ce qui est typique des séquences éloignées ou des gènes avec des domaines fonctionnels partagés.

   Signification de l'E-Value :

   L'E-Value est le nombre d'alignements avec un score aussi bon ou meilleur que l'on s'attendrait à trouver par pur hasard dans la base de données. Une valeur de $1* 10^{-12}$ est extrêmement faible et signifie que la probabilité que cette correspondance soit fortuite est presque nulle. Cela indique donc une très forte évidence d'homologie (ancêtre commun).

PARTIE III : Étude de Cas et Applications Vétérinaires (6 points)

1. Conséquence pour le Chien Alpha (Pro-drogue) :

   Le Chien Alpha possède une enzyme très active. Il va donc convertir la pro-drogue en forme active très rapidement. La forme active sera ensuite elle-même rapidement dégradée. Le chien risque de ne pas avoir une concentration suffisante de la forme active pour une durée assez longue. Il risque donc une perte d'efficacité thérapeutique (sous-dosage fonctionnel).

2. Technique de séparation (pI) :

   La technique permettant la séparation des protéines selon leur Point Isoélectrique (pI) est l'Isoélectrofocalisation (IEF), qui constitue la première dimension de l'Électrophorèse Bidimensionnelle (2D-E).

3. Rôle de l'Outgroup :

   L'Outgroup est une souche virale (ou une espèce) dont on sait, par des preuves indépendantes, qu'elle est moins apparentée à toutes les souches étudiées. En se connectant à la base de l'arbre, l'Outgroup enracine l'arbre et indique le point de divergence le plus ancien, permettant ainsi de situer la racine de l'arbre et de déterminer la direction temporelle de l'évolution du virus.

PARTIE IV : Vrai / Faux

N°	Vrai	Faux
1		X
2	X	
3		X
4	X	
5		X
6	X	
7		X
8	X	
9		X
10	X	
11		X
12	X	
13		X
14		X
15	X	
16		X
17	X	
18		X
19	X	
20		X

PARTIE V :

N°	Question	Bonne(s) Réponse(s)	Justification
1	La Bioinformatique :	**B**	Elle utilise des outils informatiques pour analyser, gérer et interpréter les données biologiques (séquences, structures, fonctions).
2	Quelle base de données est utilisée pour stocker des séquences d'ADN ?	**B**	**GenBank** est la base de données de référence pour les séquences nucléotidiques. PDB est pour les structures 3D, UniProt pour les protéines.
3	Qu'est-ce qu'un alignement de séquences en bioinformatique ?	**A**	C'est une méthode fondamentale pour comparer des séquences afin d'identifier les similarités et les homologies.
4	Qu'est-ce que le BLAST ?	**A**	**BLAST** (*Basic Local Alignment Search Tool*) est l'outil heuristique le plus populaire pour l'alignement **local** de séquences.
5	Quelle plateforme est utilisée pour l'analyse statistique en biologie ?	**B**	**R** et sa librairie **Bioconductor** sont la plateforme standard et la plus puissante pour l'analyse statistique des données génomiques et biologiques.
6	Qu'est-ce que le séquençage de nouvelle génération (NGS) ?	**A**	C'est un ensemble de technologies qui permettent un **séquençage haut débit** (millions de fragments en parallèle), par opposition au séquençage de Sanger.
7	Quelle méthode est utilisée pour annoter des fonctions de gènes ?	**A**	L'annotation des fonctions repose souvent sur la comparaison (alignement) du gène avec des bases de données de

			fonctions connues (Gene Ontology - **GO**, ou **KEGG**).
8	Qu'est-ce que le code génétique ?	**A**	C'est l'ensemble des règles qui définissent comment les **codons** (triplets de nucléotides) sont traduits en **acides aminés**.
9	Combien de nucléotides forment un codon ?	**B**	Un codon est un **triplet** de nucléotides, soit 3.
10	Combien de codons différents existe-t-il ?	**C**	Il y a $4^3 = \mathbf{64}$ combinaisons possibles de triplets (4 bases possibles à 3 positions).
11	Quel codon code pour le démarrage de la traduction chez les eucaryotes ?	**C**	Le codon **AUG** est le codon initiateur, codant pour la Méthionine.
12	Quels sont les codons STOP (non-sens) ?	**B**	Les trois codons STOP sont **UAA, UAG, UGA**.
13	Le code génétique est-il redondant ?	**B**	**Oui**, car plusieurs codons différents peuvent coder le même acide aminé (il y a 61 codons pour seulement 20 acides aminés).
14	Quel acide aminé est codé par le plus grand nombre de codons ?	**A**	La **Leucine** et la **Sérine** sont codées par six codons différents, ce qui représente le maximum.
15	Qu'entend-on par "biais d'utilisation des codons" ?	**B**	C'est la tendance d'un organisme à utiliser préférentiellement certains codons synonymes pour un acide aminé donné.
16	Qu'est-ce que le GC skew ?	**B**	C'est la mesure de l'asymétrie entre les bases Guanine (G) et Cytosine (C) sur un brin d'ADN, souvent liée aux processus de réplication.

17	Qu'est-ce qu'un motif biologique en bioinformatique ?	**B**	C'est un modèle abstrait ou une séquence conservée qui représente une région fonctionnelle ou structurelle récurrente dans les séquences (ADN ou protéines).
18	Quel est le but principal de l'analyse des motifs en génomique ?	**A**	L'identification des motifs conservés permet de **déduire les régions fonctionnelles** (sites de liaison, domaines protéiques, signaux de régulation, etc.).
19	L'orthologie décrit des gènes similaires chez des espèces différentes, issus d'un événement de duplication.	**B**	L'orthologie résulte de la **spéciation**. La similarité due à la **duplication** au sein d'un même génome est la **paralogie**.
20	L'Outgroup est utilisé dans les arbres phylogénétiques pour déterminer le sens du temps évolutif (racinement).	**A**	L'Outgroup, par sa divergence précoce connue, sert à **raconter** l'arbre en indiquant la direction du temps évolutif.

Evaluation TD

🔢 Evaluation des travaux dirigees : Amplification PCR et Qualité de Séquençage

Partie A : Amplification Exponentielle par PCR

Un laboratoire d'analyse vétérinaire doit détecter la présence d'un gène de virulence chez un agent pathogène. Après extraction et quantification, le tube réactionnel initial contient une quantité estimée de 30 copies d'ADN matrice (le gène cible).

L'objectif est d'atteindre un niveau d'amplification suffisant pour la détection (environ 1.5 million de copies).

Rappel : Le principe de la PCR est une amplification exponentielle où, à chaque cycle, la quantité d'ADN est théoriquement doublée. La formule de l'amplification est : $N_{final} = N_{initial} *2^{C}$, où C est le nombre de cycles.

Questions :

1. Nombre de Copies : Combien de copies d'ADN cible le technicien obtiendra-t-il après 15 cycles complets de PCR ?
2. Seuil de Détection : Si le seuil de détection du système (la quantité minimale nécessaire pour observer un signal) est de 1,536,000 copies. Combien de cycles C sont nécessaires pour atteindre ce seuil, en partant des 30 copies initiales ? (*Aide : vous pouvez utiliser l'approximation* $log_2 \{X\} = ln(X)\backslash ln(2)$ *1.44 * ln(X*).

Partie B : Qualité de Séquençage (Score Phred)

La qualité des données de séquençage NGS (Next-Generation Sequencing) est souvent exprimée par le Score Phred (Q).

Ce score est lié à la probabilité d'erreur (P) de l'appel de base par la formule :

$$Q = -10 \log_{10} (P)$$

Un score Q30 est souvent requis pour les analyses génomiques de haute qualité.

Questions :

1. Probabilité d'Erreur : Quelle est la probabilité d'erreur P associée à un nucléotide ayant un Score Phred de Q30 ?
2. Fiabilité : Si le séquenceur attribue un Score Phred de Q20 à une base, cela correspond à une erreur sur combien de bases séquencées ?

🔑 Solution Détaillée de l'Exercice

Solution Partie A : Amplification Exponentielle par PCR

1. Nombre de Copies après 15 cycles

- $N_{initial}$ = 30 copies
- C = 15 cycles

$N_{final} = 30 * 2^{15}$

$N_{final} = 30 * 32,768$

$$\mathbf{N_{final} = 983,040 \text{ copies}}$$

Réponse : Le technicien obtiendra 983 040 copies après 15 cycles, ce qui est légèrement inférieur au seuil de détection de 1.5 million.

2. Seuil de Détection (Nombre de Cycles C)

- $N_{initial}$ = 30 copies
- N_{final} = 1,536,000 copies

Nous devons résoudre l'équation pour C :

$1,536,000 = 30 * 2^{C}$

Étape 1 : Isoler la puissance de 2

Étape 2 : Utiliser le logarithme

Puisque le nombre de cycles doit être un entier et que la détection n'est complète qu'après le cycle, nous devons arrondir à l'entier supérieur.

Réponse : Il faut 16 cycles pour atteindre ou dépasser le seuil de détection.

Solution Partie B : Qualité de Séquençage (Score Phred)

1. Probabilité d'Erreur P pour Q_{30}

- Q = 30

Formule : $Q = -10 \log_{10} (P)$

Étape 1 : Isoler $-10 \log_{10} (P)$

Étape 2 : Calculer P

$P = 10^{-3}$

P = 0.001 ou 0.1%

Réponse : Un Score Phred Q30 signifie que la probabilité que la base lue soit incorrecte est de 1 sur 1000 (P=0.001).

2. Fiabilité de Q20

- Q = 20

Étape 1 : Calculer P

$\log_{10} (P) = -2$

$P = 10^{-2}$

P = 0.01

Étape 2 : Interpréter la probabilité

Une probabilité d'erreur $P=0.01$ signifie 1 erreur pour 100 bases séquencées (en moyenne).

Réponse : Un Score Phred Q20 correspond à une erreur attendue sur 100 bases séquencées. (Ce niveau est généralement le minimum acceptable pour les *reads* d'entrée.)

Lexique Français-Anglais

Lexique Technique Français-Anglais

Terme Français	Terme Anglais	Définition Anglaise
I. MATÉRIEL GÉNÉTIQUE ET VARIATION		
ADN (Acide Désoxyribonucléique)	DNA (Deoxyribonucleic Acid)	Hereditary material that carries genetic instructions for all organisms.
ARN (Acide Ribonucléique)	RNA (Ribonucleic Acid)	Single-stranded molecule involved in gene expression (transcription, translation).
Nucléotide	Nucleotide	The basic structural unit of DNA and RNA, composed of a base, a sugar, and a phosphate group.
Gène	Gene	A segment of DNA that contains instructions for making a specific protein or functional RNA molecule.
Génome	Genome	The entire set of genetic material in an organism.
Chromosome	Chromosome	A thread-like structure of nucleic acids and protein found in the nucleus, carrying genetic information.
Mutation	Mutation	A permanent alteration in the DNA sequence.
SNP (Polymorphisme d'un seul nucléotide)	SNP (Single Nucleotide Polymorphism)	A common variation in a DNA sequence that affects a single base pair.

Locus	Locus	The specific physical location of a gene or other DNA sequence on a chromosome.
Homologie	Homology	Similarity between biological structures or sequences due to shared ancestry.
Orthologie	Orthology	Genes in different species that evolved from a common ancestral gene via speciation.
Paralogie	Paralogy	Genes related by duplication within a genome.
II. PROCESSUS BIOLOGIQUES FONDAMENTAUX		
Réplication	Replication	The process by which DNA makes a copy of itself.
Transcription	Transcription	The process of copying a segment of DNA into RNA (mRNA).
Traduction	Translation	The process by which a sequence of mRNA nucleotides is converted into a sequence of amino acids (protein).
Codon	Codon	A sequence of three nucleotides on an mRNA that codes for a specific amino acid.
Codon-stop	Stop Codon	A triplet sequence (UAA, UAG, or UGA) that signals the termination of translation.
Protéine	Protein	Large biomolecules composed of amino acid chains, performing

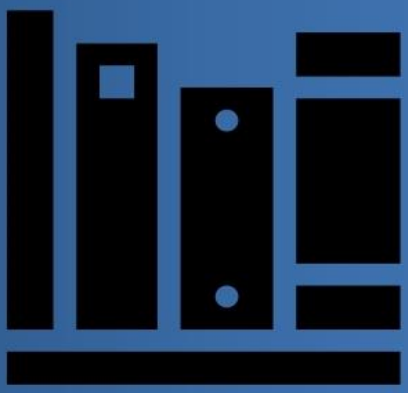
		structural and enzymatic functions.
Acide Aminé	Amino Acid	The monomer unit that makes up proteins.
III. BIO-INFORMATIQUE ET PHYLOGÉNIE		
Bio-informatique	Bioinformatics	The application of computational techniques to analyze and manage biological data.
In silico	In silico	Performed on a computer or via computer simulation.
Alignement de Séquences	Sequence Alignment	A method to arrange DNA, RNA, or protein sequences to identify regions of similarity.
Alignement Global	Global Alignment	Alignment of two sequences across their entire length (e.g., Needleman-Wunsch).
Alignement Local	Local Alignment	Alignment of segments of sequences to find regions of high similarity (e.g., Smith-Waterman).
Heuristique	Heuristic	A computational method that is faster but does not guarantee the optimal solution (e.g., BLAST).
BLAST	BLAST	(Basic Local Alignment Search Tool) A heuristic program for comparing a query sequence against a database.
E-Value	E-Value (Expectation Value)	The number of random hits expected to achieve a score as

		good or better than the observed one.
Phylogénie	Phylogeny	The study of evolutionary relationships among groups of organisms (taxa).
Dendrogramme	Dendrogram	A tree diagram used to illustrate the arrangement of clusters produced by hierarchical clustering.
Phylogramme	Phylogram	A phylogenetic tree where branch lengths are proportional to the amount of evolutionary change.
Cladogramme	Cladogram	A phylogenetic tree where branch lengths are arbitrary, focusing only on the order of branching (topologies).
Outgroup	Outgroup	A more distantly related group of organisms that serves as a reference point for rooting a phylogenetic tree.
IV. TECHNIQUES DE LABORATOIRE		
PCR (Réaction en Chaîne par Polymérase)	PCR (Polymerase Chain Reaction)	A technique used to amplify small segments of DNA across several orders of magnitude.
Taq Polymérase	Taq Polymerase	A heat-stable DNA polymerase used in PCR.
Amorce (Oligonucléotide)	Primer	A short, synthetic DNA strand that serves as a starting point for DNA synthesis.

Thermocycleur	Thermal Cycler	A laboratory apparatus used to amplify DNA and RNA samples by cycling through thermal stages.
Séquençage	Sequencing	The process of determining the precise order of nucleotides within a DNA or RNA molecule.
ddNTP (Didésoxyribonucléotide)	ddNTP (Dideoxynucleotide)	Nucleotides used in Sanger sequencing to terminate the growth of the DNA chain.
Électrophorèse Bidimensionnelle (2D-E)	Two-Dimensional Electrophoresis (2D-E)	A technique separating proteins based on two properties: isoelectric point (pI) and molecular weight (MW).
Protéomique	Proteomics	The large-scale study of proteins, particularly their structure and functions.
V. INGÉNIERIE ET APPLICATIONS		
Génomique	Genomics	The study of genomes, including their structure, function, evolution, and mapping.
Pharmacogénomique	Pharmacogenomics	The study of how genes affect a person's response to drugs, allowing personalized medicine.
Thérapie Génique	Gene Therapy	The introduction of genetic material into cells to replace or correct faulty genes.
CRISPR/Cas9	CRISPR/Cas9	A powerful genome editing tool that allows precise modification of genetic material.

Clonage	Cloning	The process of producing genetically identical individuals or molecules.
TNCS (Transfert Nucléaire de Cellule Somatique)	SCNT (Somatic Cell Nuclear Transfer)	A technique for creating a clone by transferring a somatic cell nucleus into an enucleated egg cell.
Lignée Germinale	Germline	The cell line that gives rise to gametes (sperm and egg), meaning changes are heritable.

References et bibliographiques



🕮 Références Bibliographiques et Ressources pour Étudiants

Ces références sont structurées pour fournir à la fois les bases théoriques et les outils pratiques indispensables en Biologie Moléculaire, Génétique et Bio-informatique.

I. Manuels Fondamentaux (Théorie et Biologie Moléculaire)

Ces ouvrages sont des références mondiales pour les concepts du Dogme Central, des mutations et des structures macromoléculaires.

Réf.	**Auteur(s) / Éditeur(s)**	**Titre et Édition (Français si disponible)**	**Utilité pour le Support**
B1	Alberts, B. *et al.*	**Biologie Moléculaire de la Cellule** (6e édition ou plus récente, ou l'équivalent : *Molecular Biology of the Cell*).	Référence incontournable pour la structure de l'ADN, la réplication, la transcription, la traduction, et le cycle cellulaire.
B2	Lodish, H. *et al.*	**Biologie Cellulaire et Moléculaire** (ou l'équivalent : *Molecular Cell Biology*).	Excellent complément sur les mécanismes de régulation génique et la fonction des protéines.
B3	Griffiths, A. J. F. *et al.*	**Introduction à l'Analyse Génétique** (ou l'équivalent : *Introduction to Genetic Analysis*).	Approfondissement des concepts de gène, de mutation, de variation génétique (SNP) et de génomique.

II. Ouvrages Spécialisés (Bio-informatique, Techniques et Applications)

Ces références sont plus axées sur la méthodologie de l'analyse des séquences et les applications technologiques.

Réf.	**Auteur(s) / Éditeur(s)**	**Titre et Édition (Français si disponible)**	**Utilité pour le Support**
S1	Durbin, R. *et al.*	**Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids** (Anglais).	Référence académique pour l'algorithmique (Needleman-Wunsch, Smith-Waterman) et les modèles d'alignement.
S2	Lesk, A. M.	**Introduction à la Bio-informatique** (ou *Introduction to Bioinformatics*).	Un bon manuel d'introduction aux bases de données, à l'alignement et à la phylogénie.
S3	Brown, T. A.	**Gene Cloning and DNA Analysis: An Introduction** (Anglais).	Excellent pour comprendre les techniques de PCR, séquençage et les principes de l'ingénierie génétique (Clonage, Thérapie Génique).

III. Références en Ligne (Outils et Bases de Données)

La pratique de la bio-informatique nécessite l'utilisation directe de plateformes et de bases de données.

Réf.	Plateforme / Organisation	Ressource	Utilité Pratique
O1	**NCBI (National Center for Biotechnology Information)**	**GenBank & BLAST**	Outil indispensable pour l'alignement de séquences (BLAST) et l'accès aux banques de séquences d'ADN et de protéines.
O2	**UniProt / Expasy**	**UniProt Knowledgebase**	La base de données de référence pour les séquences de protéines, les annotations fonctionnelles et les données protéomiques.
O3	**EBI (European Bioinformatics Institute)**	**Clustal Omega**	Outil d'alignement multiple de séquences en ligne (pour les exercices de phylogénie).

IV. Articles Fondateurs et Clés

Ces articles historiques sont souvent cités dans les cours et sont cruciaux pour comprendre l'origine des techniques.

Réf.	**Auteur(s) / Année**	**Sujet / Titre**	**Contribution au Domaine**
A1	**Sanger, F. *et al.* (1977)**	Le séquençage didésoxy (Méthode de Sanger).	Article décrivant la première méthode de séquençage d'ADN rapide.
A2	**Mullis, K. B. (1990)**	Article ou brevet sur la PCR.	Invention de la technique d'amplification de l'ADN *in vitro* (PCR).
A3	**Altschul, S. F. *et al.* (1990)**	Algorithme BLAST.	Publication décrivant le fonctionnement statistique de l'outil BLAST pour l'alignement local rapide.
A4	**Jinek, M. *et al.* (2012)**	Découverte du système CRISPR/Cas9.	Article clé décrivant le mécanisme d'édition du génome par CRISPR/Cas9.

@@@