République Algérienne Démocratique et Populaire
Université Larbi Ben M'hidi Oum El-Bouaghi
Faculté des Sciences Exactes et Sciences de la Nature et de la Vie
Département des Sciences de la Nature et de la Vie

Cours de génomique et bioinformatique

Destiné aux étudiants de la 2ième Année Master Biochimie Appliquée

Dr. Benslama Ouided

MCA. Université Larbi Ben M'Hidi Oum El Bouaghi

Année universitaire 2022-2023

Avant-Propos

L’idée de ce polycopié est née de l’expérience d’enseigner le cours de génomique et bioinformatique à l'université de Larbi Ben M’Hidi à Oum El Bouaghi pour les étudiants de deuxième année Master biochimie appliquée. Ce polycopié est en réalité une collection de notes de cous que j’ai essayé de développer, en se basant sur des revues et des travaux pertinents dans la discipline de la génomique et la bioinformatique, et de les rassembler afin de les rendre accessibles à un public plus large.

Ce cours est organisé en cinq chapitres inspiré du canevas de la matière. Il fournit une vue d'ensemble et une introduction au domaine de la génomique structurale et fonctionnelle d’une part et au domaine de la bioinformatique d’une autre part.

La génomique est l'étude de génomes entiers d'organismes et intègre des éléments de la génétique. La génomique utilise une combinaison d'ADN recombinant, de méthodes de séquençage d'ADN et de bioinformatique pour séquencer, assembler et analyser la structure et la fonction des génomes. Elle diffère de la « génétique classique » en ce qu'elle considère l'ensemble du matériel héréditaire d'un organisme, plutôt qu'un gène ou un produit génique à la fois. De plus, la génomique se concentre sur les interactions entre les locus et les allèles au sein du génome.

La bioinformatique, liée à la génétique et à la génomique, est une sous-discipline scientifique qui consiste à utiliser la technologie informatique pour collecter, stocker, analyser et diffuser des données et des informations biologiques, telles que des séquences d'ADN et d'acides aminés ou des annotations sur ces séquences. Ainsi, les outils bioinformatiques aident à la comparaison des données génétiques et génomiques et plus généralement à la compréhension des aspects évolutifs de la biologie moléculaire. À un niveau plus intégratif, il aide à analyser et à cataloguer les voies et réseaux biologiques qui constituent une partie importante de la biologie des systèmes.

À l'aide de la bioinformatique, les chercheurs en génomique analysent d'énormes quantités de données de séquences d'ADN pour trouver des variations qui affectent la santé, la maladie ou la réponse aux médicaments.

Dr. BENSLAMA Ouided

Table de matières

4.5.Phosphorylation des histones
5. Récepteurs nucléaires et contrôle de l'initiation transcriptionnelle
5.1.Structure du récepteur nucléaire
5.2.Coactivateurs de récepteurs nucléaires
5.3.Corepresseurs des récepteurs nucléaires
6. Contrôle post-transcriptionnel de l'expression génique
6.1. Épissage alternatif d'ARN
6.2. Contrôle de la stabilité de l'ARN
7. Contrôle post-transcroptionelles de l'expression génique
8. Contrôle traductionnel de l'expression génique
9. Exemple de régulation artificielle des gènes

Chapitre 3 : Variabilité des génomes : Polymorphisme

1. Introduction
2. Sources de variations : les mutations
2.1.Types de mutation
2.1.1. Substitution
2.1.2. Insertion
2.1.3. Suppressions
2.1.4. Décalage de cadre
2.2.Les effets des mutations sur les gènes
2.3.Les effets des mutations sur les organismes multicellulaires
3. Phénotypage moléculaire : Les marqueurs
3.1.Marqueurs morphologiques
3.2.Marqueurs cytologiques
3.3.Marqueurs biochimiques
3.4.Marqueurs moléculaires (marqueurs à base d'ADN)
3.4.1. Marqueurs RFLP (Restriction fragment length polymorphism)
3.4.1.1.Principe du RFLP
3.4.1.2.Applications du RFLP
2.4.2. Marqueurs RAPD (Rapid amplified polymorphic DNA)
2.4.3. Marqueurs AFLP (Amplified fragment length polymorphism)
2.4.3.1.Principe de l'AFLP
2.4.3.2. Étapes de l'AFLP
2.4.3.3.Applications
2.5. Marqueurs ADN microsatellites
2.6. Marqueurs SNP (single nucleotide polymorphism)

Chapitre 4. Construction d'une banque génomique

1. Définition d'une banque génomique
2. Avantages d'une bibliothèque génomique
3. Type de banques génomiques
3.1. Banque d'ADN génomique
3.2.Banque d'ADNc (banque d'ADN complémentaire)
4. Clonage moléculaire
4.1.Technique du clonage
4.2.Protocole du clonage
4.2.1. Préparation du vecteur de clonage
4.2.2. Préparation de l'insert

4.2.3. Assemblage (ligature)
4.2.4. Transformation
4.3.Criblage (screening) des clones
4.4.Sélection des clones cibles
4.4.1. Hybridation fluorescente in situ (FISH)
4.4.1.1.Etape de la méthode FISH
5. Génomique fonctionnelle
5.1.Techniques de la génomique fonctionnelle
5.1.1. Puces à ADN
5.1.1.1. Définition d'une puce à ADN
5.1.1.2.Principe de la puce à ADN
5.1.1.3.Étapes impliquées dans les biopuces à base d'ADNc
5.1.1.4.Applications de la technique des puces à ADN
5.1.2. Technologies de séquençage de nouvelle génération

Chapitre 5. Bioinformatique fonctionnelle

1. Introduction
2. Analyse d'une famille de séquence protéique
2.1.Raisons de choisir l'étude de séquences protéiques
2.2.Principales bases de données de séquences de protéines
2.2.1. Modèles et profils PROSITE
2.2.2. Pfam
2.2.3. SMART
2.2.4. TIGRFAM
2.2.5. FingerPRINTS
3. Banques de données
3.1.Définition d'une base de données
3.2.Types des bases de données
3.2.1. Les bases de données primaires
3.2.2. Les bases de données secondaires
3.3.GenBank
3.3.1. Définition
3.3.2. Organisation de la base de données GenBank
4. Notion d'analyse phylogénétique
4.1.Introduction à la phylogénétique moléculaire
4.2.Terminologie
4.3.Arbres enracinés et Arbres non enracinés
4.4.Enracinement d'un arbre phylogénétique par un outgroup
4.5.Enracinement d'un arbre phylogénétique par l'approche de l'enracinement au milieu « midpoint rooting »
4.6.Types topologiques des arbres phylogénétiques
4.7.Procédure de l'établissement d'un arbre phylogénétique
4.7.1. Calcul de la distance entre les OTUs (séquences)
a- Cas de séquences nucléiques
4.8.Détermination d'une méthode de construction de l'arbre phylogénétique
4.8.1. Méthodes basées sur les distances
4.8.1.1.Unweight Pair Group Method with Arithmetic mean (UPGMA)
4.8.1.2.Neighbor-Joining (NJ)

Chapitre 01

Organisation des génomes et structures des gènes

Chapitre 1. Organisation des génomes et structure des gènes

3. Définition du génome

Le terme génome a été inventé en 1920 pour décrire "l'ensemble de chromosomes haploïdes, qui, avec le protoplasme pertinent, spécifie les fondements matériels de l'espèce". Bien que la génétique mendélienne ait été redécouverte en 1900 et que les chromosomes aient été identifiés comme porteurs de l'information génétique en 1902, on ne savait pas en 1920 si l'information génétique était portée par l'ADN ou le composant protéique des chromosomes. De plus, le mécanisme par lequel la cellule copie des informations dans de nouvelles cellules et convertit ces informations en fonctions était inconnu pendant plusieurs décennies après l'invention du terme « génome ».

Aujourd'hui, cependant, nous sommes inondés de données génomiques. Une version récente de la base de données GenBank , version 210.0 (publiée le 15 octobre 2015), contient plus de 621 milliards de paires de bases provenant de 2 557 génomes eucaryotes, 432 génomes archéens et 7 474 génomes bactériens, ainsi que des dizaines de milliers de virus. Nous avons également maintenant une compréhension beaucoup plus large et plus détaillée de la façon dont le génome est exprimé et de la façon dont différents facteurs biologiques et environnementaux contribuent à ce processus. Même ainsi, près d'un siècle après avoir inventé le terme, la définition standard du génome reste très similaire à son prédécesseur de 1920. Par exemple, sur son site Web Genetics Home Reference, la définition des National Institutes of Health (NIH) se lit comme suit : « L'ensemble complet d'ADN d'un organisme, y compris tous les gènes, constitue le génome. Chaque génome contient toutes les informations nécessaires pour construire et maintenir cet organisme ».

4. Organisation des génomes

4.1.Génome des eucaryotes : l’homme comme exemple

Le génome haploïde humain se compose d'environ 3 x10^9 paires de bases d'ADN. L'ADN génomique existe sous forme de morceaux d'ADN linéaires uniques associés à une protéine appelée complexe nucléoprotéique. Le complexe ADN-protéine est à la base de la formation des chromosomes, pratiquement tout l'ADN génomique est réparti entre les 23 chromosomes qui résident dans le noyau cellulaire. Une très petite fraction du génome se trouve également dans un morceau d'ADN circulaire de 16 000 paires de bases qui se trouve dans les mitochondries. L'ADN en double hélice de la chromatine est répliqué avec la fibre de chromatine se condensant en corps discrets, les chromosomes, chacun composé de deux

chromatides identiques. Les deux chromatides sœurs se séparent, l'une se déplaçant vers chaque pôle de la cellule, où elles font partie du noyau nouvellement formé de chaque cellule fille. Les cellules qui composent la majeure partie du corps d'un organisme multicellulaire, les cellules somatiques, possèdent deux copies de chaque chromosome et sont dites diploïdes (2n). L'ovule et le sperme par exemple, produits par la méiose et n'ayant qu'un seul exemplaire de chaque chromosome, sont haptoïdes (n). L'ADN de la chromatine et des chromosomes est étroitement lié à une famille de protéines chargées positivement, les histones, qui s'associent fortement aux nombreux groupes phosphate chargés négativement de l'ADN. Les histones et l'ADN s'associent dans des complexes appelés nucléosomes dans lesquels le brin d'ADN s'enroule autour d'un noyau de molécules d'histone.

4.2.Éléments fonctionnels et distribution de l'ADN dans le génome

La fonction principale de l'ADN génomique est de transporter et de stocker des informations génétiques exprimées sous forme d'ARN, puis de protéines fonctionnelles. Pour que l'expression génique se produise correctement, il doit y avoir des éléments régulateurs présents sur le génome et le génome doit être fidèlement répliqué et séparé entre les cellules filles.

4.3.Éléments d'ADN requis pour la réplication et la ségrégation du génome

D'après des études sur des eucaryotes unicellulaires (levures), au moins trois types d'éléments d'ADN sont nécessaires à la réplication et à l'héritage stable des chromosomes : les séquences à réplication autonome (ARS), les centromères et les télomères. Les séquences de réplication autonome (ARS) sont les sites sur lesquels la réplication de l'ADN est initiée sur les chromosomes. Les centromères sont des séquences d'ADN nécessaires à la ségrégation des chromosomes répliqués dans les cellules filles. Télomères (voir conférence "Synthèse de l'ADN") La télomérase reconnaît les extrémités des chromosomes également appelées télomères. Les séquences d'ADN des télomères ont été déterminées dans plusieurs organismes et consistent en de nombreuses répétitions d'une séquence longue de 6 à 8 bases, [TTGGGG]n. Les chromosomes artificiels de levure ou YAC peuvent être construits en combinant de grands segments d'ADN humain (50 000 paires de bases ou plus) avec un marqueur sélectionnable et les trois éléments essentiels décrits ci-dessus. Ces chromosomes artificiels peuvent ensuite être propagés et amplifiés dans des cellules de levure. Cette technologie est utilisée dans le séquençage du génome humain.

4.4.Répartition des éléments constructifs du génome

Une partie prédominante du génome humain est constituée de séquences répétitives de divers types englobant de grandes duplications segmentaires, des répétitions intercalées dérivées de transposons et des répétitions en tandem [8]. Ces derniers comprennent les satellites, les minisatellites et les microsatellites connus également sous le nom de séquences répétées simples (SSR). En revanche, les séquences codantes des quelque 25 000 gènes humains représentent environ 1 % du génome, et environ la même part du génome peut être attribuée à leurs séquences régulatrices (Fig. 1). Cette distinction entre les séquences répétitives et codantes/régulatrices ne signifie pas nécessairement que ces séquences sont physiquement séparées les unes des autres dans le génome. Il arrive souvent que certaines de ces répétitions se produisent dans les gènes et même dans leurs séquences codantes et remplissent des fonctions régulatrices. Il arrive également que les répétitions augmentent la probabilité de mutations délétères dans leurs gènes hôtes, augmentant ainsi le risque de maladie.

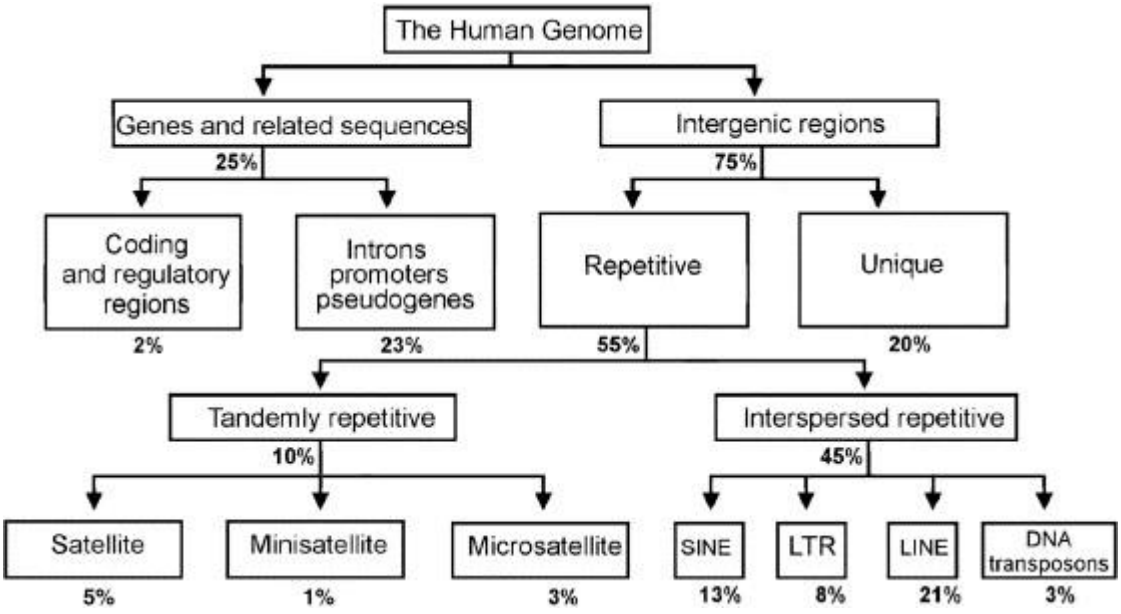


Figure 1. Composition du génome humain. Les parts en pourcentage de diverses séquences fonctionnelles et non fonctionnelles sont indiquées (24).

Le génome humain est fondamentalement uniforme et seulement environ 0,3 % de sa séquence diffère entre les individus d'une population. Cependant, l'information génétique est à l'origine d'une diversité beaucoup plus élevée observée au niveau du transcriptome. Le terme transcriptome est jeune et sa première apparition remonte à 1997 pour décrire l'ensemble des gènes exprimés à partir du génome de la levure. Selon une définition plus récente, le terme désigne l'ensemble complet des éléments transcrits du génome. En plus des ARNm, il comprend

également divers ARN non codants jouant des fonctions structurelles ou régulatrices dans les cellules. Par conséquent, il existe des milliers de transcriptomes différents dans des centaines de types de cellules et d'organes différents dans leurs divers états physiologiques et pathologiques. La variété des transcrits par rapport aux gènes augmente d'environ 50 % en raison de l'épissage alternatif, les nombreux transcrits antisens sont synthétisés, et tous les ARN non codant pour les protéines, dont la famille des microARN récemment reconnue apportent également une contribution significative au pool de transcrits humains.

On sait depuis un certain temps que les ARNm humains représentent environ 2 à 3 % de l'ARN cellulaire et qu'ils peuvent être caractérisés comme appartenant à différentes classes d'abondance. Un petit nombre d'ARNm sont synthétisés en plusieurs milliers de copies, d'autres se produisent en centaines de copies et la majorité est présente en moins de dix copies par cellule. Au total, jusqu'à 500 000 molécules d'ARNm peuvent exister dans une seule cellule humaine. Ils sont généralement les produits d'environ 25 à 50 % de gènes humains exprimés dans la plupart des tissus et des types de cellules. Ce n'est que dans le tissu cérébral que le nombre de gènes exprimés est beaucoup plus élevé. Les chiffres ci-dessus donnent une idée de la complexité du transcriptome et montrent que sa caractérisation presque complète, comme pour la séquence du génome humain, peut-être une tâche énorme.

4.4.1. **Séquences uniques**

Bien qu'à l'origine défini comme une unité fonctionnelle de l'hérédité, un gène peut être défini comme un segment exprimé d'ADN contenant des séquences régulatrices de la transcription. Il existe 26 588 transcrits codant pour des protéines avec 12 000 gènes supplémentaires dérivés par ordinateur. Ceci est cohérent avec l'estimation de 30 000 à 40 000 du consortium international de séquençage du génome humain. De plus, plus de 700 gènes d'ARN non codants ont été identifiés avec plus de 5000 gènes apparentés, dont la plupart sont des pseudogènes. La proportion du génome constituée de gènes serait estimée à **15-20%.**

Cependant, environ 90% de l'ADN des gènes codant pour les protéines sont non codants, y compris les séquences régulatrices en amont et en aval et les introns. Par conséquent, seuls 1,5 à 2% (45 à 60 Mb) de l'ADN ont une fonction de codage. Les séquences régulatrices comprennent des promoteurs communs reconnus par des facteurs de transcription situés à des distances spécifiques en amont du site de début de transcription. Ces séquences incluent les boîtes TATA, CCAATT et GC. Il existe également des séquences promotrices spécifiques aux tissus. Les activateurs et les silencieux sont des éléments agissant en cis qui fonctionnent dans

diverses orientations et emplacements à l'intérieur ou à proximité d'un gène et qui régulent à la hausse et à la baisse l'expression des gènes, respectivement. De nombreuses séquences codantes peuvent être incluses parmi les membres de familles de gènes comme décrit ci-dessous. De plus, il peut y avoir des séquences à copie unique dans l'ADN de l'espaceur sans fonction connue (déterminable).

Plus de 50 % du génome eucaryote est constitué d'ADN dont la séquence est unique et le génome humain code pour environ **100 000 protéines**. Les parties codantes moyennes d'un gène (les exons) consistent en environ 2 000 paires de bases d'ADN dont la séquence est unique. Ce nombre représente moins de 7 % de l'ADN total composant le génome humain et moins de 14 % de cet ADN est unique. La plupart des séquences codantes sont interrompues par de 1 à 50 séquences non codantes ou introns. La longueur totale des introns qui interrompent un gène dépasse généralement de loin la longueur totale des exons. Étant donné que les séquences qui régulent l'expression des gènes représentent également certaines des séquences uniques, la quantité réelle d'ADN codant pour des produits de gènes fonctionnels est probablement inférieure à 3 % de l'ADN génomique total.

4.4.2. **Séquences répétitives en tandem**

Les mammifères ont environ 3 milliards de paires de bases par génome haploïde abritant environ 20 000 à 25 000 gènes. Une partie mineure du génome (5-10 %) est codante pour les séquences, et la partie restante est non codante représentant l'ADN répétitif. La comparaison de la taille du génome de différents eucaryotes montre que la quantité d'ADN non codant est très variable et constitue de 30 % à environ 99 % du génome total. Les séquences répétitives non codantes sont des éléments dynamiques, qui remodèlent le génome de leur hôte en générant des réarrangements, un brassage des gènes et en modulant le schéma d'expression. Ce dynamisme des répétitions conduit à une divergence évolutive qui peut être utilisée dans l'identification des espèces, l'inférence phylogénétique et pour étudier le processus de mutations sporadiques et la sélection naturelle. Ces séquences répétitives sont principalement composées de répétitions entrecoupées et en tandem. Ce dernier comprend les satellites, les minisatellites et les microsatellites. Les ADN satellites sont principalement associés à l'hétérochromatine centromérique et celle-ci est de plus en plus utilisée comme outil polyvalent pour l'analyse du génome, la cartographie génétique et pour comprendre l'organisation chromosomique. D'autre part, les minisatellites et les microsatellites sont dispersés dans tout le génome et sont hautement polymorphes dans toutes les populations étudiées. Cet arrangement a conduit à leur utilisation intensive en tant que marqueurs génétiques pour la prise d'empreintes digitales, le génotypage

et l'analyse médico-légale dans le système humain. Sur la base de leurs arrangements, les séquences d'ADN répétitives sont classées en deux types (**Figure 1**).

2.4.2.1 Macrosatellites

Les macrosatellites sont de très longs réseaux, jusqu'à des centaines de kilobases, d'ADN répétés en tandem. Les trois bandes satellites observées par centrifugation à densité flottante représentent des sections du génome humain contenant de l'ADN hautement répété qui altère en fait la proportion de nucléotides du reste du génome. Cependant, toutes les séquences satellites ne sont pas résolues par centrifugation en gradient de densité. L'ADN satellite alpha ou l'ADN alphoïde constitue la majeure partie de l'hétérochromatine centromérique sur tous les chromosomes.

2.4.2.2. Minisatellites

Les minisatellites sont des séquences d'ADN répétées en tandem, donnant une longueur totale de moins de 1 kbp à 15 kbp. Un sous-ensemble de minisatellites comprend les réseaux hautement polymorphes de courtes répétitions en tandem sans fonction connue qui servent de marqueurs d'ADN utiles appelés répétitions en tandem à nombre variable (VNTR). Ces séquences contiennent généralement 1 à 5 kpb d'ADN d'unités répétitives de 15 à 100 nucléotides. Plusieurs minisatellites partagent suffisamment d'homologie de séquence pour être analysés par une seule sonde produisant des empreintes ADN. Un exemple est une séquence centrale de 10 à 15 pb de minisatellites de myoglobine, qui comprend une séquence centrale presque invariante (GGGCAGGANG) parmi plusieurs locus VNTR polymorphes.

Les séquences d'ADN télomériques contiennent un autre sous-ensemble de minisatellites. Les séquences télomériques contiennent 10 à 15 kb de répétitions d'hexanucléotides, le plus souvent TTAGGG dans le génome humain, aux extrémités des chromosomes. Ces séquences sont ajoutées par la télomérase pour assurer la réplication complète du chromosome. Les télomères des cellules somatiques sont généralement plus courts que dans les cellules germinales, illustré par leur taille décroissante au sein des cellules B humaines et des cellules cutanées avec l'âge. Chez l'homme, il a été postulé que la perte de télomères est associée au vieillissement et à la tumorigenèse.

2.4.2.2. Microsatéllites

Les microsatellites sont de petits réseaux de courtes répétitions en tandem simples, principalement de 4 pb ou moins. Différents réseaux se trouvent dispersés dans tout le génome,

bien que les répétitions dinucléotidiques CA / TG soient les plus courantes, produisant 0,5% du génome. Séries de As et Ts sont également courants. Les microsatellites n'ont pas de fonctions connues. Cependant, les paires de dinucléotides CA/TG peuvent former la conformation de l'ADN-Z in vitro, ce qui est peut-être indicatif de la fonction. La variation du nombre de copies d'unités répétées des microsatellites se produit apparemment par un glissement de réplication produisant des marqueurs d'ADN hautement polymorphes appelés polymorphismes répétés en tandem courts (STRP). Les STRP sont couramment utilisés dans les kits commerciaux d'empreintes génétiques. L'expansion des répétitions de trinucléotides dans les gènes a été associée à des troubles génétiques tels que la maladie de Huntington, la dystrophie musculaire myotonique, l'ataxie de Friedreich et le syndrome de l'X fragile.

2.4.2.3. Familles de gènes

Les familles de gènes consistent généralement en un ensemble de gènes présentant une homologie de séquence élevée sur toute leur longueur, principalement dans les exons des familles de gènes codant pour des protéines. Les membres de familles de gènes, ou éventuellement des groupes séparés de la même famille de gènes, sont considérés comme paralogues et sont dérivés d'un gène ou d'un locus ancestral par duplication, et sont donc liés sur le plan évolutif et fonctionnel. La duplication d'un gène, cependant, peut produire un pseudogène non fonctionnel. De plus, il existe des gènes produisant des produits avec une faible homologie de séquence globale, mais qui sont homologues au niveau de domaines conservés fonctionnellement ou de motifs d'acides aminés courts, formant collectivement un type supplémentaire de famille de gènes. Un groupe de gènes avec des produits fonctionnellement et structurellement apparentés avec de faibles homologies de séquence et dépourvus de motifs d'acides aminés conservés peut être appelé une superfamille de gènes.

2.4.2.3.1. Familles de gènes avec des produits essentiellement identiques

Si la cellule justifie de nombreuses protéines ou molécules d'ARN, une solution pourrait être la production de multiples copies fonctionnelles d'un gène. Le génome humain, et les génomes eucaryotes en général, ont amplifié un certain nombre de gènes dont les produits sont responsables de fonctions générales telles que la réplication de l'ADN et la synthèse des protéines.

Les gènes des histones sont hautement conservés chez les eucaryotes et jouent un rôle fondamental dans la structure de la chromatine. La famille des histones se compose de cinq

gènes qui ont tendance à être liés, bien que dans des tableaux différents de nombres de copies variables dispersés dans le génome humain.

Les gènes individuels d'une famille d'histones particulière codent pour des produits essentiellement identiques (c'est-à-dire que les gènes H4 produisent la même protéine H4). L'analyse de clones génomiques humains individuels a identifié des gènes d'histones isolés (par exemple H4), des grappes de deux ou plusieurs gènes d'histones ou des grappes de tous les gènes d'histones (par exemple H3-H4-H1-H3-H2A-H2B) (**Figure 2**).

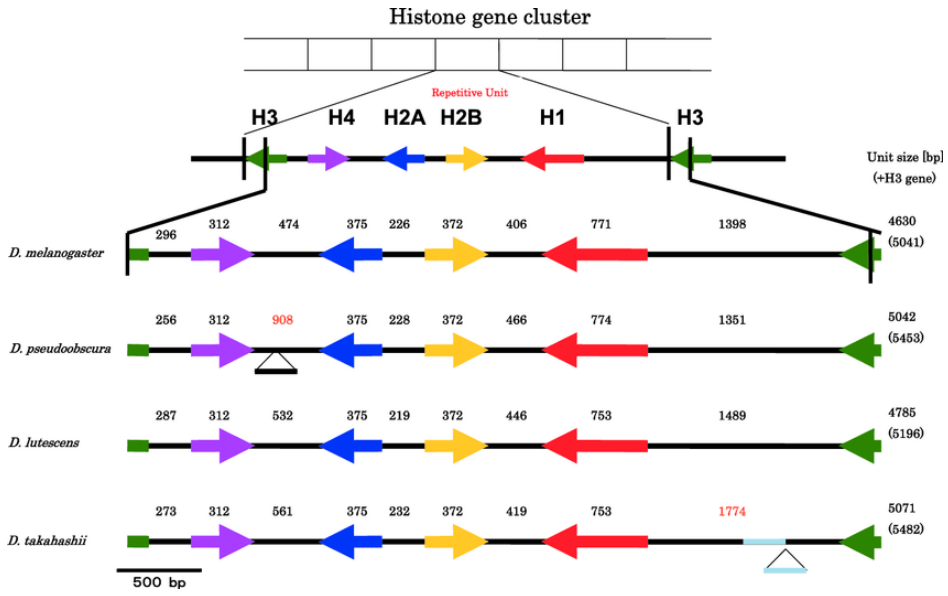


Figure 2. Organisation de l'unité répétitive du gène de l'histone chez *D. melanogaster*, *D. pseudoobscura, D. lutescens* et *D. takahashii*. Les nombres indiquent la taille (pb) des régions ou des unités. Les flèches indiquent le sens de la transcription. L'insertion trouvée dans l'espaceur entre H1 et H3 chez *D. takahashii* est répétitive (couleur bleu clair) (12).

La majorité des gènes d'histones forment un grand groupe sur le chromosome humain 6 (6p21.3) et un petit groupe en 1q21. De plus, les gènes d'histone manquent d'introns; une caractéristique rare pour les gènes eucaryotes. Les gènes qui codent pour l'ARN ribosomique (ARNr), y compris les unités d'espacement, totalisent environ 0,4 % de l'ADN du génome humain. Les gènes individuels d'une famille d'ARNr particulière sont essentiellement identiques. Les gènes d'ARNr 28S, 5.8S et 18S sont regroupés avec des unités d'espacement (ETS (espaceur transcrit externe), ITS (espaceur transcrit interne)), dans des réseaux en tandem d'environ 60 copies chacun produisant environ 2 Mbp d'ADN. Ces amas sont présents sur les bras courts de cinq chromosomes acrocentriques et forment les régions organisatrices nucléolaires, soit environ 300 copies. Ces trois gènes d'ARNr sont transcrits en une seule unité

(donnant de l'ARNr 41S) puis clivés. Les gènes de l'ARNr 5S sont regroupés sur le chromosome 1q. Il existe environ 30 gènes d'ARN de transfert humain (ARNt). Les gènes d'ARNt et leurs pseudogènes sont dispersés sur au moins sept chromosomes. De plus, des gènes d'ARNt ont été trouvés dans divers clusters, c'est-à-dire des gènes génomiques clonés des fragments contenant plusieurs gènes d'ARNt ont été isolés. La dispersion des pseudogènes d'ARNt peut s'être produite par rétroposition médiée par l'ARN. Ceci est cohérent avec l'hypothèse selon laquelle diverses familles SINE ont été dérivées de gènes d'ARNt.

On pense que les petites molécules d'ARN nucléaire (ARNsn) fonctionnent dans le traitement de l'ARN. Il existe six familles de gènes snRNA apparentés, appelés U1 à U6, qui sont dispersés parmi les chromosomes. Cependant, différents modèles de cluster ont été observés pour ces gènes sur différents chromosomes. Par exemple, 35 à 100 gènes U1 fonctionnels, partageant tous 20 kb de séquences flanquantes presque identiques en 5 'et 3', sont regroupés de manière lâche dans le chromosome 1p36 et contiennent plus de 44 kb de séquences intergéniques, tandis que 10 à 20 gènes U2 sont regroupés dans une unité de répétition serrée et pratiquement parfaite de 6 ko sur 17q21–q22

De plus, plus d'une sous-famille d'un snRNA U a été identifiée ; U3 comprend au moins deux sous-familles, qui diffèrent par les séquences flanquantes. De plus, des pseudogènes d'ARNsn ont été identifiés et on pense qu'ils sont dispersés dans le génome par rétroposition. Les gènes d'ARNt sont également trouvés regroupés avec des gènes d'ARN U ; par exemple, la bande chromosomique 1p36 contient 15 à 30 copies de chacun des gènes U1, Glu-ARNt et Asn-ARNt.

2.4.2.3.2. Familles de gènes avec des homologies de séquence élevées

Il existe de nombreuses familles de gènes dans le génome humain partageant une importante homologie intrafamiliale. Ceux-ci sont généralement dispersés, mais peuvent contenir des membres liés. L'une des familles de gènes les plus étudiées est la famille de l'hémoglobine. L'hémoglobine humaine est une protéine tétramère constituée de deux sous-unités α-globine et deux sous-unités β-globine. Il existe plusieurs polypeptides possibles construisant la molécule d'hémoglobine avec des propriétés physiologiques et une régulation ontologique différentes. Cela s'est probablement produit à la suite d'une duplication de gènes permettant une divergence de séquences pour procurer une nouvelle fonction. Les deux familles de globine existent sous forme de grappes de gènes et de pseudogènes sur des chromosomes séparés. Le groupe de gènes α-globine est sur le chromosome humain 16 et le groupe β-globine est sur 11. Bien que liés par

la séquence, il existe une plus grande homologie intra-que intercluster. Par conséquent, les duplications intracluster sont postérieures à la duplication du gène ancestral produisant l'a- et la β-globine. La régulation ontologique est apparemment coordonnée sur chaque cluster par des séquences en amont, assurant l'expression du gène le mieux adapté au besoin en oxygène (un fœtus, par exemple, existe dans un environnement relativement hypoxique). La divergence antérieure à l'hémoglobine est la divergence de l'hémoglobine et de la myoglobine à partir d'un gène ancestral. La myoglobine est une protéine monomérique codée par un seul gène sur le chromosome humain 22 et stocke l'oxygène dans les muscles, tandis que l'hémoglobine est le transporteur d'oxygène dans le sang. Les proto-oncogènes sont également des membres de la famille des gènes. Ces gènes contribuent à la néoplasie lorsque leur expression (séquence ou niveau) est altérée. Les produits géniques ont des fonctions cellulaires normales telles que les facteurs de croissance sécrétés (par exemple, la famille des gènes Wnt), les récepteurs de surface cellulaire (par exemple, la famille des gènes erbB), les transducteurs de signal intracellulaire (par exemple, la famille des gènes ras) et les protéines de liaison à l'ADN (par exemple, la famille des gènes myc).

La famille de gènes Wnt se compose d'au moins 15 gènes structurellement apparentés fonctionnant dans divers aspects de la croissance et de la différenciation. Ils contiennent un peptide signal sécrétoire N-terminal, un domaine court de faible conservation de séquence et un bloc hautement conservé (allant de 40 à 95%) d'environ 300 acides aminés avec des motifs hautement conservés, et conservation de l'espacement de 22 résidus de cystéine. Les gènes Wnt correspondent à différents chromosomes, certains démontrant la conservation de la synténie.

Il existe quatre gènes erbB. Ce sont des récepteurs du facteur de croissance épidermique (EGFR) regroupés, comme les autres récepteurs tyrosine kinases, dans une famille basée sur l'homologie de séquence de leurs domaines kinases, leur structure et la similarité structurale de leurs ligands. Les gènes myc font partie de la famille de base des facteurs de transcription hélice-boucle-hélice. Les membres fonctionnels, y compris c-myc, L-myc et N-myc, ne sont pas liés génétiquement, les deux derniers démontrant des modèles d'expression plus restreints, mais ils partagent une structure à trois exons et deux introns. Une analyse détaillée de la séquence des gènes myc suggère que le progéniteur des gènes N-myc et L-myc était un gène c-myc dupliqué. Les gènes ras représentent une sous-famille de protéines de liaison à la guanosine triphosphate (GTP), dispersées dans le génome. N-ras, H-ras et K-ras sont des gènes étroitement apparentés codant pour un produit p21ras. Les autres membres de cette famille comprennent TC21 et R-ras.

2.4.2.3.3. Familles de gènes avec une faible homologie de séquence mais des domaines fonctionnellement conservés

Certaines séquences du génome humain partagent des domaines d'acides aminés hautement conservés avec de faibles homologies globales. Ceux-ci ont souvent une fonction de développement. Il existe neuf gènes de boîte appariée dispersés (Pax) qui contiennent des domaines de liaison à l'ADN hautement conservés avec six hélices α. Les gènes homeobox ou Hox partagent une séquence codante commune de 60 acides aminés. Chez l'homme, il existe quatre groupes de gènes Hox, sur différents chromosomes. Cependant, les gènes individuels du cluster présentent une plus grande homologie avec un gène homologue sur un autre cluster qu'avec les autres gènes du même cluster.

2.4.2.3.4. Familles de gènes avec des produits différents mais des motifs d'acides aminés courts conservés

Certains gènes sont considérés comme des familles basées non pas sur l'homologie de séquence de longueur entière, mais sur des motifs d'acides aminés courts conservés. Les gènes de la boîte DEAD codent pour des produits avec une activité d'hélicase à ARN et partagent huit motifs d'acides aminés courts, y compris la boîte DEAD (Asp-Glu-Ala-Asp). Cependant, il existe d'autres familles de gènes avec des motifs d'acides aminés, tels que la boîte WD, qui assurent différentes fonctions. Les gènes de la boîte WD sont caractérisés par entre quatre et huit répétitions en tandem d'une séquence centrale de longueur fixe se terminant par un dipeptide WD.

2.4.3. Superfamilles de gènes

Les séquences d'ADN qui donnent des produits fonctionnellement et structurellement apparentés avec une faible homologie de séquence et dépourvus de motifs d'acides aminés significativement conservés peuvent être regroupées en une superfamille de gènes. Cependant, différentes familles de gènes peuvent comprendre une superfamille. Les gènes de la superfamille des immunoglobulines codent pour des protéines qui forment des dimères constitués de domaines variables extracellulaires à l'extrémité N-terminale et de domaines constants à l'extrémité C-terminale. Les membres de la superfamille des immunoglobulines comprennent l'immunoglobuline, l'antigène leucocytaire humain (HLA), le récepteur des lymphocytes T (TCR), les gènes T4 et T8. Un autre exemple comprend trois superfamilles de récepteurs de facteurs de croissance :

1. Protéines avec une structure centrale de sept séquences α-hélicoïdales transmembranaires ;
2. Grandes glycoprotéines possédant généralement une seule séquence transmembranaire et une activité tyrosine kinase (comprend la famille EGFR/erbB décrite ci-dessus);
3. Protéines transmembranaires uniques dépourvues d'activité kinase.

2.4.4. Éléments transposables

Le génome humain contient des séquences répétées intercalées dont le nombre de copies a été largement amplifié par le mouvement dans tout le génome. Ces séquences sont appelées éléments transposables. Presque toute la transposition s'est produite via un ARN intermédiaire produisant des classes de séquences appelées rétrotransposons ou rétroposons (**Figure 3**). Cependant, il existe également des preuves d'une ancienne famille de transposons à médiation par l'ADN (pogo) dans le génome humain. Les éléments d'ADN intercalés courts et longs (respectivement SINE et LINE) sont les principales familles d'éléments transposables dans le génome humain. Ceux-ci sont appelés rétroposons, car ils n'ont pas les longues répétitions terminales (LTR) des rétrovirus et sont amplifiés via un ARN intermédiaire. Les LINE sont parfois appelées rétrotransposons non LTR car les séquences dans les éléments codent pour les enzymes utilisées dans le processus de rétroposition.

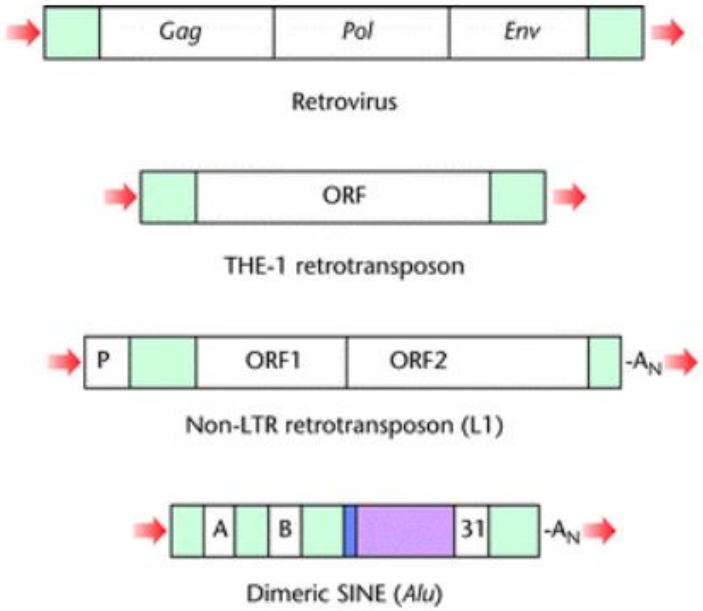


Figure 3. Éléments transposables médiés par l'ARN dans le génome humain. Chacun contient les répétitions directes latérales caractéristiques (flèches). Le rétrovirus endogène humain contenant de longues répétitions terminales (LTR) (régions vert pâle), gag (gène d'antigène spécifique de groupe), pol (gène de polymérase) et env (gène d'enveloppe). Le

rétrotransposon THE-1 se compose d'un cadre de lecture ouvert (ORF) et de LTR. Le rétrotransposon non-LTR (L1) contient des séquences internes du promoteur de l'ARN polymérase II (P), deux cadres de lecture ouverts et une queue poly(A). L'élément Alu a une structure dimère de moitiés homologues séparées par une région médiane riche en A (bleu). La moitié gauche contient les séquences promotrices de l'ARN polymérase III A- et B-box, et la moitié droite contient 31 pb internes supplémentaires. Pour les éléments L1 et Alu, les régions vert pâle et mauve sont des séquences uniques à ces éléments (12).

L'élément Alu est estimé à plus d'un million de copies dans le génome humain représentant la famille SINE primaire. Les comparaisons de séquences suggèrent que les répétitions Alu sont dérivées du gène ARN 7SL. Chaque élément Alu est d'environ 280 bp avec une structure dimère, contient le promoteur de l'ARN polymérase III, et a généralement une queue riche en A et des répétitions directes flanquantes (générées lors de l'intégration). Environ 5000 éléments Alu se sont intégrés dans le génome humain à la suite de la divergence entre les humains et les grands singes. Environ 25% des intégrations Alu les plus récentes ont produit des polymorphismes d'insertion présence/absence qui sont utiles comme marqueurs d'ADN pour l'étude de la médecine légale et de la génétique des populations humaines. Les récentes intégrations germinales d'éléments Alu ont donné lieu à des phénotypes pathogènes. Les éléments en aluminium prédominent dans des bandes R chromosomiques et s'insèrent préférentiellement dans des séquences riches en A, y compris les Atails des intégrations Alu précédentes. Les éléments Alu et les LINE semblent s'amplifier par l'activité de quelques loci maîtres, laissant la grande majorité de ces répétitions comme inactives pseudogènes.

Les LINE (ou éléments L1) sont estimés à plus de 500 000 copies et se trouvent principalement dans les bandes G chromosomiques. Une LIGNE pleine longueur est d'environ 6,1 kpb, bien que la plupart soient des pseudogènes tronqués (voir ci-dessous) avec diverses extrémités 5' en raison d'une transcription inverse incomplète. Environ 1 à 2% des 3500 LINE pleine longueur estimées ont des séquences fonctionnelles du promoteur de l'ARN polymérase II ainsi que deux cadres de lecture ouverts intacts nécessaires pour générer de nouvelles copies L1. Les LINE individuelles contiennent une queue poly(A) et sont flanquées de répétitions directes. L'activité de mobilisation de LINE a été vérifiée dans les tissus germinaux et somatiques.

Les SINE et les LINE contribuent en outre à l'évolution du génome en produisant des sites pour une recombinaison homologue inégale. Dans le seul gène du récepteur des lipoprotéines de basse densité (LDL), plusieurs altérations ont été attribuées à la recombinaison à divers sites Alu, entraînant une hypercholestérolémie familiale. Le génome humain contient également des familles de séquences apparentées aux rétrovirus. Ceux-ci sont caractérisé par des séquences

codant pour des enzymes pour la rétroposition et contenant des LTR. Cependant, la plupart de ces séquences sont des éléments de type rétrovirus tronqués et mutés disparus.

Les rétrovirus endogènes peuvent avoir été initialement incorporés dans le génome suite à une infection rétrovirale des cellules germinales. De plus, des LTR solitaires de ces éléments peuvent être localisés dans tout le génome. Il existe plusieurs familles de rétrovirus endogènes humains (HERV) peu abondants (10 à 1000 copies), avec des éléments individuels allant de 6 à 10 kb. Dans l'ensemble, les éléments LTR englobent environ 8% du génome.

L'élément humain de type transposon (THE-1) contient les longues répétitions terminales de génomes rétroviraux intégrés, mais manque de séquences codant pour les enzymes impliquées dans la rétroposition. Par conséquent, cette famille d'ADN intercalée est provisoirement caractérisée comme un rétrotransposon. La séquence THE-1 de 2,3 kb est estimée à 10 000 copies avec 10 000 LTR solitaires supplémentaires.

Il existe des pseudogènes qui sont le résultat d'une rétroposition (rétropseudogènes). Ces pseudogènes manquent d'introns et des séquences d'ADN flanquantes du locus fonctionnel et ne sont donc pas des produits de duplication de gène. La génération de ces types d'éléments dépend de la transcriptase inverse d'autres éléments tels que les LINE.

Les séquences répétitives à fréquence de réitération moyenne (MER) représentent un large groupe de familles de séquences intercalées non caractérisées. Les mécanismes d'amplification de ces séquences ne sont pas clairs et peuvent donc justifier ou non leur inclusion en tant qu'élément transposable. Cependant, il a été suggéré que certaines de ces séquences sont répliquées et disséminées par des virus à ADN, indiqué par la présence de séquences MER dans les recombinants SV40. Le nombre de copies des familles MER varie de 200 à 10 000, produisant collectivement 100 000 à 200 000 copies.

2.4.5. Pseudogènes

Les pseudogènes sont des copies non fonctionnelles d'un gène contenant une partie ou la totalité de la séquence d'origine. Les pseudogènes peuvent survenir par duplications en tandem, accumulant des mutations en raison de l'absence de pression de sélection, et sont généralement reconnaissables par l'absence d'un cadre de lecture ouvert. Un exemple est les pseudogènes de la globine. Un pseudogène ou rétropseudogène traité est dérivé d'un ARN intermédiaire. Les traits caractéristiques de ces séquences sont qu'elles manquent de séquences régulatrices, et sont donc normalement incapables d'expression, et elles manquent d'introns qui ont été épissés

pendant le traitement de l'ARN. Il peut y avoir jusqu'à 20 à 30 rétropseudogènes issus d'un gène fonctionnel parental, par ex. la protéine ribosomique L32 et la glycéraldéhyde-3-phosphate déshydrogénase. Cependant, les SINE et les LINE représentent les familles les plus abondantes de rétropseudogènes. Les pseudogènes traités peuvent également être dérivés de gènes d'ARN. Les preuves des pseudogènes «rétro» de l'ARNt sont les séquences CCA à l'extrémité 3 'qui sont ajoutées après la transcription à l'ARNt fonctionnel.

3. Analyse des génomes

3.4. Exemple d'analyse de génome de modèle procaryote : les bactéries

En ce moment même, une technologie prometteuse pour obtenir plus d'informations sur les maladies humaines est le système CRISPR-Cas9 (Clustered Regularly Interspaced Short Polindromic Repeats/ CRISPR-associated), qui s'est avéré être un système immunitaire procaryote contre les virus. Ce système consiste en un petit groupe de gènes cas (codant pour des protéines associées à CRISPR) et une séquence d'ADN spécifique, appelée locus CRISPR, qui comprend de courtes répétitions séparées par des espaceurs uniques. Lors d'une infection virale, son espaceur unique s'intègre dans le locus bactérien CRISPR. Par la suite, ce locus est transcrit en ARN CRISPR précurseur (pré-ARNc). Après le traitement, l'ARNc mature peut reconnaître et détruire l'acide nucléique cible en interagissant avec les protéines Cas. Ainsi, le locus CRISPR contient des informations sur les infections virales antérieures, permettant ainsi aux bactéries de reconnaître et d'inactiver le virus en cas de réinfection. Actuellement, certaines études scientifiques montrent qu'il est possible de modifier les composants protéiques et ARN du système bactérien CRISPR afin de reconnaître et de couper l'ADN au locus souhaité. En raison de ces propriétés, il est possible d'appliquer ce système in vitro dans la lignée cellulaire humaine, afin d'étudier les maladies humaines sans conséquences négatives.

3.5. Exemple d'analyse de génome de modèles eucaryotes

Bien que les génomes humains soient identiques à environ 99,9%, les 0,1% restants sont la raison de la différence entre les personnes causée par différentes variantes. Depuis 2003, la séquence complète du génome humain, son annotation et l'avancement accru des technologies de séquençage. Bien que la technique de détection des variants devienne maintenant une routine, la question clé depuis de nombreuses années concerne la fonction des variants détectés. La source d'informations importantes sur la génomique fonctionnelle sont plusieurs projets à grande échelle, par exemple, le projet ENCODE, dont l'objectif principal était d'identifier tous les éléments fonctionnels, y compris les éléments régulateurs dans les régions codantes et non

codantes. Selon un autre, le 1000 Genomes Project, il existe environ 20 000 à 23 000 variantes dans les régions synonymes et non synonymes du génome humain. Même si tous ne sont pas fonctionnellement significatifs, 530 à 610 des variants ont un impact fonctionnel en provoquant des suppressions et des insertions dans le cadre, des codons d'arrêt prématurés, des décalages de cadre ou en perturbant les sites d'épissage. Malgré de nombreuses études, les scientifiques sont toujours confrontés à un énorme défi pour démêler la signification de la séquence et décider si une variante trouvée est pathogène ou non. Une variante pathogène peut entraîner une maladie ou provoquer un certain nombre de troubles. Cependant, la compréhension des mécanismes pathogènes crée une opportunité de prévenir des conséquences graves en développant de nouveaux outils de diagnostic et en concevant des traitements très efficaces pour la maladie. Pour atteindre cet objectif, il est nécessaire d'effectuer une analyse fonctionnelle du génome à grande échelle qui implique différents domaines d'étude : génomique, épigénomique, transcriptomique, protéomique et interactomique.

Les modèles animaux ont longtemps été appliqués dans différentes études pour l'étude des mécanismes biologiques et pathogènes, ainsi que pour le développement de traitements efficaces. Selon le but de l'étude, différents modèles animaux peuvent être utilisés, bien que la souris (*Mus musculus*), la mouche des fruits (*Drosophila melanogaster*) et le poisson zèbre (*Danio rerio*) soient les plus couramment utilisés dans les études de génome fonctionnel. Cette étude présente de nombreux avantages. Par exemple, la mutation peut être induite artificiellement et le phénotype mutant peut être reconnu facilement, les gènes peuvent être clonés en utilisant des procédures standard, l'animal produit un grand nombre de descendants dans un laps de temps relativement court. Il existe deux stratégies principales d'utilisation du modèle animal : "knock out" - suppression du gène d'intérêt, "knock in" - incorporation de la même mutation que celle observée chez l'homme. Par exemple, un certain nombre d'études ont été menées pour créer des modèles animaux de maladies humaines par mutagenèse chimique (par exemple, N-éthyl nitrosourée ; ENU) qui provoque des mutations ponctuelles alléliques aléatoires chez la souris. Cependant, la principale limitation des modèles animaux est le phénotype qui souvent ne reflète pas les êtres humains.

3.6. Techniques d’analyse du génome

Bien que les chercheurs puissent facilement planifier leur essai dans le cas de variants particuliers, ils sont confrontés à certains défis dans l'étude de variants non spécifiés. Le séquençage est considéré comme la méthode "de référence" pour l'identification de variants connus et non spécifiés dans l'ADN génomique. Conformément à l'énoncé précédent, les

techniques de Sanger ou Next-Generation Sequencing (NGS) peuvent être utilisées. Le concept derrière ces deux méthodes est similaire. Au cours de la réaction en chaîne par polymérase, qui consiste en plusieurs cycles de réplication séquentielle de l'ADN, l'ADN polymérase catalyse l'incorporation complémentaire de désoxyribonucléoside 5'-triphosphates (dNTP) marqués par fluorescence dans la matrice d'ADN. Pour chaque cycle, une couleur du fragment d'ADN marqué est enregistrée par un détecteur, déterminant ainsi le nucléotide dans la séquence. La principale différence entre la technologie conventionnelle (c. Ces deux méthodes de séquençage sont largement utilisées dans le monde entier. Même ainsi, on considère que dans un projet à petite échelle, il est plus éligible d'utiliser le système de séquençage Sanger en raison de sa précision. D'autre part, dans les projets à grande échelle, cette méthode de recherche serait coûteuse et prendrait du temps, c'est pourquoi le NGS doit être appliqué. Les progrès technologiques généraux réalisés dans certaines stratégies de séquençage d'ADN de nouvelle génération ont un impact énorme sur la recherche génétique.

Afin de comprendre la structure, la fonction ou l'évolution du génome, il ne suffit pas d'obtenir les données de séquençage de l'ADN par le NGS : mais il faut aussi une analyse approfondie et précise à l'aide d'approches bioinformatiques. La clé d'une analyse de séquence réussie consiste à aligner la séquence d'intérêt avec une autre séquence dont la fonction est connue (généralement appelée génome de référence). Cela peut être très utile lorsque la fonction du gène est inconnue mais qu'elle est liée à l'évolution d'un autre gène dont la fonction est définie. Dans un tel cas, on peut soupçonner que le gène inconnu a la même fonction ou une fonction similaire. De plus, les séquences peuvent être scannées afin de trouver les correspondances significatives entre les composants d'une séquence qui ont été précédemment décrites comme ayant un impact énorme sur la fonction génomique. Afin de comparer les données, il est nécessaire de rechercher des informations dans différentes bases de données biomédicales. L'une des plus grandes sources d'informations biomédicales et génomiques est le NCBI (National Center for Biotechnology Information), qui donne accès à d'autres bases de données telles que PubMed, Entrez Gene, OMIM, Variation Viewer, dbSNP et autres.

3.3.1. Epigénomique

Pour l'analyse fonctionnelle, il est important de prendre en compte les modifications épigénétiques telles que la méthylation de l'ADN et les modifications des histones, car elles affectent l'expression des gènes sans aucune modification de la séquence d'ADN sous-jacente. La méthylation de l'ADN, qui se produit généralement dans le contexte de dinucléotides CpG densément situés (c'est-à-dire des îlots CpG), est en corrélation avec la suppression de la

transcription. Afin de détecter l'état de méthylation de l'ADN, les cytosines non méthylées sont converties en uracile en utilisant du bisulfite de sodium, car la cytosine méthylée résiste à cet impact. De plus, les enzymes de restriction dépendantes de la méthylation (MDRE) sont très efficaces pour l'analyse de la méthylation de l'ADN. Ces enzymes, e. g., HpaII et MspI, reconnaissent et digèrent simplement l'ADN méthylé. Habituellement, la MDRE ou même la conversion au bisulfite plus fréquemment utilisée est la première étape de nombreuses méthodes ultérieures telles que la PCR spécifique à la méthylation, le séquençage, le réseau de billes, etc.

3.6.2. Transcriptomique

Lorsque le génome humain a été entièrement séquencé, l'attention s'est portée sur l'identification et l'annotation de ses éléments d'ADN fonctionnels, y compris ceux qui régulent l'expression des gènes. L'identification de tels éléments est une étape d'une importance vitale pour élucider les voies pathogènes qui affectent la santé humaine.

Tous les processus au niveau de l'ARN, y compris l'activation ou l'inhibition de la transcription, le traitement de l'ARNm et son transport, sont régulés par différents éléments fonctionnels de l'ADN génomique. Néanmoins, la régulation la plus élevée se produit au niveau de l'initiation de la transcription par plusieurs éléments régulateurs, appelés séquence régulatrice agissant en cis et facteurs trans. Les trans-facteurs tels que les facteurs de transcription (TF), les activateurs et les répresseurs (y compris les co-activateurs et les co-répresseurs) interagissent avec des régions d'ADN spécifiques, i. e., séquence régulatrice agissant en cis qui comprend un promoteur central (avec une boîte TATA et d'autres éléments de liaison), un promoteur proximal, un activateur, un silencieux, un isolant et une région de contrôle de locus (LCR). L'étude de ces éléments régulateurs peut être un défi pour les scientifiques en raison des difficultés à identifier la position des sites d'initiation de la transcription (TSS) et des sites de liaison des facteurs de transcription (TFBS) dans le promoteur principal. Cependant, il existe plusieurs approches expérimentales et bioinformatiques. Tout d'abord, une approche bioinformatique comparative est nécessaire pour l'étude des éléments régulateurs. Ce type de recherche est généralement basé sur la construction d'alignements entre des séquences orthologues car l'homologie de séquence fournit des preuves précieuses pour l'analyse de la fonction des gènes. Néanmoins, une compréhension plus approfondie des éléments de régulation nécessite des investigations en laboratoire. On pense que chaque TFBS pourrait être détecté par la méthode ChIP mentionnée ci-dessus. Théoriquement, en fonction de l'immunoprécipitation de la protéine cible, les promoteurs centraux, les amplificateurs, les silencieux, les isolants et les LCR pourraient être déterminés. De plus, les marqueurs

épigénétiques peuvent être utiles pour détecter les TSS dans les loci promoteurs et amplificateurs principaux, car les TSS des gènes activement transcrits sont marqués par H3K4me3 et H3K27ac, tandis que les activateurs par H3K4me1 et H3K27ac. Un autre dosage fonctionnel très fréquent de l'élément régulateur est basé sur la transgenèse d'un gène rapporteur spécifique (par exemple, le gène de la protéine fluorescente verte - GFP ou luciférase) dans la séquence régulatrice cible. Après la traduction, l'activité du gène rapporteur est mesurée, par ex. par exemple, par fluorescence de la GFP, dans le but de déterminer si la région examinée contient des éléments qui altèrent l'expression du gène rapporteur.

3.6.3. Proteomique et interactomique

Du point de vue fonctionnel, l'analyse de la protéomique et de l'interactomique est aussi vitale que l'analyse décrite précédemment de la génomique, de l'épigénomique et de la transcriptomique, car certaines études montrent que l'expression des gènes au niveau de l'ADN ou de l'ARNm est sensiblement inchangée, bien qu'elle affecte la fonction protéique et vice versa. Les protéines remplissent une vaste gamme de fonctions dans les organismes, bien que l'expression anormale des protéines qui se produit en raison de modifications post-transcriptionnelles ou d'une interaction protéique avec une autre protéine ou des acides nucléiques perturbe la fonction cellulaire.

Selon l'intention de l'expérience, il existe deux stratégies bien connues pour la quantification des protéines : les immunodosages ou les méthodes de détection sans anticorps. Le dosage immunologique, tel que le dosage immuno-enzymatique (ELISA), est une méthode largement utilisée en raison de sa grande sensibilité et de sa forte spécificité. Cependant, les chercheurs peuvent parfois être confrontés au problème lorsqu'il n'existe aucun anticorps pour la protéine d'intérêt. Dans de tels cas, la solution est des méthodes sans anticorps. Premièrement, par rapport à la méthode de séparation unidimensionnelle des protéines, l'électrophorèse bidimensionnelle sur gel (2-DE), qui sépare les protéines par deux propriétés dans les gels 2D, est plus efficace. Cependant, l'outil analytique le plus courant et le plus complet pour la détection, l'identification et la quantification des protéines est la spectrométrie de masse (MS) qui mesure le rapport masse/charge (m/z) des ions. L'avancement de la SM donne la possibilité d'atteindre un plus grand débit d'échantillons avec une précision et une exactitude élevée. De plus, on considère que la méthodologie MS est rapide et fiable pour les études à grande échelle. De plus, en raison de ses avantages, la SEP est très souvent associée à une autre technique. Par exemple, certaines études consistent en une purification à base d'anticorps et une analyse par spectrométrie de masse appelée dosage immunologique par spectrométrie de masse (MSIA).

Références

1. Atchley WR and Fitch WM (1995) Myc and Max: molecular evolution of a family of proto-oncogene products and their dimerization partner. Proceedings of the National Academy of Sciences of the USA 92: 10217–10221.
2. Avery OT, MacLeod CM, McCarty M. Studies on the chemical nature of the substance inducing transformation of Pneumococcal types. *J Exp Med*, 1944;79:137–159. pmid:19871359
3. Batzer M.A. , P.L. Deininger Nat. Rev. Genet., 3 (2002), pp. 370-379
4. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. *Nucleic Acids Res*. 2005;33:D34–D38. pmid:15608212
5. Bergstein I, Eisenberg LM, Bhalerao J et al. (1997) Isolation of two novel WNT genes, WNT14 and WNT15, one of which (WNT15) is closely linked to WNT3 on human chromosome 17q21. Genomics 46: 450–458
6. Bromham L. (2002). The Human Zoo: Endogenous Retroviruses in the Human GenomeTrends Ecol. Evol. 179197
7. Craig JM and Bickmore WA (1994) The distribution of CpG islands in mammalian chromosomes. Nature Genetics 7: 376–382.
8. Crick FHC. On protein synthesis. *Symp Soc Exp Biol*. 1958;12:138–163. pmid:13580867
9. D.L. Black. Annu. Rev. Biochem., 72 (2003), pp. 291-336
10. Dagan T. , R. Sorek, E. Sharon, G. Ast, D. Graur Nucleic Acids Res., 32 (2004), pp. D489-D492
11. Deininger PL and Batzer MA (1993) Evolution of retroposons. Evolutionary Biology 27: 157–196.
12. Elgar G. Vavouri T. 2008 Tuning in to the signals: noncoding sequence conservation in vertebrate genomes Trends Genet. 24734452
13. G. Vansant, W.F. Reynolds Proc. Natl. Acad. Sci. USA, 92 (1995), pp. 8229-8233
14. Goldman AD, Landweber LF (2016) What Is a Genome? PLoS Genet 12(7): e1006181. https://doi.org/10.1371/journal.pgen.1006181
15. Hentschel CC and Birnstiel ML (1981) The organization and expression of histone gene families. Cell 25: 301–313.
16. Hochgeschwender U. Brennan M. B.1991Identifying Genes Within The Genome: New Ways For Finding The Needle In A HaystackBioessays13139144
17. International Human Genome Sequencing Consortium (2001) Nature 409, 860–921
18. Jasinska A, G. Michlewski, M. de Mezer, K. Sobczak, P. Kozlowski, M. Napierala, W.J. Krzyzosiak Nucleic Acids Res., 31 (2003), pp. 5463-5468
19. Jasinska A., Krzyzosiak JW. (2004). Repetitive sequences that shape the human transcriptome. FEBS Letters Volume 567, Issue 1, 136-141
20. Kapitonov V. , J. Jurka. J. Mol. Evol., 42 (1996), pp. 59-65
21. Kashi Y., D. King, M. Soller Trends Genet., 13 (1997), pp. 74-78
22. Lederberg J, McCray AT. 'Ome Sweet 'Omics: A Genealogical Treasury of Words. *The Scientist*. 2001;15:8.
23. Lindgren V, Ares M Jr, Weiner AM and Francke U (1985) Human genes for U2 small nuclear RNA map to a major adenovirus 12 modification site on chromosome 17. Nature 314: 115–116.
24. Makalowski W. , G.A. Mitchell, D. Labuda Trends Genet., 10 (1994), pp. 188-193

25. McBride OW, Pirtle IL and Pirtle RM (1989) Localization of three DNA segments encompassing tRNA genes to human chromosomes 1, 5, and 16: proposed mechanism and significance of tRNA gene dispersion. Genomics 5: 561–573.
26. Michalowski S., J.W. Miller, C.R. Urbinati, M. Paliouras, M.S. Swanson, J. Griffith Nucleic Acids Res., 27 (1999), pp. 3534-3542
27. Michel JB, Shen YK, Aiden AP, Veres A, Gray MK, et al. Quantitative analysis of culture using millions of digitized books. *Science*. 2011;331:176–182. pmid:21163965
28. Napierala M., W.J. Krzyzosiak J. Biol. Chem., 272 (1997), pp. 31079-31085
29. Numata, K., Kanai, A., Saito, R., Kondo, S., Adachi, J., Wilming, L.G., Hume, D.A., Hayashizaki, Y. and Tomita, M. (2003) RIKEN GER Group; GSL Members. Genome Res. 13, 1301–1306
30. Robertson HM (1996) Members of the pogo superfamily of DNA-mediated transposons in the human genome. Molecular and General Genetics 252: 761–766.
31. Ruan Y., P. Le Ber, H.H. Ng, E.T. Liu Trends Biotechnol., 22 (2004), pp. 23-30
32. Slamovits C. H. Rossi M. S.2002Satellite DNA: Agent Of Chromosomal Evolution In Mammals. Mastozoología Neotropical 9297308
33. Sobczak K., M. de Mezer, G. Michlewski, J. Krol, W.J. Krzyzosiak Nucleic Acids Res., 31 (2003), pp. 5469-5482
34. Sobczak K., W.J. Krzyzosiak J. Biol. Chem., 277 (2002), pp. 17349-17358
35. Sorek R. , G. Ast, D. Graur Genome Res., 12 (2002), pp. 1060-1067
36. Stephens JC, Cavanaugh ML, Gradie MI, Mador ML and Kidd KK (1990) Mapping the human genome: current status. Science 250: 237–244.
37. Subramanian S., R.K. Mishra, L. Singh Genome Biol., 4 (2003), p. R13
38. Subramanian S., V.M. Madgula, R. George, R.K. Mishra, M.W. Pandit, C.S. Kumar, L. Singh Bioinformatics, 19 (2003), pp. 549-552
39. Sutton WS. On the morphology of the chromosome group in Brachystola magna. *Biol*. *Bull*. 1902;4:24–39
40. van de Lagemaat, J.R. Landry, D.L. Mager, P. Medstrand Trends Genet., 19 (2003), pp. 530-536 L.N.
41. van der Drift P, Chan A, van Roy N, Laureys G, Westerveld A, Speleman F and Versteeg R (1994) A multimegabase cluster of snRNA and tRNA genes on chromosome 1p36 harbours an adenovirus/SV40 hybrid virus integration site. Human Molecular Genetics 3: 2131–2136
42. Velculescu V.E., L. Zhang, W. Zhou, J. Vogelstein, M.A. Basrai, D.E. Bassett Jr., P. Hieter, B. Vogelstein, K.W. Kinzler Cell, 88 (1997), pp. 243-251
43. Watson JD, Crick FHC. Genetical Implications of the structure of Deoxyribonucleic Acid. *Nature*. 1953;171:964–967. pmid:13063483
44. Welcsh P.L. , M.C. King Hum. Mol. Genet., 10 (2001), pp. 705-713
45. Wren J.D., E. Forgacs, J.W. Fondon III, A. Pertsemlidis, S.Y. Cheng, T. Gallardo, R.S. Williams, R.V. Shohet, J.D. Minna, H.R. Garner Am. J. Hum. Genet., 67 (2000), pp. 345-356
46. Yelin R., D. Dahary, R. Sorek, E.Y. Levanon, O. Goldstein, A. Shoshan, A. Diber, S. Biton, Y. Tamir, R. Khosravi, S. Nemzer, E. Pinner, S. Walach, J. Bernstein, K. Savitsky, G. Rotman Nat. Biotechnol., 21 (2003), pp. 379-386
47. Yulug I.G. , A. Yulug, E.M. Fisher. Genomics, 27 (1995), pp. 544-548

Chapitre 02
Régulation de l'expression génique

Chapitre 2. Régulation de l’expression génique

6. Introduction

Chaque cellule somatique du corps contient généralement le même ADN. (Quelques exceptions incluent les globules rouges, qui ne contiennent pas d'ADN dans leur état mature, et certaines cellules du système immunitaire qui réarrangent leur ADN tout en produisant des anticorps.) En général, les gènes qui déterminent si vous avez les yeux verts ou les cheveux bruns, ou comment rapidement vous métabolisez les aliments sont les mêmes dans les cellules oculaires et les cellules hépatiques, même si ces organes fonctionnent très différemment. Si chaque cellule a le même ADN, comment se fait-il que les cellules diffèrent dans leur structure et leur fonction ? Pourquoi les cellules de l'œil diffèrent-elles si radicalement des cellules du foie ?

Bien que chaque cellule de votre corps contienne les mêmes séquences d'ADN, chaque cellule n'active pas ou n'exprime pas le même ensemble de gènes. En fait, seul un petit sous-ensemble de protéines est fabriqué par une cellule. En d'autres termes, dans une cellule donnée, tous les gènes codés dans l'ADN ne sont pas transcrits en ARNm ou traduits en protéine. Les cellules de l'œil fabriquent un certain sous-ensemble de protéines et les cellules du foie fabriquent un sous-ensemble différent de protéines. De plus, à différents moments, les cellules hépatiques peuvent fabriquer différents sous-ensembles de protéines hépatiques. L'expression de gènes spécifiques est un processus hautement régulé avec de nombreux niveaux et étapes de contrôle. Cette complexité assure l'expression de chaque protéine dans les cellules appropriées au bon moment.

7. Aspects généraux de la régulation de l’expression des gènes par les facteurs de transcription

Pour qu'une cellule fonctionne correctement, les protéines nécessaires doivent être synthétisées au bon moment. Toutes les cellules contrôlent ou régulent la synthèse des protéines à partir d'informations codées dans leur ADN. Le processus d'« activation » d'un gène pour produire de l'ARNm et des protéines s'appelle l'expression génique. Qu'il s'agisse d'un organisme unicellulaire simple ou d'un organisme multicellulaire complexe, chaque cellule contrôle le moment où ses gènes sont exprimés, la quantité de protéine fabriquée et le moment où il est temps d'arrêter de fabriquer cette protéine car elle n'est plus nécessaire.

La régulation de l'expression des gènes économise de l'énergie et de l'espace. Il est plus efficace sur le plan énergétique d'activer les gènes uniquement lorsqu'ils sont nécessaires. De plus, exprimer uniquement un sous-ensemble de gènes dans chaque cellule économise de l'espace car l'ADN doit être déroulé de sa structure étroitement enroulée pour transcrire et traduire l'ADN. Les cellules devraient être énormes si chaque protéine était exprimée dans chaque cellule tout le temps. Le contrôle de l'expression des gènes est extrêmement complexe. Les dysfonctionnements de ce processus sont préjudiciables à la cellule et peuvent conduire au développement de nombreuses maladies, dont le cancer.

Chez les eucaryotes, le contrôle de l'expression des gènes est plus complexe et peut se produire à de nombreux niveaux différents. Les gènes eucaryotes ne sont pas organisés en opérons, chaque gène doit donc être régulé indépendamment. De plus, les cellules eucaryotes possèdent beaucoup plus de gènes que les cellules procaryotes. La régulation de l'expression génique peut se produire à n'importe quelle étape lorsque l'ADN est transcrit en ARNm et que l'ARNm est traduit en protéine. Pour plus de commodité, la régulation est divisée en cinq niveaux : épigénétique, transcriptionnel, post-transcriptionnel, traductionnel et post-traductionnel (**Figure 1**).

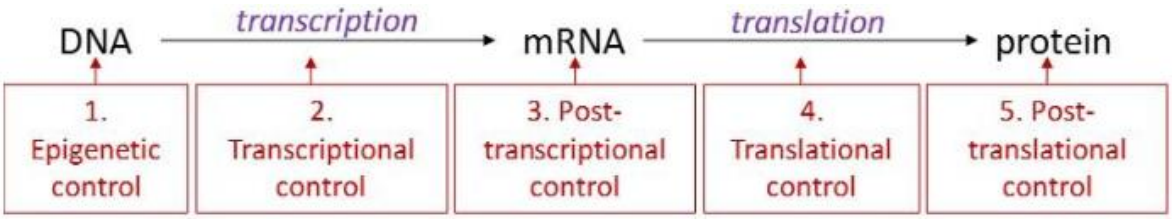


Figure 1. La régulation de l'expression des gènes chez les eucaryotes peut se produire à cinq niveaux différents. Ici, le dogme central est schématisé avec des flèches indiquant où chaque type de régulation eucaryote de l'expression génique l'interrompt (**32**).

7.1.Initiation de la transcription chez les eucaryotes

Contrairement à la polymérase procaryote qui peut se lier seule à une matrice d'ADN, les eucaryotes ont besoin de plusieurs autres protéines, appelées facteurs de transcription, pour se lier d'abord à la région promotrice, puis aider à recruter la polymérase appropriée.

7.1.1. Contrôle de l'initiation de la transcription eucaryote

La transcription des différentes classes d'ARN chez les eucaryotes est réalisée par trois polymérases différentes (voir page Synthèse d'ARN). L'ARN pol I synthétise les ARNr, sauf

pour l'espèce 5S. L'ARN pol II synthétise les ARNm et certains petits ARN nucléaires (ARNsn) impliqués dans l'épissage de l'ARN. L'ARN pol III synthétise l'ARNr 5S et les ARNt. La grande majorité des ARN eucaryotes sont soumis à un traitement post-transcriptionnel.

7.1.1.1.Les éléments promoteurs

Les contrôles les plus complexes observés dans les gènes eucaryotes sont ceux qui régulent l'expression des gènes transcrits par ARN pol II, les gènes ARNm. Presque tous les gènes d'ARNm eucaryotes contiennent une structure de base constituée d'exons codants et d'introns non codants et de promoteurs basaux de deux types et d'un nombre quelconque de domaines régulateurs transcriptionnels différent. Les éléments promoteurs basaux sont appelés boîtes CCAAT et boîtes TATA en raison de leurs motifs de séquence. La boîte TATA réside 20 à 30 bases en amont du site de départ de la transcription et est similaire en séquence à la boîte de Pribnow procaryote (consensus TATAT/AAT/A, où T/A indique que l'une ou l'autre des bases peut être trouvée à cette position).

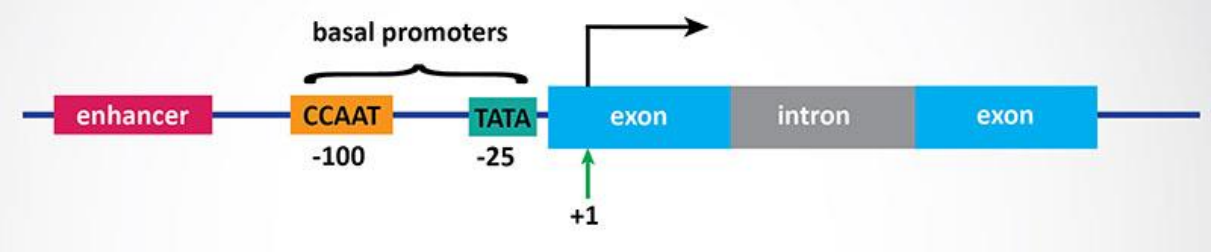


Figure 2. Structure typique d'un gène d'ARNm eucaryote. Les gènes d'ARNm eucaryotes ont la structure régulatrice générale composée des deux éléments promoteurs basaux, la boîte TATA et la boîte CCAAT. De plus, il peut y avoir un ou plusieurs éléments activateurs associés à la région régulatrice du gène (**17**).

De nombreuses protéines identifiées comme TFIIA, B, C, etc. (pour les facteurs de transcription régulant l'ARN pol II), ont été observées pour interagir avec la TATA-box. La boîte CCAAT (consensus GGT/CCAATCT) réside de 50 à 130 bases en amont du site d'initiation de la transcription. La protéine identifiée comme C/EBP (pour CCAAT-box/Enhancer Binding Protein) se lie à l'élément CCAAT-box.

Il existe également de nombreuses autres séquences régulatrices dans les gènes d'ARNm qui se lient à divers facteurs de transcription (voir le schéma ci-dessous). Ces séquences régulatrices sont majoritairement situées en amont (5') du site d'initiation de la transcription, bien que certains éléments se trouvent en aval (3') ou même à l'intérieur des gènes eux-mêmes. Le

nombre et le type d'éléments régulateurs à trouver varient avec chaque gène d'ARNm. Différentes combinaisons de facteurs de transcription peuvent également exercer des effets régulateurs différentiels sur l'initiation de la transcription. Les différents types cellulaires expriment chacun des combinaisons caractéristiques de facteurs de transcription ; c'est le principal mécanisme de spécificité de type cellulaire dans la régulation de l'expression des gènes d'ARNm.

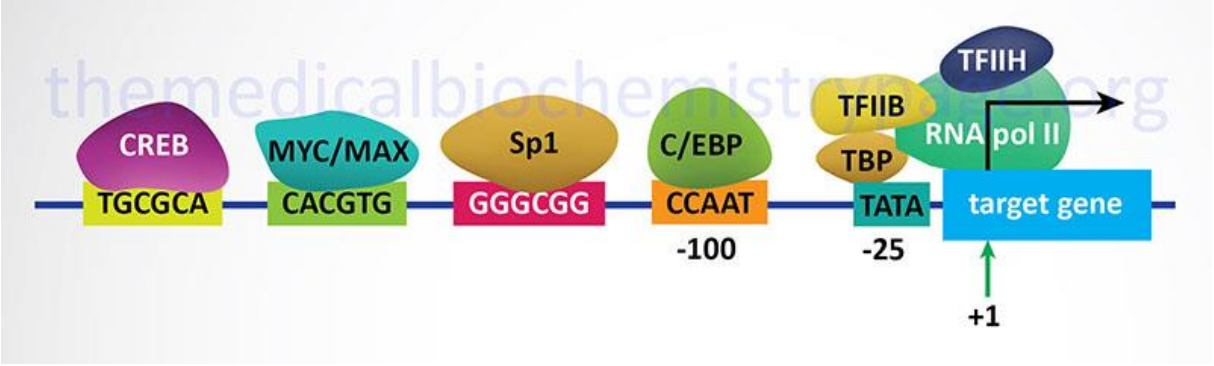


Figure 3. Structure de la région en amont d'un gène d'ARNm eucaryote typique. Le diagramme indique que les éléments basaux de la boîte TATA et de la boîte CCAAT résident près des positions nucléotidiques -25 et -100, respectivement. Il a été démontré que le facteur de transcription TFIID est la protéine de liaison à la boîte TATA, TBP. Plusieurs sites de liaison de facteurs de transcription supplémentaires ont été inclus et se sont avérés résider en amont des 2 éléments basaux et du site d'initiation de la transcription. L'emplacement et l'ordre des sites de liaison au facteur de transcription diversement indiqués ne sont que schématiques et non indicatifs comme étant typiques de tous les gènes d'ARNm eucaryotes. Il existe une vaste gamme de facteurs de transcription différents qui régulent la transcription des 3 classes de gènes eucaryotes codant pour les ARNm, les ARNt et les ARNr. CREB : protéine de liaison à l'élément de réponse à l'AMPc. C/EBP : protéine de liaison CCAAT-box/enhancer (**36**).

7.1.1.2. Les trois ARN polymérases eucaryotes

Les caractéristiques de la synthèse d'ARNm eucaryote sont nettement plus complexes que celles des procaryotes. Au lieu d'une seule polymérase comprenant cinq sous-unités, les eucaryotes ont trois polymérases composées chacune de 10 sous-unités ou plus. Chaque polymérase eucaryote nécessite également un ensemble distinct de facteurs de transcription pour l'amener à la matrice d'ADN.

L'ARN polymérase I est située dans le nucléole, une sous-structure nucléaire spécialisée dans laquelle l'ARN ribosomique (ARNr) est transcrit, traité et assemblé en ribosomes (tableau 1).

Les molécules d'ARNr sont considérées comme des ARN structuraux car elles ont un rôle cellulaire mais ne sont pas traduites en protéines. Les ARNr sont des composants du ribosome et sont essentiels au processus de traduction. L'ARN polymérase I synthétise tous les ARNr à l'exception de la molécule d'ARNr 5S. La désignation "S" s'applique aux unités "Svedberg", une valeur non additive qui caractérise la vitesse à laquelle une particule sédimente pendant la centrifugation.

L'ARN polymérase II est située dans le noyau et synthétise tous les pré-ARNm nucléaires codant pour les protéines. Les pré-ARNm eucaryotes subissent un traitement intensif après la transcription mais avant la traduction. Pour plus de clarté, la discussion de ce module sur la transcription et la traduction chez les eucaryotes utilisera le terme « ARNm » pour décrire uniquement les molécules matures et traitées qui sont prêtes à être traduites. L'ARN polymérase II est responsable de la transcription de l'écrasante majorité des gènes eucaryotes.

L'ARN polymérase III est également située dans le noyau. Cette polymérase transcrit une variété d'ARN structuraux qui comprend le pré-ARNr 5S, les pré-ARN de transfert (pré-ARNt) et les petits pré-ARN nucléaires. Les ARNt ont un rôle critique dans la traduction ; ils servent de molécules adaptatrices entre la matrice d'ARNm et la chaîne polypeptidique en croissance. Les petits ARN nucléaires ont une variété de fonctions, y compris «l'épissage» des pré-ARNm et la régulation des facteurs de transcription.

7.1.1.3.Facteurs de transcription pour l'ARN polymérase II : Mise en place du Complexe Pré-Initiation

La complexité de la transcription eucaryote ne s'arrête pas aux polymérases et aux promoteurs. Une armée de facteurs de transcription basaux, d'amplificateurs et de silencieux aide également à réguler la fréquence à laquelle le pré-ARNm est synthétisé à partir d'un gène. Les amplificateurs et les silencieux affectent l'efficacité de la transcription mais ne sont pas nécessaires pour que la transcription se poursuive. Les facteurs de transcription basaux sont cruciaux dans la formation d'un complexe de pré-initiation sur la matrice d'ADN qui recrute ensuite l'ARN polymérase II pour l'initiation de la transcription.

Pour que la transcription se produise, l'ARN polymérase II doit être recrutée à l'emplacement approprié - autour du site de démarrage de la transcription, sur le promoteur principal. Les premières étapes de la transcription eucaryote impliquent l'assemblage régulé des facteurs généraux de transcription (GTFS). Ces protéines servent de plate-forme pour le recrutement de l'ARN polymérase II.

Les GTF comprennent les facteurs TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TFIIH, ARN polymérase (ARN pol II). Nous nous concentrerons uniquement sur les fonctions de 2 GTF :

Comme son nom l'indique « TATA-binding protein », le TBP est une protéine spécifique à la séquence qui se lie à la boîte TATA. Les études de cristallographie aux rayons X du TBP montrent qu'il a une forme en forme de selle qui s'enroule partiellement autour de la double hélice.

La liaison du TFIID au promoteur central est suivie du recrutement d'autres GTF et éventuellement de l'ARN pol II. La combinaison de tous les GTF avec l'ARN Pol II est le complexe de pré-initiation (PIC). Le PIC adopte d'abord un état inactif, le complexe « fermé », qui est incompétent pour initier la transcription. Ce complexe est « prêt pour la transcription ».

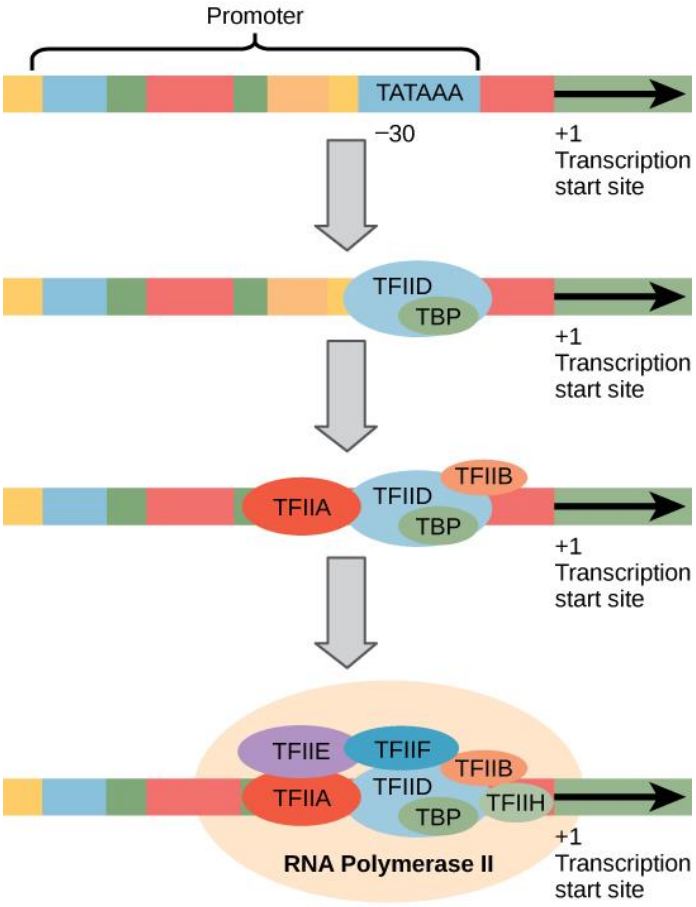


Figure 4. Un promoteur généralisé d'un gène transcrit par l'ARN polymérase II est montré. Les facteurs de transcription reconnaissent le promoteur. L'ARN polymérase II se lie alors et forme le complexe d'initiation de la transcription **(17).**

Les noms des facteurs de transcription basaux commencent par "TFII" (il s'agit du facteur de transcription de l'ARN polymérase II) et sont spécifiés par les lettres A à J. Les facteurs de transcription se mettent systématiquement en place sur la matrice d'ADN, chacun stabilisant davantage le complexe de préinitiation et contribuant au recrutement de l'ARN polymérase II.

Les processus d'amener les ARN polymérases I et III à la matrice d'ADN impliquent des collections légèrement moins complexes de facteurs de transcription, mais le thème général est le même. La transcription eucaryote est un processus étroitement régulé qui nécessite une variété de protéines pour interagir entre elles et avec le brin d'ADN. Bien que le processus de transcription chez les eucaryotes implique un investissement métabolique plus important que chez les procaryotes, il garantit que la cellule transcrira précisément les pré-ARNm dont elle a besoin pour la synthèse des protéines.

7.2.Allongement et terminaison chez les eucaryotes

Après la formation du complexe de pré-initiation, la polymérase est libérée des autres facteurs de transcription et l'allongement peut se dérouler comme chez les procaryotes avec la polymérase synthétisant le pré-ARNm dans la direction 5 'vers 3'. Comme discuté précédemment, l'ARN polymérase II transcrit la majeure partie des gènes eucaryotes, donc cette section se concentrera sur la façon dont cette polymérase accomplit l'allongement et la terminaison (**Figure 5**).

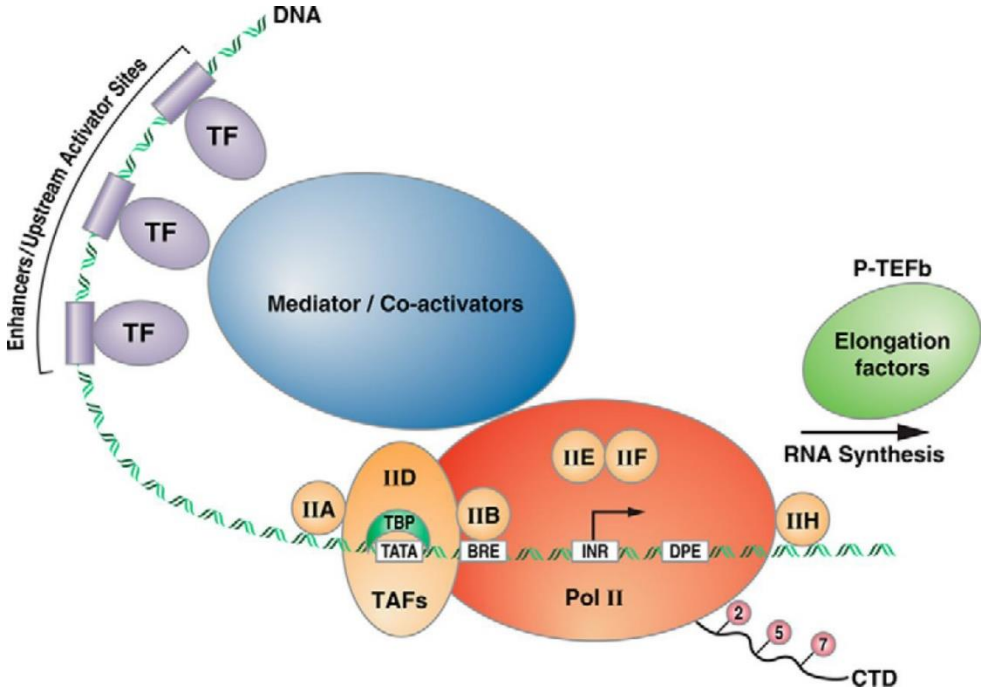


Figure 5. Facteur de l'allongement de la transcription **(25).**

Après l'initiation, la mécanique de l'allongement de la transcription est similaire à celle du procaryote, cependant, une grande différence est la modification de l'ARNm tel qu'il émerge de l'enzyme ARN pol II. La première modification se produit à l'extrémité 5 '.

Une fois que l'extrémité 5 'd'un ARN naissant s'étend sans RNAP II d'environ 20 à 30 nt, il est prêt à être coiffé par une structure de 7-méthylguanosine.

Il consiste en un nucléotide guanine connecté à l'ARNm via une liaison triphosphate 5 'à 5' inhabituelle (**Figure 6**). Cette guanosine est méthylée en position 7 directement après coiffage in vivo par une méthyltransférase. Il est appelé bouchon de 7-méthylguanylate, abrégé m7G.

Bien que le processus enzymatique d'élongation soit essentiellement le même chez les eucaryotes et les procaryotes, la matrice d'ADN est plus complexe. Lorsque les cellules eucaryotes ne se divisent pas, leurs gènes existent sous la forme d'une masse diffuse d'ADN et de protéines appelées chromatine. L'ADN est étroitement emballé autour de protéines histones chargées à intervalles répétés. Ces complexes ADN-histone, appelés collectivement nucléosomes, sont régulièrement espacés et comprennent 146 nucléotides d'ADN enroulés autour de huit histones comme un fil autour d'une bobine.

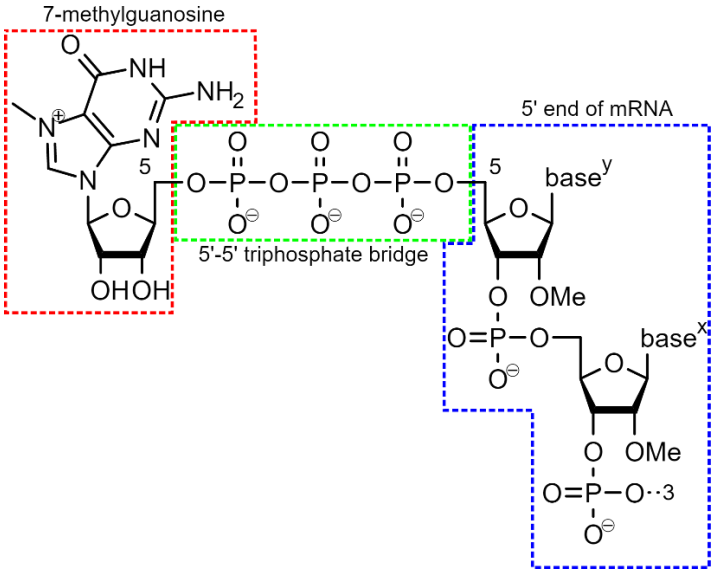


Figure 6. Structure CAP 5′. (***Depuis Naturwiki, CC BY-SA 4.0 <https://creativecommons.org/licenses/by-sa/4.0>, via Wikimedia Commons***)

Le processus comporte trois étapes. Tout d'abord, l'ARN triphosphatase élimine le groupe triphosphate 5'-terminal. Deuxièmement, la guanylation par le GTP est catalysée par une enzyme de coiffage, formant une liaison « arrière » inhabituelle 5'-5' entre la nouvelle guanine

et le premier nucléotide du transcrit d'ARN. Enfin, la guanine-7-méthyltransférase méthyle la guanine nouvellement attachée.

Une fois le CAP fabriqué, il est reconnu et lié par un complexe de protéines (CAP Binding Protein -CBP) qui reste associé au capuchon jusqu'à ce que l'ARNm ait été transporté dans le cytoplasme. L'extrémité 3 'du gène (dans le 3'UTR) est la séquence signature pour signaler la fin de la transcription et de la polyadénylation. Il se compose d'un site Poly A flanqué d'un signal de polyadénylation (AATAAA) et d'un élément en aval riche en GT.

Pour que la synthèse des polynucléotides se produise, la machinerie de transcription doit écarter les histones chaque fois qu'elle rencontre un nucléosome. Ceci est accompli par un complexe protéique spécial appelé FACT, qui signifie «facilite la transcription de la chromatine». Ce complexe éloigne les histones de la matrice d'ADN lorsque la polymérase se déplace le long de celle-ci. Une fois le pré-ARNm synthétisé, le complexe FACT remplace les histones pour recréer les nucléosomes.

La terminaison de la transcription est différente pour les différentes polymérases. Contrairement aux procaryotes, l'élongation par l'ARN polymérase II chez les eucaryotes a lieu 1 000 à 2 000 nucléotides au-delà de l'extrémité du gène en cours de transcription. Cette queue de pré-ARNm est ensuite éliminée par clivage pendant le traitement de l'ARNm. D'autre part, les ARN polymérases I et III nécessitent des signaux de terminaison. Les gènes transcrits par l'ARN polymérase I contiennent une séquence spécifique de 18 nucléotides qui est reconnue par une protéine de terminaison. Le processus de terminaison dans l'ARN polymérase III implique une épingle à cheveux d'ARNm similaire à la terminaison rho-indépendante de la transcription chez les procaryotes.

8. Structure de la chromatine et contrôle de l'expression des gènes

La chromatine est un complexe protéine-ADN présent chez les eucaryotes qui contient toutes les informations génétiques de l'organisme. Il est composé d'ADN étroitement enroulé autour d'un octamère d'histone, ou de deux copies de quatre protéines histones H2A, H2B, H3 et H4. Ce complexe ADN-histone s'appelle le nucléosome et permet à toutes les informations génétiques de s'intégrer dans le noyau de chaque cellule eucaryote. Pour que la cellule exprime certains gènes, la chromatine doit être rendue accessible aux facteurs de transcription. La chromatine peut être modifiée par des mécanismes épigénétiques comme les modifications d'histones lors de la transcription afin d'inhiber ou d'exprimer certains gènes. Ce processus est appelé remodelage de la chromatine.

est un processus biologique qui joue un rôle majeur dans l'expression des gènes, la réparation de l'ADN et l'apoptose. L'ADN est chargé négativement en raison des nombreux phosphates chargés négativement dans le squelette. Inversement, les protéines histones sont chargées positivement, de sorte que l'ADN et les protéines histones sont naturellement capables de s'enrouler ensemble.

On suppose que le remodelage de la chromatine est déterminé par le code des histones, qui est une théorie très complexe indiquant que la combinaison des modifications post-traductionnelles des histones affecte directement la transcription et l'expression génétiques. L'hypothèse implique qu'il existe un plan élaboré par des enzymes spécifiques qui permettent l'ajout ou la suppression de groupes méthyle ou acétyle aux histones, ou d'identifier les domaines de l'histone à modifier épigénétiquement, modifiant finalement l'expression des gènes (Figure 7).

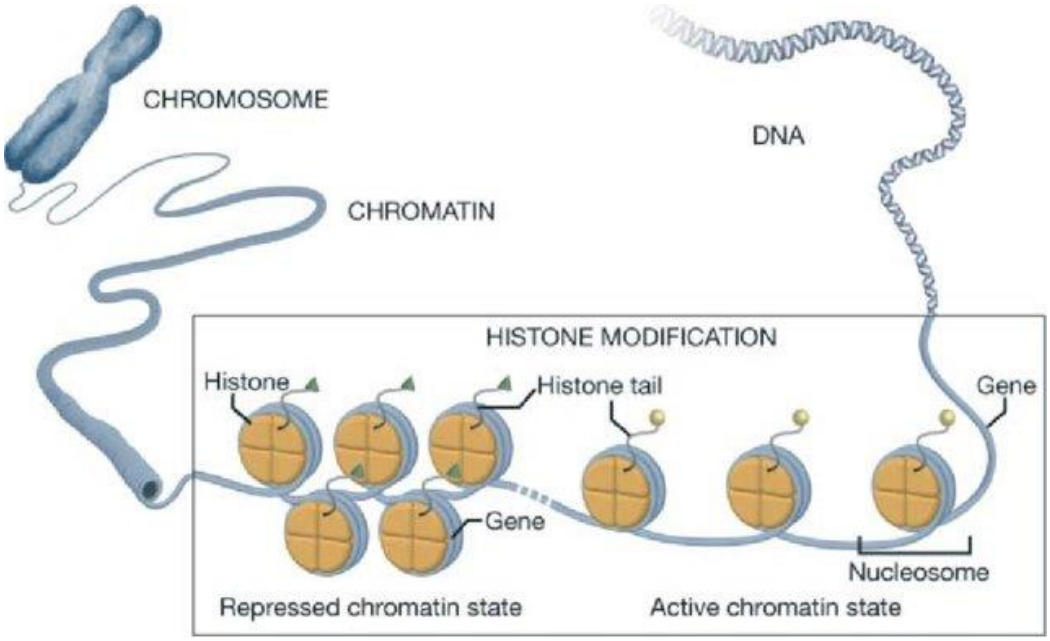


Figure 7. Structure de la chromatine et modifications des histones. L'ADN est enroulé autour d'octamères d'histones dans des nucléosomes. L'état de la chromatine est influencé par les modifications post-traductionnelles des queues d'histones **(28).**

Le code d'histone est incroyablement complexe, étant donné qu'il existe 19 lysines connues sur l'histone H3 seule connue pour être méthylées, et chacune peut être non méthylée, monométhylée, diméthylée ou triméthylée. Il existe de nombreuses autres modifications, notamment l'acétylation de la lysine, la méthylation de l'arginine, la phosphorylation de la thréonine/sérine/tyrosine sur l'histone H3. Des modifications supplémentaires survenant sur

d'autres protéines histones doivent également être prises en compte dans la complexité du code des histones.

9. Modifications des histones, structure de la chromatine, régulation transcriptionnelle

9.1. Acétylation des histones

Les protéines histones sont soumises à un certain nombre de modifications et ces modifications sont connues pour affecter la structure de la chromatine. L'acétylation des histones est connue pour entraîner une structure de chromatine plus ouverte et ces histones modifiées se trouvent dans les régions de la chromatine qui sont transcriptionnellement actives. Inversement, la sous-acétylation (désacétylation) des histones est associée à une chromatine fermée et à une inactivité transcriptionnelle. Une corrélation directe entre l'acétylation des histones et l'activité transcriptionnelle a été démontrée lorsqu'il a été découvert que des complexes protéiques, précédemment connus pour être des activateurs transcriptionnels, se sont avérés avoir une activité histone acétyltransférase (HAT). Et comme prévu, les complexes répresseurs transcriptionnels contiennent une activité d'histone désacétylase (HDAC) (Figure 8).

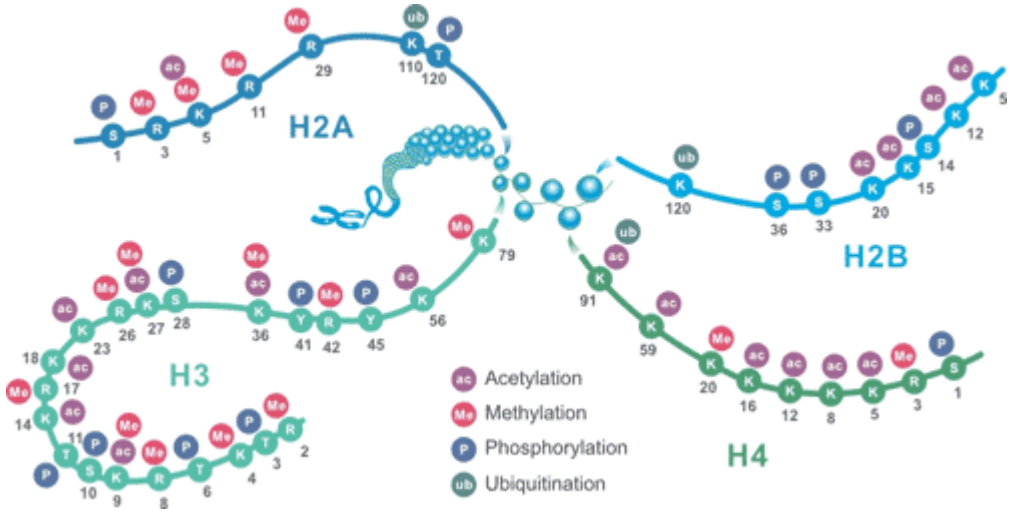


Figure 8. Le schéma du site de modification des histones communes **(10).**

Les enzymes qui acétylent le groupe ε-amino des résidus de lysine dans les protéines en général et les histones en particulier sont des membres de la grande famille des lysine acétyltransférases (KAT) qui est composée de 17 gènes chez l'homme. De nombreuses protéines non histones qui sont acétylées sont impliquées dans la réplication, la recombinaison et la réparation de l'ADN, ainsi que dans les facteurs de transcription et de nombreux autres types de protéines. L'analyse mondiale des protéines a identifié plus de 1 700 protéines humaines qui sont modifiées par

l'acétylation des résidus de lysine. Les 17 gènes KAT humains ont été classés en cinq sous-familles en fonction de l'homologie de séquence, des caractéristiques structurelles partagées et des propriétés d'acétylation du substrat. Les histones acétyltransférases de mammifères (HAT) sont localisées au niveau nucléaire (souvent appelées HAT de type A) ou localisées dans le cytoplasme (souvent appelées HAT de type B). Tous les HAT nucléaires contiennent un bromodomaine leur permettant de reconnaître et d'interagir avec les lysines acétylées dans les substrats d'histones. Les HAT cytoplasmiques sont responsables de l'acétylation des protéines histones nouvellement synthétisées avant leur transport dans le noyau.

La sous-famille HAT1 est composée de deux membres, HAT1 et HAT4 (la désignation officielle du gène est NAA60 pour N(alpha)-acétyltransférase 60, sous-unité catalytique NatF). Les protéines de la sous-famille HAT1 sont toutes deux des enzymes cytoplasmiques qui acétylent les protéines histones nouvellement synthétisées. La protéine HAT1 acétyle les histones K5 et K12 en histone H4. Il convient de noter que certaines désignations incluent le gène HAT1 dans la sous-famille GCN5/PCAF (GNAT).

La sous-famille GCN5 / PCAF (également connue sous le nom de N-acétyltransférase liée à GCN5, GNAT) est ainsi appelée en raison de la caractérisation initiale de l'activité histone acétyltransférase de la protéine codée par le gène GCN5 (contrôle général non dépressible 5) dans les protozoaires, Tetrahymena thermophila. Le nom du gène PCAF est dérivé du facteur associé à p300/CBP. La sous-famille GCN5/PCAF comprend les deux gènes dont le nom de groupe est dérivé, GCN5 (KAT2A) et PCAF (KAT2B). Les protéines codées KAT2A et KAT2B acétylent les histones H3 et H4.

La sous-famille MYST porte le nom des quatre premiers membres du groupe; MOZ, YBF2/SAS3, SAS2 et TIP60. La sous-famille MYST humaine est composée de cinq protéines, KAT5 (TIP60), KAT6A (MOZ), KAT6B, KAT7 et KAT8. La protéine codée par KAT5 acétyle les histones H2A et H4. Les protéines codées KAT6A, KAT6B, KAT7 et KAT8 acétylent les histones H3 et H4.

La sous-famille p300/CBP comprend les deux protéines, p300 et CBP, qui ont dérivé le nom de la sous-famille. Le nom de la protéine p300 est dérivé de sa masse moléculaire et la protéine est codée par le gène EP300 (adenovirus E1A binding protein p300). La protéine p300 est également définie par la nomenclature KAT standard en tant que KAT3B. Le nom de la protéine CBP est dérivé de la protéine de liaison CREB (cAMP-response element binding protein). La protéine CPB est codée par le gène CREBBP qui est également identifié par la nomenclature

KAT standard comme KAT3A. Les protéines codées par CREBBP/KAT3A et EP300/KAT3B acétylent les quatre histones du nucléosome, H2A, H2B, H3 et H4.

La sous-famille SRC constitue les corégulateurs des récepteurs nucléaires qui ont une activité histone acétyltransférase. Le nom SRC est dérivé de l'identification originale du coactivateur 1 des récepteurs stéroïdiens (SRC-1). SRC-1 est codé par le gène NCOA1. La sous-famille SRC est composée de trois membres codés par les gènes NCOA1, NCOA2 (à l'origine GRIP1 pour la protéine 1 interagissant avec le récepteur des glucocorticoïdes et également TIF2 pour le facteur intermédiaire transcriptionnel 2) et NCOA3 (initialement identifié comme SRC-3). Chacun des HAT de la sous-famille SRC acétyle les histones H3 et H4.

9.2.Désacétylation des histones

La désacétylation des histones est nécessaire pour réguler les effets positifs ou négatifs sur l'expression des gènes exercés par l'acétylation des histones. La désacétylation des histones est catalysée par une grande superfamille d'enzymes composée des gènes de sirtuine (SIRT) et des gènes d'histone désacétylase (HDAC). Les gènes HDAC sont ensuite divisés en trois sous-familles identifiées comme classe I, classe II et classe IV. La sous-famille HDAC de classe II est ensuite divisée en sous-familles de classe IIA et de classe IIB. La sous-famille des gènes de la sirtuine humaine est composée de sept gènes identifiés comme SIRT1-SIRT7. La sous-famille HDAC I est composée de quatre gènes. La sous-famille HDAC IIA est composée de quatre gènes. La sous-famille HDAC IIB est composée de deux gènes. La sous-famille HDAC IV est composée d'un seul gène, HDAC11. On sait peu de choses sur les fonctions globales de la protéine HDAC11. Les gènes sirtuines sont souvent appelés la sous-famille des HDAC de classe III. Toutes les enzymes HDAC sont des désacétylases dépendantes du Zn2+, tandis que les sirtuines sont des enzymes dépendantes du NAD+ (Figure 9).

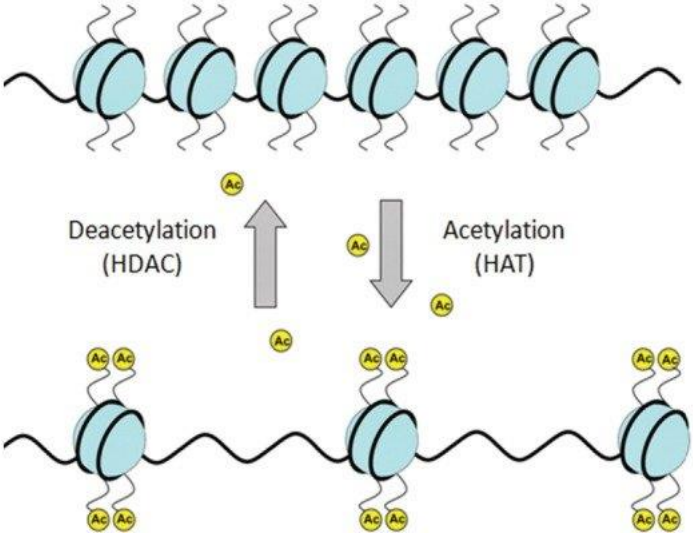


Figure 9. Acétylation et désactylation des histones. L'histone acétyltransférase (HAT) ajoute des groupes acétyle (Ac) sur les queues d'histone, ce qui entraîne une ouverture du nucléosome permettant ainsi aux facteurs de transcription d'accéder à l'ADN et d'initier la transcription des gènes. Les histone désacétylases (HDAC) éliminent l'Ac des queues d'histone, conduisant à une fermeture structure de la chromatine **(29).**

Toutes les enzymes HDAC de classe I sont des enzymes localisées dans le noyau exprimées de manière ubiquitaire. De plus, HDAC1, HDAC2 et HDAC3 sont des composants de complexes multiprotéiques, alors que HDAC8 ne l'est pas. Les protéines HDAC1 et HDAC2 forment à la fois des homodimères et des hétérodimères entre elles. HDAC1 et HDAC2 se trouvent dans au moins trois complexes corépresseurs multiprotéiques distincts. Ces complexes de corépresseurs sont le remodelage et la désacétylation des nucléosomes (NRD ; également appelé NuRD), CoREST [corépresseur de REST (facteur de transcription silencieux RE1)], mSin3 et la désacétylase associée à la Nanog et à l'Oct4 (NODE). En plus de l'activité de désacétylation fournie par HDAC1 et HDAC2, le complexe CoREST recrute l'histone déméthylase (voir section suivante) KDM1 qui déméthyle le résidu K4 diméthylé dans l'histone H3. HDAC3 est un composant du corépresseur du récepteur nucléaire (NCoR ou NCOR1) et du médiateur de silençage des complexes corépresseurs transcriptionnels de l'acide rétinoïque et du récepteur de l'hormone thyroïdienne (SMRT ou NCOR2).

Les protéines HDAC de classe IIA ont toutes des profils d'expression spécifiques aux tissus ainsi que des fonctions distinctes. Les quatre protéines de la sous-famille de classe IIA font la navette entre le cytoplasme et le noyau. Ce processus de navette est régulé par leur état de phosphorylation. Étant donné que les protéines HDAC de classe IIA ont toutes une substitution d'acide aminé (Tyr pour His) dans leurs domaines catalytiques, ces HDAC ont peu d'activité de

désacétylase intrinsèque. La fonction principale des HDAC de classe IIA est la liaison des résidus de lysine acétylée dans d'autres protéines, recrutant ainsi des complexes modifiant la chromatine sur des gènes cibles spécifiques. Les HDAC de classe IIA fonctionnent comme des désacétylases grâce à leur capacité à recruter des complexes de corépresseurs contenant HDAC3 vers des promoteurs distincts.

Les protéines HDAC de classe IIB font également la navette entre le cytoplasme et le noyau bien qu'elles ne se trouvent principalement que dans le cytoplasme. Une caractéristique de cette classe d'enzymes est qu'elles ont toutes des domaines catalytiques dupliqués. Une fonction majeure de HDAC6 cytoplasmique est la clairance des protéines mal repliées par la voie de l'autophagie ou par la formation d'aggresomes.

9.3. Méthylation des histones

Une autre modification d'histone connue pour affecter la structure de la chromatine est la méthylation. La méthylation des histones, ainsi que de nombreuses autres protéines non histones, se produit sur les résidus lysine et arginine. La méthylation de la lysine des histones peut entraîner trois états distincts, la monométhylation, la diméthylation ou la triméthylation. Cependant, avec la méthylation de l'histone lysine, il n'y a pas de corrélation directe entre la modification et un effet spécifique sur la transcription. La méthylation de l'histone lysine (K) à certaines positions est associée à des régions de chromatine muette transcriptionnellement, tandis que la méthylation à d'autres positions est associée à des régions transcriptionnellement actives de l'ADN. Il a été démontré que la méthylation de l'histone arginine (R) est associée à la promotion d'une structure de chromatine ouverte et, par conséquent, entraîne une activation de la transcription. La méthylation des résidus de lysine (K) dans l'histone H3 (en particulier K9 et K27) et l'histone H4 (K20) est associée à des régions de chromatine silencieuse par transcription. Ces sites de méthylation spécifiques sont identifiés comme H3K9, H3K27 et H4K20. Inversement, la méthylation à H3K4, H3K36 et H3K79 est associée à des domaines transcriptionnellement actifs dans la chromatine. Cependant, ces associations ne sont pas concrètes étant donné que la méthylation de H3K9 a été trouvée dans des gènes transcriptionnellement actifs et que la méthylation de H3K36 s'est avérée être associée à la répression de l'initiation de la transcription intragénique.

Toutes les enzymes lysine méthyltransférase appartiennent à la grande famille d'enzymes identifiée comme la famille de la lysine (K) méthyltransférase (KMT). Les histones lysine méthyltransférases sont également identifiées comme des HMTases (pour histone

méthyltransférases). Les humains expriment une famille de 34 gènes codant pour la protéine lysine méthyltransférase, qui ne méthylent pas toutes les histones. Les enzymes qui effectuent la méthylation de l'histone lysine sont toutes des membres de la famille des méthyltransférases contenant le domaine SET (à l'exception d'une enzyme : DOT1L). Le domaine SET est ainsi appelé car il a été identifié à l'origine dans trois protéines de *Drosophila melanogaster* identifiées comme suppresseur de la variante de panachure 3-9 [Su (var) 3-9], amplificateur de zeste et Trithorax. Le domaine SET est composé d'environ 130 acides aminés. Il existe quatre histones méthyltransférases supplémentaires qui appartiennent à une famille de protéines différente identifiée comme la famille des facteurs de transcription contenant les domaines PR et SET, identifiée comme la famille PRDM. Le domaine PR de tous les membres de la famille PRDM contient un domaine à doigts de zinc. La famille de domaines PR / SET contient 17 membres avec PRDM2 (également identifié comme KMT8), PRDM8, PRDM9 (voir la figure ci-dessous) et éventuellement PRDM14 possédant une activité d'histone méthyltransférase.

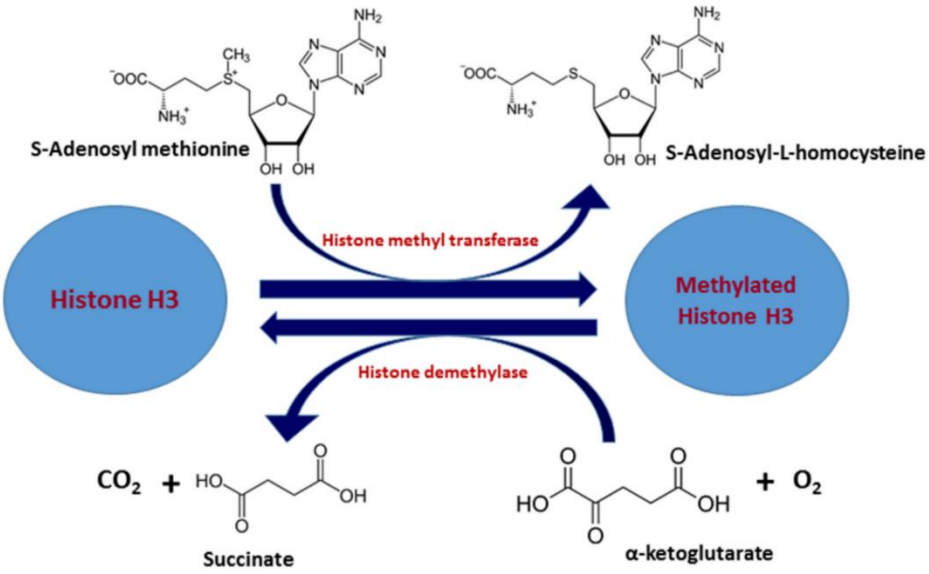


Figure 10. Processus de méthylation et de déméthylation de la protéine lysine. La méthylation et la déméthylation de la lysine de la protéine histone (ainsi que d'autres protéines) sont catalysées par une famille de lysine méthyl transférases et de lysine déméthylases **(31).**

9.4.Ubiquitylation des histones

Les protéines histones peuvent également être modifiées par l'ajout de la petite protéine ubiquitine. En ce qui concerne les histones, l'ubiquitine se trouve sur toutes les histones nucléosomales (H2A, H2B, H3 et H4) ainsi que sur l'histone de liaison, H1. Cependant, la grande majorité des histones ubiquitylées sont H2A et H2B et elles sont toutes deux de la forme

monoubiquitine. La monoubiquitylation de H2A se produit à Lys 119 (K119) et celle de H2B est K120.

Bien que la monoubiquitylation de H2A et H2B prédomine, une polyubiquitylation est observée. La polyubiquitylation de K36 dans l'histone H2A et le variant H2AX est associée à des réponses aux dommages à l'ADN et cette modification est nécessaire pour que les processus de réparation soient initiés. Les histones H3 et H4 sont également connues pour être polyubiquitylées mais les fonctions biologiques précises de ces histones modifiées ne sont pas entièrement élucidées. Lorsqu'il est ubiquitylé, H2A est associé à la répression de la transcription. L'effet exactement opposé est observé lorsque l'histone H2B est ubiquitylée, conduisant à une stimulation de l'activité des gènes.

L'une des raisons pour lesquelles l'histone H2B monoubiquitylée est associée à une activité transcriptionnelle est que cette modification favorise la méthylation de l'histone H3 en K4 et K79, qui, comme indiqué ci-dessus, est associée à une structure de chromatine ouverte. Étant donné que l'ubiquitylation de H2A est principalement associée au silençage génique, il n'est pas surprenant que les ligases d'ubiquitine H2A soient associées à des complexes corépresseurs transcriptionnels.

Il a été démontré qu'au moins sept ligases d'ubiquitine différentes ubiquitylent les histones. La grande majorité de ces caractérisations ont été réalisées avec des études sur la monoubiquitylation de H2A et H2B. La monoubiquitylation de H2A et H2B est connue pour être réversible et les enzymes qui catalysent l'élimination sont des peptidases identifiées comme des enzymes de déubiquitylation (DUB). Au moins six enzymes DUB différentes ont été identifiées comme étant impliquées dans l'élimination de la monoubiquitine de H2A et H2B.

9.5.Phosphorylation des histones

La phosphorylation des histones est connue pour se produire sur les quatre histones nucléosomales, H2A, H2B, H3 et H4. La phosphorylation des histones se produit sur les résidus Ser, Thr et Tyr par l'action de plusieurs kinases. L'élimination de la phosphorylation est catalysée par les phosphatases. La phosphorylation des histones se produit principalement, mais pas exclusivement, en réponse à des signaux extérieurs tels que la stimulation du facteur de croissance ou des inducteurs de stress tels qu'un choc thermique. Les histones phosphorylées sont localisées dans des gènes qui deviennent transcriptionnellement actifs en conséquence de ces signaux extérieurs. La phosphorylation des protéines histones est également nécessaire pour

réguler d'autres formes de modification des histones. Par exemple, la phosphorylation de Ser 1 (S1) dans l'histone H4 empêche l'acétylation de cette histone.

Il a été démontré que de nombreux résidus dans les quatre histones nucléosomiques sont phosphorylés, entraînant une altération de l'activité transcriptionnelle. Les sites de phosphorylation dans l'histone H2A comprennent Ser 1 (S1), S16 et Thr 119 (T119). Les conséquences de la modification H2AS1 sont l'inhibition de la transcription, alors que H2AT119 est associé à la régulation de la structure de la chromatine au cours de la mitose. L'histone H2B est phosphorylée sur S14, S32, S36 et Tyr 37 (Y37). La modification H2BS14 est impliquée dans l'induction de l'apoptose. La phosphorylation de H2B S32 est catalysée par PKC en réponse à des dommages à l'ADN. La phosphorylation de H2B S36 est catalysée par l'AMPK en réponse aux voies de réponse au stress cellulaire. L'histone H3 est phosphorylée sur de nombreux résidus, notamment T3, T6, S10, T11, S28, Y41 et T45. L'histone H4 est phosphorylée sur S1, S47, His 18 (H18) et H75. La phosphorylation des résidus d'histidine dans l'histone H4 est associée à la facilitation de la réplication de l'ADN.

En plus de la régulation de la transcription résultant de la phosphorylation des histones, cette modification est également associée aux processus de remodelage de la chromatine et de réparation des dommages à l'ADN. Un gène H2A particulier, identifié comme H2AFX, code pour une histone indépendante de la réplication (protéine identifiée comme H2AX ou H2a.X) qui est impliquée de manière critique dans la réponse des cellules aux cassures double brin de l'ADN (DSB). La phosphorylation de Ser 139 (S139) dans H2AX génère l'histone modifiée identifiée comme γH2AX. La phosphorylation de H2AX se produit tout au long du cycle cellulaire en réponse à divers événements de réponse aux dommages à l'ADN (DDR) tels que la jonction d'extrémités non homologues (NHEJ), la recombinaison homologue et la réparation de l'ADN couplée à la réplication.

Après réparation de l'ADN endommagé, γH2AX est retiré de l'ADN afin de mettre fin à la rétention des enzymes de réparation des dommages à l'ADN. En plus de l'élimination de la chromatine, γH2AX est déphosphorylé par un certain nombre de phosphatases, dont PP2A. Il a également été démontré que la protéine H2AX est phosphorylée sur Tyr 142 (Y142), ce qui donne l'isoforme identifiée comme H2AXY142. La phosphorylation de l'histone H2B sur Ser 14 (S14) est également associée à des réponses aux dommages à l'ADN et à l'induction de l'apoptose.

L'importance de la phosphorylation des histones en réponse aux dommages à l'ADN peut être démontrée chez les patients atteints du syndrome de Coffin-Lowry qui résulte de défauts du gène RPS6KA3 (protéine ribosomale S6 kinase A3; également connue sous le nom de kinase ribosomale S6 2 : RSK2). Le syndrome de Coffin-Lowry est une forme rare de déficience intellectuelle liée à l'X caractérisée par des malformations squelettiques, un retard de croissance, un déficit auditif, des troubles du mouvement paroxystique et une déficience cognitive chez les hommes atteints.

10. Récepteurs nucléaires et contrôle de l'initiation transcriptionnelle

10.1. Structure du récepteur nucléaire

La superfamille des récepteurs nucléaires est une famille de facteurs de transcription qui sont largement exprimés dans tout le corps. Cette famille fonctionne dans des voies de signalisation bien organisées qui dépendent fortement du microenvironnement tissulaire et, lorsqu'elles sont perturbées, de manière endogène ou exogène, peuvent provoquer un dysfonctionnement des organes, un cancer ou une perte d'intégrité tissulaire. Une intervention pharmacologique inhibant les voies de signalisation des membres de cette famille a été utilisée pour le traitement de nombreuses maladies. Sur la base de l'évolution et de la réponse thérapeutique robuste aux thérapies anti-androgènes, nous examinons différents agents actuellement utilisés à différents stades de la progression du cancer de la prostate ainsi que de nouvelles cibles explorées en raison d'une augmentation de la résistance au traitement (Figure 1).

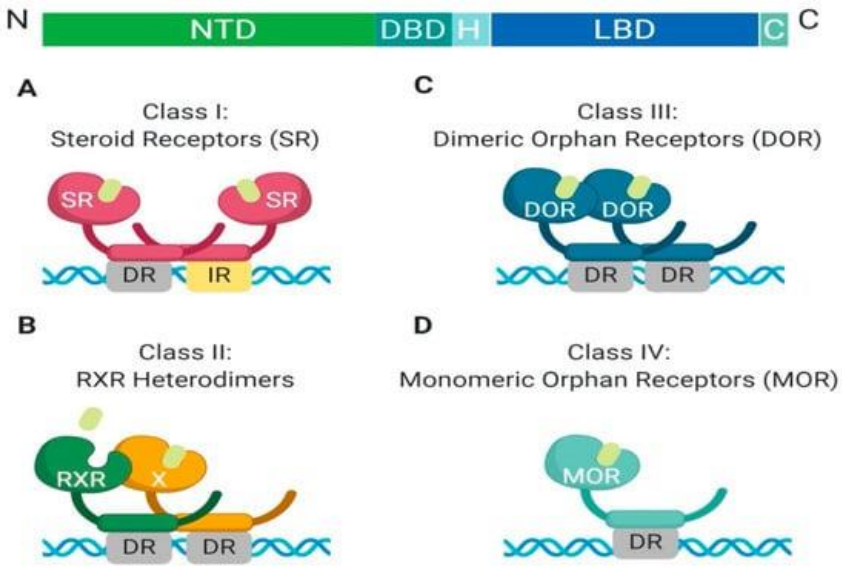


Figure 11. Illustration schématique de la superfamille des récepteurs nucléaires classiques. (A–D) représentent graphiquement les quatre classes de la superfamille des récepteurs nucléaires

qui sont définies en fonction de la dimérisation (homo, hétéro ou mono), de la liaison à l'ADN (répétition directe ou répétition inversée) et de la spécificité du ligand (requise ou non requise) . Classe I, récepteur stéroïde (également connu sous le nom de récepteurs hormonaux nucléaires); Classe II, hétérodimères RXR ; Classe III, récepteurs orphelins dimériques ; Classe IV, récepteurs orphelins monomériques. Abréviations : NTD, domaine N-terminal ; DBD, domaine de liaison à l'ADN ; H, région charnière ; LBD, domaine de liaison au ligand ; C, extrémité C-terminale variable ; DR, répétition directe ; IR, répétition inversée **(30).**

Tous les membres de la superfamille des récepteurs nucléaires contiennent un domaine N-terminal variable (NTD), un domaine de liaison à l'ADN (DBD), une région charnière, un domaine de liaison au ligand conservé (LBD) et un domaine C-terminal variable. Les deux domaines les plus conservés parmi tous les récepteurs nucléaires sont le domaine de liaison à l'ADN et le domaine de liaison au ligand. Le domaine de liaison à l'ADN contient deux motifs à doigts de zinc, qui agissent comme un crochet, qui permettent la liaison à la chromatine dans le noyau. Chaque classe a différentes séquences de reconnaissance de liaison à l'ADN, qui vont de demi-sites variables avec des répétitions inversées, des répétitions directes ou aucune répétition dans la séquence d'ADN.

Le domaine de liaison au ligand des récepteurs nucléaires reste hautement conservé en fonction mais diffère en spécificité et en affinité pour des ligands spécifiques. Toutes les classes, à l'exception des récepteurs orphelins, sont activées par des ligands. La liaison du ligand au niveau du LBD induit un changement allostérique, induisant une activation. Les ligands de chaque classe de récepteurs nucléaires ont des structures similaires. De plus, la classification du ligand détermine à quelle classe de récepteurs nucléaires chacun appartient. Par exemple, les ligands exprimés de manière endogène pour ces récepteurs peuvent être des hormones, des métabolites ou des ligands enzymatiques, ainsi que des ligands non identifiés.

Une autre caractéristique qui différencie les membres de la classe est la dimérisation des partenaires au sein du noyau. Les classes I à III nécessitent une dimérisation, contrairement à la classe IV. De plus, les classes I et III nécessitent une homodimérisation, qui peut fournir une liaison plus forte du doigt de zinc à l'ADN, tandis que la classe II nécessite une hétérodimérisation.

Des modifications ont été apportées à chaque sous-classe en fonction des nouvelles informations recueillies grâce à l'analyse structurelle et aux données de séquençage. Pour cette revue, nous nous concentrerons sur les subdivisions classiques de la superfamille des récepteurs

nucléaires définies par les caractéristiques de la structure et de la fonction de la superfamille des récepteurs nucléaires telles que la dimérisation, les motifs et la spécificité de liaison à l'ADN et l'activation de la liaison au ligand.

10.2. Coactivateurs de récepteurs nucléaires

Le premier coactivateur de récepteur nucléaire à être identifié était le coactivateur de récepteur stéroïdien-1 (SRC-1). À ce jour, plus de 400 corégulateurs (coactivateurs et corépresseurs) ont été identifiés. On sait maintenant qu'il existe trois familles de gènes SRC. SRC-1 (codé par le gène NCOA1), SRC-2 (également connu sous le nom de GRIP1 pour la protéine 1 interagissant avec le récepteur des glucocorticoïdes et TIF2 pour le facteur intermédiaire transcriptionnel 2) codé par le gène NCOA2 et SRC-3 (également connu sous le nom de AIB1 pour amplifié dans le cancer du sein 1 et TRAM-1 pour la molécule activatrice des récepteurs des hormones thyroïdiennes 1) codée par le gène NCOA3. Les trois membres de la famille SRC contiennent des domaines homologues et partagent entre 50% et 54% de similarité de séquence d'acides aminés. Il existe également une famille diversifiée d'enzymes qui interagissent avec les SRC et les modifient, notamment les histones acétyltransférases (HAT), les histones méthyltransférases (HMT), les kinases, les phosphatases, les ubiquitine ligases et les petites ligases modificatrices liées à l'ubiquitine (SUMO).

Le récepteur gamma activé par les proliférateurs de peroxysomes, le coactivateur 1 alpha (PGC-1α) est un autre corégulateur NR critique. Il a été démontré que PGC-1α est impliqué dans la régulation du métabolisme et de l'homéostasie énergétique. En effet, les niveaux d'expression de PGC-1α ont été associés à des maladies génétiques associées à une altération de la fonction mitochondriale, notamment le diabète de type 2 et l'obésité. Un autre coactivateur important est le CBP [CREBP (cAMP response-element binding protein)-binding protein]. Le CBP est étroitement lié à un autre coactivateur appelé p300. La désignation de p300 se rapporte à la taille moléculaire de la protéine initialement caractérisée. La protéine p300 est codée par le gène EP300. CBP et p300 possèdent une activité intrinsèque d'histone acétyltransférase (HAT) qui conduit à la relaxation de la structure de la chromatine près d'un gène cible NR. D'autres complexes de remodelage de la chromatine, tels que l'arginine méthyltransférase 1 associée au coactivateur (CARM1), peuvent également stimuler la transcription génique par les NR ainsi que d'autres facteurs de transcription en combinaison avec la famille des coactivateurs SRC.

En plus d'agir en tant que coactivateurs pour les NR, les protéines de la famille SRC interagissent également avec de nombreux types différents de facteurs de transcription et

potentialisent leur activité transcriptionnelle. Ceux-ci incluent p53, les transducteurs de signal et les activateurs de la transcription (STAT), le facteur nucléaire-κB (NF-κB), le facteur 1 inductible par l'hypoxie (HIF1) et le facteur nucléaire hépatocytaire-4 (HNF4) pour n'en nommer que quelques-uns. Plusieurs stimuli extracellulaires, tels que des facteurs de croissance et des cytokines, qui activent les récepteurs transducteurs de signaux transmembranaires, générant des codes de phosphorylation sur les SRC qui conduisent à une affinité accrue du coactivateur pour le récepteur des androgènes (AR), le récepteur des œstrogènes alpha (ERα) et le récepteur de la progestérone (RP).

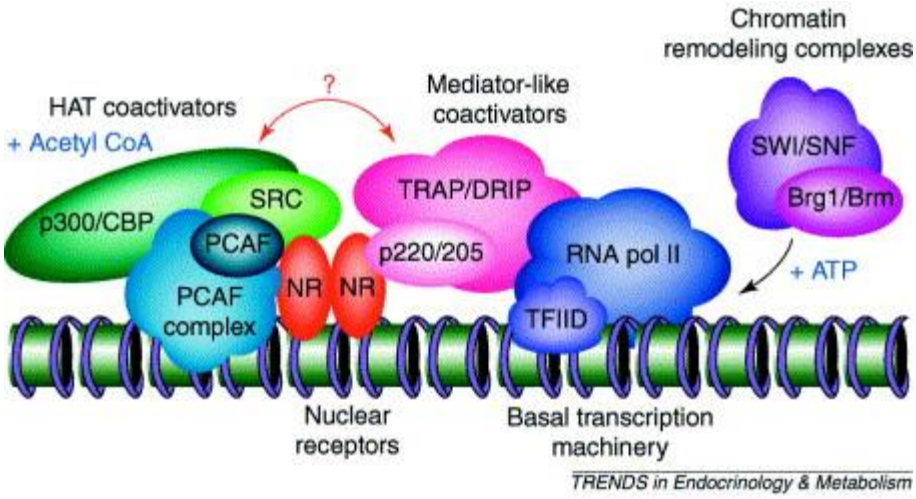


Figure 12. De multiples interactions physiques et fonctionnelles entre les récepteurs nucléaires, les coactivateurs, les remodeleurs de la chromatine et la chromatine conduisent à une séquence ordonnée d'événements aboutissant à la transcription de gènes régulés par les hormones, y compris : (1) l'interaction dépendante du ligand des coactivateurs avec les NR liés à la chromatine, (2) le remodelage de la chromatine dépendant de l'ATP par des complexes de remodelage de la chromatine, (3) l'acétylation des histones par les coactivateurs HAT et (4) les contacts entre les NR et la base machinerie transcriptionnelle par les coactivateurs de type médiateur (par exemple TRAP / DRIP). La question de savoir si des interactions fonctionnelles se produisent entre les coactivateurs SRC-p300/CBP-PCAF et le complexe TRAP/DRIP n'est actuellement pas claire, comme l'indique le point d'interrogation. Abréviations : acétyl CoA, acétyl coenzyme A (le donneur d'acétyle pour les réactions d'acétylation) ; CBP, protéine liant CREB; CREB, protéine de liaison à l'élément de réponse à l'AMPc ; HAT, histone acétyltransférase; NR, récepteur nucléaire ; PCAF, facteur associé à p300/CBP ; p220/205, TRAP220/DRIP205 ; ARN pol II, ARN polymérase II; SRC, coactivateur des récepteurs stéroïdiens ; TFIID, le facteur de transcription IID [qui contient la protéine de liaison TATA (TBP) et les facteurs associés au TBP] ; TRAP/DRIP, protéines associées aux récepteurs des hormones thyroïdiennes/protéines interagissant avec les récepteurs de la vitamine D **(21).**

10.3. Corepresseurs des récepteurs nucléaires

En règle générale, il a été établi que lorsque les récepteurs nucléaires sont exempts de ligand activateur, ils interagissent préférentiellement avec les complexes corépresseurs pour médier la répression transcriptionnelle. Le corépresseur du récepteur nucléaire 1 (NCoR1) et le médiateur de silençage des récepteurs rétinoïques et thyroïdiens (SMRT) sont les complexes de corépresseurs NR les mieux caractérisés. La protéine NCoR1 est codée par le gène NCOR1 et la protéine SMRT est codée par le gène NCOR2. Le complexe protéique central NCoR/SMRT se compose de NCoR/SMRT, de la transducine β-like 1/related 1 (TBL1/TBLR1 : codée par le gène TBL1), de l'histone désacétylase 3 (codée par le gène HDAC3) et du suppresseur de la voie de la protéine G. 2 (codé par le gène GPS2). NCoR et SMRT servent de sites d'amarrage pour l'assemblage du complexe corépresseur. NCoR/SMRT se lient à divers récepteurs nucléaires et s'associent à chacune des autres sous-unités complexes.

Lorsque le NR interagit avec le ligand, l'activation transcriptionnelle résulte de la capacité du complexe NR-ligand à recruter des protéines coactivatrices et à déplacer les protéines corépresseurs. Les corépresseurs des récepteurs nucléaires peuvent inhiber l'activité transcriptionnelle de la plupart des membres de la superfamille NR. Comme toujours en biologie, il existe quelques exceptions à la règle générale des corépresseurs de liaison NR sans ligand. Ces exceptions incluent LCoR (corépresseur du récepteur nucléaire dépendant du ligand ; codé par le gène LCOR), RIP140 (protéine interagissant avec le récepteur 140 ; codé par le gène NRIP1) et le répresseur de l'activité du récepteur des œstrogènes (REA ; codé par le gène de la prohibitine 2 , PHB2). Ces répresseurs se lient aux récepteurs nucléaires de manière ligand-dépendante et entrent en compétition avec les coactivateurs en les déplaçant. En outre, il existe plusieurs facteurs corégulateurs, tels que les complexes de remodelage de la chromatine dépendants de l'ATP SWI/SNF (commutation de type d'accouplement/saccharose non fermentant, complexe de remodelage de la chromatine), qui se sont avérés impliqués dans la régulation à la fois de la transcription activation et répression.

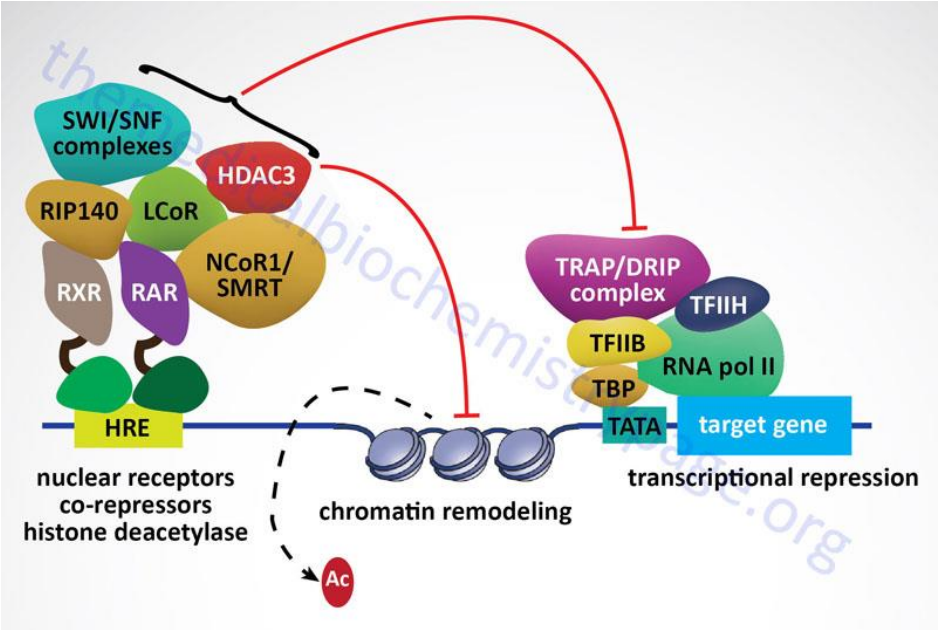


Figure 13. Modèle d'interactions des récepteurs nucléaires (NR) avec les corépresseurs : un exemple des complexes de corépresseurs de transcription associés au complexe de facteurs de transcription hétérodimériques RXR et RAR au niveau d'un HRE, et à plusieurs facteurs de transcription basaux associés à l'ARN pol II au niveau d'un site de début de transcription du gène cible. La présence d'histone désacétylases (par exemple HDAC3) conduit à l'élimination de tout site d'acétylation d'histone activant la chromatine, provoquant la formation d'une structure de chromatine réprimée par la transcription (**36**).

6. Contrôle post-transcriptionnel de l'expression génique

La régulation post-transcriptionnelle se produit après la transcription de l'ARNm mais avant le début de la traduction. Cette régulation peut se produire au niveau du traitement de l'ARNm, du transport du noyau au cytoplasme ou de la liaison aux ribosomes.

6.1. Épissage alternatif d'ARN

L'épissage alternatif de l'ARN est un mécanisme qui permet de retirer différentes combinaisons d'introns, et parfois d'exons, du transcrit primaire (Figure 17.11). Cela permet à différents produits protéiques d'être produits à partir d'un gène. L'épissage alternatif peut agir comme un mécanisme de régulation des gènes. L'épissage différentiel est utilisé pour produire différents produits protéiques dans différentes cellules ou à différents moments dans la même cellule. L'épissage alternatif est maintenant compris comme un mécanisme commun de régulation des gènes chez les eucaryotes ; jusqu'à 70% des gènes chez l'homme sont exprimés sous forme de protéines multiples par épissage alternatif.

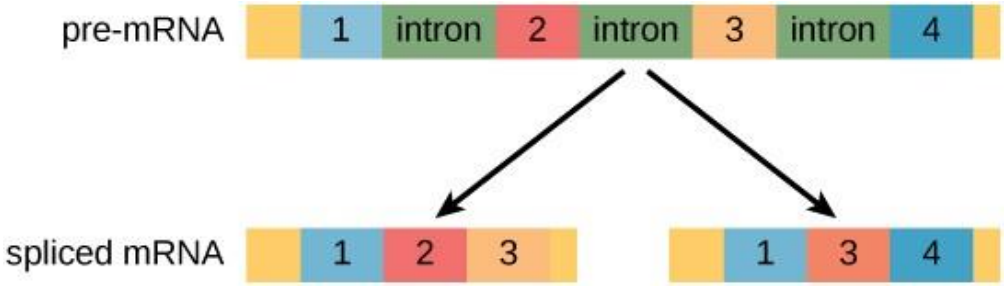


Figure 14. Avant qu'un ARN puisse être traduit, les introns doivent être éliminés par épissage. Le pré-ARNm peut être alternativement épissé pour créer différentes protéines **(5).**

6.2. Contrôle de la stabilité de l'ARN

Un autre type de contrôle post-transcriptionnel implique la stabilité de l'ARNm dans le cytoplasme. Plus un ARNm existe longtemps dans le cytoplasme, plus il doit être traduit et plus il produit de protéines. De nombreux facteurs contribuent à la stabilité de l'ARNm, notamment la longueur de sa queue poly-A (Figure 15).

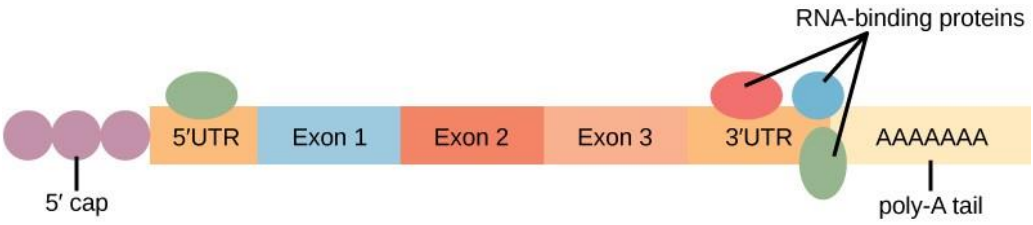


Figure 15. La région codant pour les protéines de l'ARNm est flanquée de régions non traduites (UTR) en 5 'et 3'. Les protéines de liaison à l'ARN au niveau de l'UTR 5 'ou 3' influencent la stabilité de la molécule d'ARN **(5).**

Les protéines, appelées protéines de liaison à l'ARN (RBP), peuvent se lier aux régions de l'ARN juste en amont ou en aval de la région codant pour la protéine. Ces régions de l'ARN qui ne sont pas traduites en protéine sont appelées régions non traduites ou UTR. La région juste avant la région codant pour la protéine est appelée 5′ UTR, tandis que la région après la région codante est appelée 3′ UTR (Figure 17.12). La liaison des RBP à ces régions peut augmenter ou diminuer la stabilité d'une molécule d'ARN, en fonction de la RBP spécifique qui se lie.

Les microARN, ou miARN, peuvent également se lier à la molécule d'ARN. Les miARN sont de courtes molécules d'ARN (21 à 24 nucléotides) qui sont fabriquées dans le noyau sous forme de pré-miARN plus longs, puis découpées en miARN matures par une protéine appelée dicer. Les miARN se lient à l'ARNm avec un complexe ribonucléoprotéique appelé complexe de

silençage induit par l'ARN (RISC). Le complexe RISC-miARN dégrade rapidement l'ARNm cible.

7. Contrôle post-transcroptionelles de l'expression génique

Une fois qu'un ARNm a été transporté dans le cytoplasme, il est traduit en protéines. Le contrôle de ce processus dépend largement de la molécule d'ARNm. La stabilité de l'ARNm aura un impact important sur sa traduction en une protéine. La traduction peut également être régulée au niveau de la liaison de l'ARNm au ribosome. Une fois l'ARNm lié au ribosome, la vitesse et le niveau de traduction peuvent encore être contrôlés. Un exemple de contrôle de la traduction se produit dans les protéines destinées à se retrouver dans un organite appelé le réticulum endoplasmique (ER). Les premiers acides aminés de ces protéines sont une étiquette appelée séquence signal. Dès que ces acides aminés sont traduits, une particule de reconnaissance de signal (SRP) se lie à la séquence signal et arrête la traduction tandis que le complexe ARNm-ribosome est acheminé vers le RE. Une fois arrivés, le SRP est supprimé et la traduction reprend.

10. Contrôle traductionnel de l'expression génique

La traduction de l'ARNm en protéine représente la dernière étape de la voie d'expression génique, qui médie la formation du protéome à partir de l'information génomique. La régulation de la traduction est un mécanisme utilisé pour moduler l'expression des gènes dans un large éventail de situations biologiques. Du développement embryonnaire précoce à la différenciation cellulaire et au métabolisme, la traduction est utilisée pour affiner les niveaux de protéines dans le temps et dans l'espace. Cependant, bien que de nombreux exemples aient été décrits, il reste encore beaucoup à apprendre sur les mécanismes moléculaires du contrôle traductionnel. Deux modes généraux de contrôle peuvent être envisagés : le contrôle global, dans lequel la traduction de la plupart des ARNm dans la cellule est régulée ; et un contrôle spécifique de l'ARNm, par lequel la traduction d'un groupe défini d'ARNm est modulée sans affecter la biosynthèse générale des protéines ou l'état de traduction du transcriptome cellulaire dans son ensemble. La régulation globale se produit principalement par la modification des facteurs d'initiation de la traduction, tandis que la régulation spécifique de l'ARNm est pilotée par des complexes protéiques régulateurs qui reconnaissent des éléments particuliers qui sont généralement présents dans les régions non traduites (UTR) 5 'et/ou 3' de l'ARNm cible. Récemment, il a été découvert que la traduction de l'ARNm peut également être régulée par de petits MICRO-ARN (miARN) qui s'hybrident à des séquences d'ARNm fréquemment situées dans l'UTR 3'.

Un cas particulier et extrêmement intéressant de régulation spécifique de l'ARNm est la régulation locale de la traduction qui se produit dans une cellule polarisée. La traduction d'ARNm spécifiques est limitée à des emplacements définis, tels que le pôle antérieur ou postérieur d'un ovocyte, ou une synapse neuronale spécifique. Le but de cette régulation est de générer des gradients de protéines qui émanent d'une position particulière dans la cellule, ou de restreindre l'expression des protéines à une petite région définie - par exemple, à une synapse. Bien qu'un tel contrôle local de la traduction implique presque invariablement des complexes régulateurs qui s'associent aux transcrits cibles, il pourrait également utiliser des changements locaux dans l'activité des facteurs généraux de traduction.

Les caractéristiques structurelles et les séquences régulatrices de l'ARNm sont responsables de son devenir traductionnel. Celles-ci incluent : les modifications canoniques des extrémités des molécules d'ARNm - la cap structure et la poly(a) tail - qui sont de puissants promoteurs de l'initiation de la traduction ; séquences internes d'entrée du ribosome (ires), qui interviennent dans l'initiation de la traduction indépendante de la coiffe ; cadres de lecture ouverts en amont (uORF), qui réduisent normalement la traduction à partir de l'ORF principal ; les structures d'ARN secondaires ou tertiaires, telles que les épingles à cheveux et les pseudoknots, qui bloquent généralement l'initiation, mais peuvent également faire partie des éléments IRES et donc favoriser la traduction indépendante de la coiffe ; et, des sites de liaison spécifiques pour les complexes régulateurs, qui sont des déterminants cruciaux de la traduction de l'ARNm. Bien qu'en principe, la régulation puisse activer ou réprimer la traduction, la plupart des mécanismes de régulation découverts jusqu'à présent sont inhibiteurs, ce qui implique que, à moins qu'un mécanisme de régulation ne soit imposé, les ARNm sont par défaut traductionnels actifs. Cependant, cela ne signifie pas que tous les ARNm non réprimés sont activement engagés avec les ribosomes, car l'activité des facteurs d'initiation de la traduction, en particulier ceux qui favorisent le recrutement de complexes ribosomiques qui initient la traduction, est souvent limitante. En conséquence, la plupart des ARNm sont répartis entre un pool activement traduit et un pool non traduit dans le cytoplasme des cellules, et des modifications de l'activité de ces facteurs limitant la traduction entraînent des modifications de la synthèse protéique globale (Figure 16).

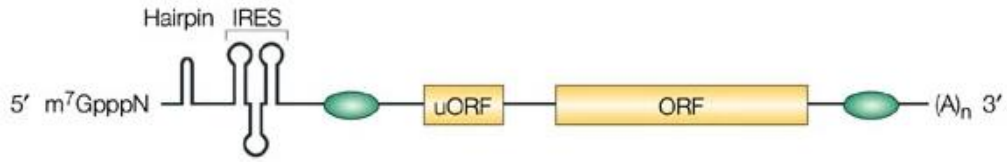


Figure 16. La structure de coiffe m7GpppN à l'extrémité 5 'de l'ARNm et la queue poly (A) ((A) n sur la figure) à l'extrémité 3 ', sont des motifs canoniques qui favorisent fortement l'initiation de la traduction. Les structures secondaires, telles que les épingles à cheveux, bloquent la traduction. Les séquences internes d'entrée des ribosomes (IRES) interviennent dans la traduction indépendante de la coiffe. Les cadres de lecture ouverts en amont (uORF) fonctionnent normalement comme des régulateurs négatifs en réduisant la traduction à partir de l'ORF principal. Les ovales verts symbolisent les sites de liaison pour les protéines et/ou les régulateurs d'ARN, qui inhibent généralement, mais parfois favorisent, la traduction **(46).**

11. Exemple de régulation artificielle des gènes

Les chercheurs ont mis au point une technique qui pourrait aider à affiner la production d'anticorps monoclonaux et d'autres protéines utiles.

Grâce à une approche basée sur les protéines CRISPR, les chercheurs ont développé une nouvelle façon de contrôler avec précision la quantité d'une protéine particulière qui est produite dans les cellules de mammifères. Cette technique pourrait être utilisée pour affiner la production de protéines utiles, telles que les anticorps monoclonaux utilisés pour traiter le cancer et d'autres maladies, ou d'autres aspects du comportement cellulaire. Les chercheurs ont montré que ce système peut fonctionner dans une variété de cellules de mammifères, avec des résultats très cohérents.

De nombreuses protéines thérapeutiques, y compris des anticorps monoclonaux, sont produites dans de grands bioréacteurs contenant des cellules de mammifères conçues pour générer la protéine souhaitée. Pour ce faire, les chercheurs ont ciblé les promoteurs des gènes qu'ils souhaitaient réguler positivement. Dans toutes les cellules de mammifères, les gènes ont une région promotrice qui se lie aux facteurs de transcription - des protéines qui initient la transcription du gène en ARN messager.

Le système conçu par les chercheurs comprend plusieurs composants. L'un est le gène à transcrire, ainsi qu'une séquence « opérateur », qui consiste en une série de sites de liaison de facteurs de transcription artificiels. Un autre composant est un ARN guide qui se lie à ces

séquences opératrices. Enfin, le système comprend également un domaine d'activation de la transcription attaché à une protéine Cas9 désactivée. Lorsque cette protéine Cas9 désactivée se lie à l'ARN guide sur le site du promoteur synthétique, le facteur de transcription basé sur CRISPR peut activer l'expression génique.

Les sites promoteurs utilisés pour ce système synthétique ont été conçus pour être distincts des sites promoteurs naturels, de sorte que le système n'affecte pas les gènes dans les propres génomes des cellules. Chaque opérateur comprend entre deux et 16 copies du site de liaison de l'ARN guide, et les chercheurs ont découvert que leur système pouvait initier la transcription génique à des taux qui correspondent linéairement au nombre de sites de liaison, leur permettant de contrôler avec précision la quantité de protéine produite.

Références

1. "Eukaryotic Transcriptional Regulation" by E. V. Wong, LibreTexts is licensed under CC BY-NC-SA
2. Berger, S. L. Histone modifications in transcriptional regulation [J]. Curr. Opin. Genet. 2002, Dev. 12, 142–148.
3. Bergtrom, Gerald, "Cell and Molecular Biology 4e: What We Know and How We Found Out" (2020). *Cell and Molecular Biology 4e: What We Know and How We Found Out – All Versions*. 13. https://dc.uwm.edu/biosci_facbooks_bergtrom/13
4. Bird, A. (2002). DNA methylation patterns and epigenetic memory. Genes Dev. 16, 6–21
5. Chapter 17. regulation of gene expression: https://rwu.pressbooks.pub/bio103/chapter/regulation-of-gene-expression/
6. Chasman DI, Flaherty KM, Sharp PA, Kornberg RD. Crystal structure of yeast TATA-binding protein and model for interaction with DNA. Proc. Natl Acad. Sci. USA. (1993);90:8174–8178.
7. Derrien, T.; Johnson, R.; Bussotti, G.; Tanzer, A.; Djebali, S.; Tilgner, H.; Guernec, G.; Martin, D.; Merkel, A.; Knowles, D.G.; et al. The gencode v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res.* **2012**, *22*, 1775–1789.
8. Ferguson-Smith, A.C. & Surani, M.A. (2001). Imprinting and the epigenetic asymmetry between parental genomes. Science 293, 1086–1089
9. Flatt, P.M. (2019) Biochemistry – Defining Life at the Molecular Level. Published by Western Oregon University, Monmouth, OR (CC BY-NC-SA). Available at: https://wou.edu/chemistry/courses/online-chemistry-textbooks/ch450-and-ch451-biochemistry-defining-life-at-the-molecular-level/?preview_id=4919&preview_nonce=cca8f0ce36&preview=true
10. Four Common Histone Modifications : https://www.cusabio.com/c-20829.html
11. Geisler, S.; Lojek, L.; Khalil, A.M.; Baker, K.E.; Coller, J. Decapping of long noncoding RNAs regulates inducible genes. *Mol. Cell* **2012**, *45*, 279–291.
12. Greer, E. L. & Shi, Y. Histone methylation: a dynamic mark in health, disease and inheritance [J]. Nat. Rev. Genet. 2012, 13, 343–57.

13. Guttman, M.; Amit, I.; Garber, M.; French, C.; Lin, M.F.; Feldser, D.; Huarte, M.; Zuk, O.; Carey, B.W.; Cassady, J.P.; et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **2009**, *458*, 223–227.
14. Hall, J.M.; McDonnell, D.P.; Korach, K.S. Allosteric regulation of estrogen receptor structure, function, and coactivator recruitment by different estrogen response elements. *Mol. Endocrinol.* **2002**, *16*, 469–486.
15. Hangauer, M.J.; Vaughn, I.W.; McManus, M.T. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS Genet.* **2013**, *9*, e1003569
16. Holliday, R. & Pugh, J.E. (1975). DNA modification mechanisms and gene activity during development. Science 187, 226–232
17. https://courses.lumenlearning.com/suny-biology1/chapter/eukaryotic-transcription/
18. Huang, P.; Chandra, V.; Rastinejad, F. Structural overview of the nuclear receptor superfamily: Insights into physiology and therapeutics. *Annu. Rev. Physiol.* **2010**, *72*, 247–272.
19. Jähner, D. et al. (1982) De novo methylation and expression of retroviral genomes during mouse embryogenesis. Nature 298, 623–628
20. Küpper, H.; Sekiya, T.; Rosenberg, M.; Egan, J.; Landy, A. A Rho-dependent termination site in the gene coding for tyrosine tRNA su3 of *Escherichia Coli*. *Nature* **1978**, *272*, 423–428.
21. Lee WK., Lee Kraus K. (2001). Nuclear receptors, coactivators and chromatin: new approaches, new insights.12: 191-197
22. Li, E. (2002). Chromatin modification and epigenetic reprogramming in mammalian development. Nat. Rev. Genet. 3, 662–673
23. Ma MK, Heath C, *et al*. Histone crosstalk directed by H2B ubiquitination is required for chromatin boundary integrity [J]. PLoS Genet. 2011 Jul; 7(7):e1002175.
24. Mangelsdorf, D.J.; Evans, R.M. The RXR heterodimers and orphan receptors. *Cell* **1995**, *83*, 841–850.
25. Mangelsdorf, D.J.; Thummel, C.; Beato, M.; Herrlich, P.; Schütz, G.; Umesono, K.; Blumberg, B.; Kastner, P.; Mark, M.; Chambon, P.; et al. The nuclear receptor superfamily: The second decade. *Cell* **1995**, *83*, 835–839.
26. Nakamura K, Kato A, *et al*. Regulation of homologous recombination by RNF20-dependent H2B ubiquitination [J]. Mol Cell. 2011, Mar 4; 41(5):515-28.
27. Niemczyk, M.; Ito, Y.; Huddleston, J.; Git, A.; Abu-Amero, S.; Caldas, C.; Moore, G.E.; Stojic, L.; Murrell, A. Imprinted chromatin around DIRAS3 regulates alternative splicing of GNG12-AS1, a long noncoding RNA. *Am. J. Hum. Genet.* **2013**, *93*, 224–235.
28. Pieterman C R C, Conemans E B,, K M A Dreijerink , J M de Laat1 , H Th M Timmers , M R Vriens and G D Valk. (2014). Thoracic and duodenopancreatic neuroendocrine tumors in multiple endocrine neoplasia type 1: Natural history and function of menin in tumorigenesis. Endocrine-Related Cancer. 21, R121–R142
29. Pons D, de Vries FR, van den Elsen PJ, Heijmans BT, Quax PH, Jukema JW. Epigenetic histone acetylation modifiers in vascular remodelling: new targets for therapy in cardiovascular disease. Eur Heart J. 2009; 30(3): 266- 277.
30. Porter BA., Ortiz MA., Bratslavsky G., Kotula L. (2019). Structure and Function of the Nuclear Receptor Superfamily and Current Targeted Therapies of Prostate Cancer. Cancers 2019, 11(12), 1852
31. Rajan PK , Udoh UA , Sanabria JD , Banerje M et al., (2020) The Role of Histone Acetylation-/Methylation-Mediated Apoptotic Gene Regulation in Hepatocellular Carcinoma. Int. J. Mol. Sci. 21, 8894

32. Regulation Of Gene Expression: Overview In Prokaryotes And Eukaryotes, Diagram, Lac Operon https://www.embibe.com/exams/regulation-of-gene-expression/

33. Riggs, A.D. X. (1975). inactivation, differentiation, and DNA methylation. *Cytogenet. Cell Genet.* **14**, 9–25
34. Shaffer, P.L.; Jivan, A.; Dollins, D.E.; Claessens, F.; Gewirth, D.T. Structural basis of androgen receptor binding to selective androgen response elements. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 4758–4763.
35. Tan, M.H.; Li, J.; Xu, H.E.; Melcher, K.; Yong, E.L. Androgen receptor: Structure, role in prostate cancer and drug discovery. *Acta Pharmacol. Sin.* **2015**, *36*, 3–23.
36. The medical biochemistry page: https://themedicalbiochemistrypage.org/signal-transduction-pathways-overview/
37. Uesaka, M.; Nishimura, O.; Go, Y.; Nakashima, K.; Agata, K.; Imamura, T. Bidirectional promoters are the major source of gene activation-associated non-coding rnas in mammals. *BMC Genom.* **2014**, *15*, 35.
38. Umesono, K.; Evans, R.M. Determinants of target gene specificity for steroid/thyroid hormone receptors. *Cell* **1989**, *57*, 1139–1146.
39. Urnov, F.D. & Wolffe, A.P. (2001). Above and within the genome: epigenetics past and present. J. Mammary Gland Biol. Neoplasia 6, 153–167
40. Wang L. Cheung, Kozo Ajiro, *et al*. Apoptotic Phosphorylation of Histone H2B Is Mediated by Mammalian Sterile Twenty Kinase [J]. Cell. 2003, 113, 507–517.
41. Wolffe, A.P. & Matzke, M.A. (1999). Epigenetics: regulation through repression. Science 286, 481–486
42. Works contributed to LibreTexts by Kevin Ahern and Indira Rajagopal. LibreTexts content is licensed by CC BY-NC-SA 3.0. The entire textbook is available for free from the authors at http://biochem.science.oregonstate.edu/content/biochemistry-free-and-easy

43. Ye Zhang, Karen Griffin, *et al*. Phosphorylation of Histone H2A Inhibits Transcription on Chromatin Templates [J]. J Biol Chem. 2004, 279(21):21866-72.
44. Zheng, G.X.; Do, B.T.; Webster, D.E.; Khavari, P.A.; Chang, H.Y. Dicer-microRNA-Myc circuit promotes transcription of hundreds of long noncoding RNAs. *Nat. Struct. Mol. Biol.* **2014**, *21*, 585–590.
45. Zhou W., Zhu P., *et al*. Histone H2A monoubiquitination represses transcription by inhibiting RNA polymerase II transcriptional elongation [J]. 2008, Mol. Cell 29, 69–80.
46. Gebauer, F., Hentze, M. Molecular mechanisms of translational control. *Nat Rev Mol Cell Biol* **5**, 827–835 (2004). https://doi.org/10.1038/nrm1488
47. A new control system for synthetic genes : https://news.mit.edu/2022/synthetic-gene-expression-control-1101.

Chapitre 03

Variabilité des génomes : Le polymorphisme

Chapitre 3 : Variabilité des génomes : Polymorphisme

4. Introduction

Dans des études antérieures, des marqueurs morphologiques et des facteurs écogéographiques ont été utilisés pour représenter la diversité, et après cela, le caryotype chromosomique a été développé. Avec le développement rapide de la biotechnologie moderne, des marqueurs biochimiques, tels que les protéines et les isozymes, ont été utilisés. Dans les années 1980, de nombreux types différents de marqueurs moléculaires d'ADN avaient été explorés, par ex. Polymorphisme de longueur de fragment de restriction (RFLP), ADN polymorphe amplifié aléatoire (RAPD), polymorphisme de longueur de fragment amplifié (AFLP), polymorphisme de conformation monocaténaire (SSCP) et ADN microsatellite. Tous ces marqueurs à base d'ADN présentent des avantages spécifiques et ont joué un rôle important dans l'évaluation de la diversité génétique chez les animaux de ferme. De plus, avec les innovations biotechnologiques et informatiques, de nouvelles stratégies telles que les puces SNP du génome entier et le codage à barres ADN ont émergé. À l'heure actuelle, les techniques de marqueurs moléculaires d'ADN sont largement appliquées dans les domaines de l'identification du germoplasme, de la phylogénétique et de l'analyse structurale génétique. Ils surmontent les limites des marqueurs morphologiques, cytologiques et biochimiques, à savoir le petit nombre de ces marqueurs et le fait qu'ils peuvent être influencés par l'environnement. L'expansion de l'information sur l'ADN facilitera l'étude de la diversité à l'échelle du génome ; ces informations sont beaucoup plus précises pour l'évaluation de la diversité génétique que les marqueurs précédents.

5. Sources de variations : les mutations

La mutation fait référence à un changement héréditaire soudain dans le phénotype d'un individu. Dans le terme moléculaire, la mutation est définie comme le changement permanent et relativement rare du nombre ou de la séquence des nucléotides. La mutation a été découverte pour la première fois par Wright en 1791 chez l'agneau mâle aux pattes courtes. Plus tard, une mutation a été rapportée par Hugo de Vries en 1900 chez Oenothera, Morgan (1910) chez Drosophila (mutant œil blanc) et plusieurs autres chez divers organismes. Le terme mutation a été inventé par de Vries.

5.1.Types de mutation

5.1.1. Substitution

Une substitution est une mutation dans laquelle il y a un échange entre deux bases (c'est-à-dire un changement dans une seule "lettre chimique" comme le passage d'un T à un C). Une telle substitution pourrait changer un codon en un codant pour un acide aminé différent et provoquer un changement dans la protéine produite. Parfois, les substitutions peuvent ne pas affecter la structure de la protéine, de telles mutations sont appelées mutations silencieuses et parfois elles peuvent changer un codon codant pour un acide aminé en un seul codon "stop" et provoquer une protéine incomplète. Cela peut sérieusement affecter la structure des protéines qui peut complètement changer l'organisme.

Exemple de mutation de substitution : L'anémie falciforme est causée par une mutation de substitution, où le codon (GAG mute en --> GTG) et conduit à un changement (Glu --> Val).

5.1.2. Insertion

Les insertions sont des mutations dans lesquelles des paires de bases supplémentaires sont insérées dans un nouvel endroit de l'ADN. Le nombre de paires de bases insérées peut aller de un à des milliers ! Exemple de mutation d'insertion : la maladie de Huntington et le syndrome du X fragile sont des exemples de mutation d'insertion dans laquelle des répétitions de trinucléotides sont insérées dans la séquence d'ADN conduisant à ces maladies.

5.1.3. Suppressions

Les délétions sont des mutations dans lesquelles une section d'ADN est perdue ou supprimée. Le nombre de paires de bases supprimées peut à nouveau aller de un à des milliers ! Les mutations d'insertion et de suppression sont souvent appelées ensemble INDELS.

Exemple de mutation par délétion : le syndrome de délétion 22q11.2 est causé par la suppression de certaines bases du chromosome 22. Cette maladie se caractérise par une fente palatine, des malformations cardiaques, des troubles auto-immuns, etc.

5.1.4. Décalage de cadre

L'ADN codant pour les protéines est divisé en codons longs de trois bases. Les insertions et les délétions dans ces codons peuvent complètement modifier un gène, de sorte que son message ne peut pas être décodé correctement. De telles mutations sont appelées mutations de décalage de cadre. Par exemple, considérez la phrase "Le chat a mangé son rat". Chaque mot représente

un codon. Si nous supprimons la première lettre et lisons la phrase de la même manière, cela n'a pas de sens. De même, si les codons se mélangent, ils n'auraient plus de sens, dans de tels décalages de cadre, une erreur similaire se produit au niveau de l'ADN, où les codons ne peuvent pas être analysés correctement. Cela donne généralement lieu à des protéines tronquées qui sont aussi inutiles et pas informatif.

Exemples de mutation Frameshift : la maladie de Tay-Sachs, plusieurs types de cancers, la maladie de Crohn, la fibrose kystique ont été associés à la mutation Frameshift.

5.2. Les effets des mutations sur les gènes

De nombreuses mutations entraînent des modifications de la séquence nucléotidique qui n'ont aucun effet sur le fonctionnement du génome. Ces mutations silencieuses comprennent pratiquement toutes celles qui se produisent dans l'ADN intergénique et dans les composants non codants des gènes et des séquences liées aux gènes. En d'autres termes, environ 98,5 % du génome humain peut être muté sans effet significatif.

Les mutations dans les régions codantes des gènes sont beaucoup plus importantes. Premièrement, nous examinerons les mutations ponctuelles qui modifient la séquence d'un codon triplet. Une mutation de ce type aura l'un des quatre effets suivants (**Figure 1**) :

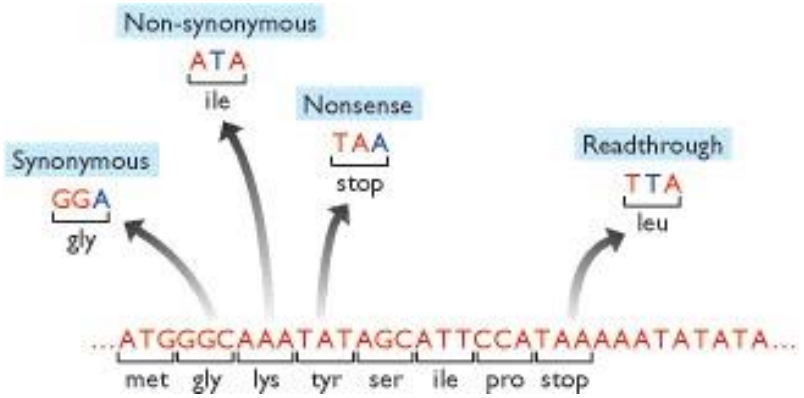


Figure 1. Effets des mutations ponctuelles sur la région codante d'un gène **(34)**

Il peut en résulter un changement synonyme, le nouveau codon spécifiant le même acide aminé que le codon non muté. Un changement synonyme est donc une mutation silencieuse car il n'a aucun effet sur la fonction de codage du génome : le gène muté code exactement pour la même protéine que le gène non muté.

Cela peut entraîner un changement non synonyme, la mutation modifiant le codon de sorte qu'il spécifie un acide aminé différent. La protéine codée par le gène muté présente donc un

seul changement d'acide aminé. Cela n'a souvent aucun effet significatif sur l'activité biologique de la protéine car la plupart des protéines peuvent tolérer au moins quelques modifications d'acides aminés sans effet notable sur leur capacité à fonctionner dans la cellule, mais des modifications de certains acides aminés, tels que ceux au niveau actif. site d'une enzyme, ont un impact plus important. Un changement non synonyme est également appelé mutation faux-sens.

La mutation peut convertir un codon qui spécifie un acide aminé en un codon de terminaison. Il s'agit d'une mutation non-sens et il en résulte une protéine raccourcie car la traduction de l'ARNm s'arrête à ce nouveau codon de terminaison plutôt que de passer au codon de terminaison correct plus en aval. L'effet de ceci sur l'activité protéique dépend de la quantité de polypeptide perdue : généralement, l'effet est drastique et la protéine est non fonctionnelle.

La mutation pourrait convertir un codon de terminaison en un codon spécifiant un acide aminé, entraînant la lecture du signal d'arrêt de sorte que la protéine est étendue par une série supplémentaire d'acides aminés à son extrémité C-terminale. La plupart des protéines peuvent tolérer de courtes extensions sans effet sur la fonction, mais des extensions plus longues peuvent interférer avec le repliement de la protéine et ainsi entraîner une activité réduite.

Les mutations de délétion et d'insertion ont également des effets distincts sur les capacités de codage des gènes (**Figure 2**). Si le nombre de nucléotides délétés ou insérés est de trois ou un multiple de trois, alors un ou plusieurs codons sont retirés ou ajoutés, la perte ou le gain résultant d'acides aminés ayant des effets variables sur la fonction de la protéine codée. Les délétions ou les insertions de ce type sont souvent sans conséquence mais auront un impact si, par exemple, des acides aminés impliqués dans le site actif d'une enzyme sont perdus, ou si une insertion perturbe une structure secondaire importante de la protéine. En revanche, si le nombre de nucléotides supprimés ou insérés n'est pas de trois ou un multiple de trois, il en résulte un décalage de cadre, tous les codons en aval de la mutation étant prélevés sur un cadre de lecture différent de celui utilisé dans le gène non muté.

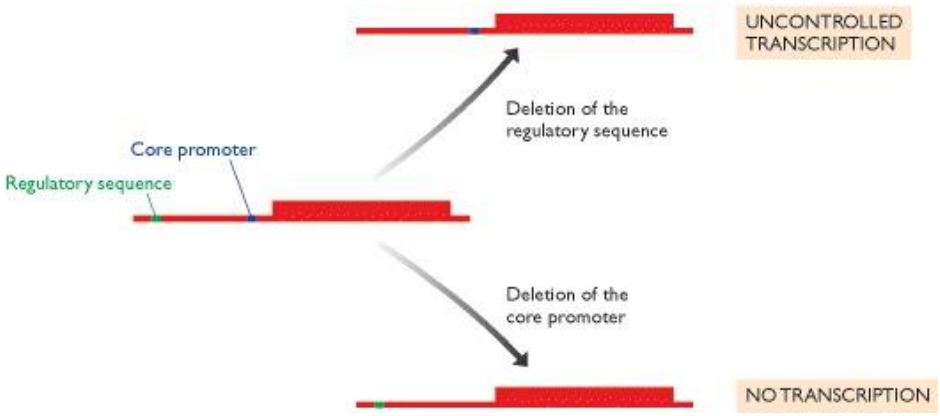


Figure 2. Deux effets possibles des mutations par délétion dans la région en amont d'un gène **(34)**

Un domaine qui a fait l'objet de meilleures recherches concerne les mutations qui se produisent dans les introns ou aux frontières intron-exon. Dans ces régions, les mutations ponctuelles seront importantes si elles modifient les nucléotides impliqués dans les interactions ARN-protéine et ARN-ARN qui se produisent lors de l'épissage de différents types d'intron. Par exemple, la mutation du G ou du T dans la copie d'ADN du site d'épissage 5 'd'un intron GU-AG, ou de l'A ou du G au site d'épissage 3', perturbera l'épissage car la limite intron-exon correcte ne sera plus reconnu. Cela peut signifier que l'intron n'est pas retiré du pré-ARNm, mais il est plus probable qu'un site d'épissage cryptique sera utilisé comme alternative. Il est également possible qu'une mutation au sein d'un intron ou d'un exon crée un nouveau site cryptique préféré à un véritable site d'épissage non lui-même muté. Les deux types d'événements ont le même résultat : relocalisation du site d'épissage actif, conduisant à un épissage aberrant. Cela peut supprimer une partie de la protéine résultante, ajouter une nouvelle séquence d'acides aminés ou entraîner un décalage de cadre. Plusieurs versions de la maladie du sang β-thalassémie sont causées par des mutations qui conduisent à la sélection du site d'épissage cryptique lors du traitement des transcrits de la β-globine.

5.3.Les effets des mutations sur les organismes multicellulaires

Passons maintenant aux effets indirects que les mutations ont sur les organismes, à commencer par les eucaryotes diploïdes multicellulaires tels que les humains. La première question à considérer est l'importance relative d'une même mutation dans une cellule somatique par rapport à une cellule germinale. Étant donné que les cellules somatiques ne transmettent pas de copies de leurs génomes à la génération suivante, une mutation de cellule somatique n'est importante que pour l'organisme dans lequel elle se produit : elle n'a aucun impact potentiel sur l'évolution. En fait, la plupart des mutations des cellules somatiques n'ont pas d'effet significatif,

même si elles entraînent la mort cellulaire, car il existe de nombreuses autres cellules identiques dans le même tissu et la perte d'une cellule est sans importance. Une exception est lorsqu'une mutation provoque un dysfonctionnement d'une cellule somatique d'une manière nuisible à l'organisme, par exemple en induisant la formation de tumeurs ou une autre activité cancéreuse.

Les mutations dans les cellules germinales sont plus importantes car elles peuvent être transmises aux membres de la génération suivante et seront alors présentes dans toutes les cellules de tout individu qui hérite de la mutation. La plupart des mutations, y compris toutes les mutations silencieuses et de nombreuses dans les régions codantes, ne modifieront toujours pas le phénotype de l'organisme de manière significative. Ceux qui ont un effet peuvent être divisés en deux catégories :

La perte de fonction est le résultat normal d'une mutation qui réduit ou abolit une activité protéique. La plupart des mutations avec perte de fonction sont récessives (Section 5.2.3), car chez un hétérozygote, la deuxième copie chromosomique porte une version non mutée du gène codant pour une protéine entièrement fonctionnelle dont la présence compense l'effet de la mutation (Figure 3). Il existe quelques exceptions où une mutation de perte de fonction est dominante, un exemple étant l'haploinsuffisance , où l'organisme est incapable de tolérer la réduction d'environ 50% de l'activité protéique subie par l'hétérozygote. C'est l'explication de quelques maladies génétiques chez l'homme, dont le syndrome de Marfan qui résulte d'une mutation du gène de la protéine du tissu conjonctif appelée fibrilline.

Les mutations de gain de fonction sont beaucoup moins fréquentes. La mutation doit être celle qui confère une activité anormale à une protéine. De nombreuses mutations de gain de fonction se trouvent dans des séquences régulatrices plutôt que dans des régions codantes, et peuvent donc avoir un certain nombre de conséquences. Par exemple, une mutation peut entraîner l'expression d'un ou plusieurs gènes dans les mauvais tissus, ces tissus acquérant des fonctions qui leur manquent normalement. Alternativement, la mutation pourrait conduire à la surexpression d'un ou plusieurs gènes impliqués dans le contrôle du cycle cellulaire, conduisant ainsi à une division cellulaire incontrôlée et donc au cancer. En raison de leur nature, les mutations de gain de fonction sont généralement dominantes.

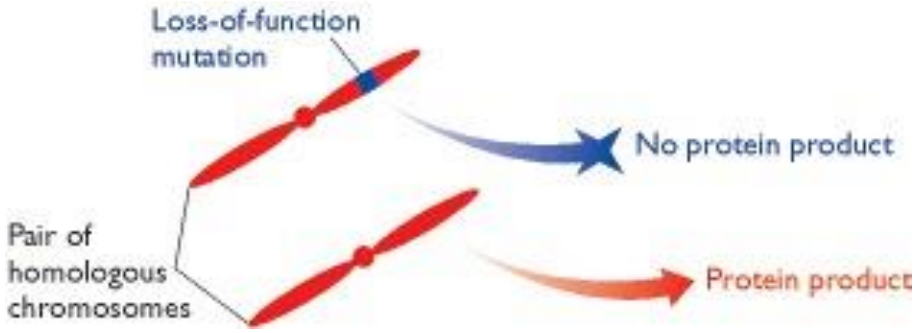


Figure 3. Une mutation avec perte de fonction est généralement récessive car une version fonctionnelle du gène est présente sur la deuxième copie du chromosome **(34)**

L'évaluation des effets des mutations sur les phénotypes d'organismes multicellulaires peut être difficile. Toutes les mutations n'ont pas un impact immédiat : certaines sont d'apparition tardive et ne confèrent un phénotype altéré que plus tard dans la vie de l'individu. D'autres affichent une non-pénétrance chez certains individus, ne s'exprimant jamais même si l'individu a une mutation dominante ou est un homozygote récessif. Chez l'homme, ces facteurs compliquent les tentatives de cartographie des mutations pathogènes par analyse généalogique car ils introduisent une incertitude quant aux membres d'une généalogie porteurs d'un allèle mutant.

6. Phénotypage moléculaire : Les marqueurs

6.1.Marqueurs morphologiques

Les marqueurs morphologiques se réfèrent normalement aux caractéristiques animales externes (c'est-à-dire la couleur du pelage, la forme du corps, la structure de la peau et les caractéristiques anatomiques), qui peuvent être obtenues par observation visuelle directe et mesure. Ils sont utilisés dans l'identification, la classification et la caractérisation de l'évolution génétique de différentes espèces ou populations. Cependant, le phénotype d'un animal est déterminé par son patrimoine génétique et l'environnement qu'il vit. L'évaluation des ressources génétiques des animaux d'élevage au moyen de marqueurs morphologiques est basée sur des jugements et des descriptions subjectifs, et les conclusions atteintes ne sont souvent pas tout à fait exactes. De plus, la mesure et l'identification des traits morphologiques des animaux prennent généralement beaucoup de temps et il n'est pas facile d'éliminer les effets des facteurs environnementaux. Par conséquent, l'application des marqueurs morphologiques est limitée dans l'évaluation des caractères quantitatifs. Cependant, il s'agit toujours d'une méthode efficace pour l'évaluation des traits qualitatifs, pour laquelle il est facile de caractériser les différences phénotypiques entre les individus par l'observation et la mesure directes.

6.2.Marqueurs cytologiques

Des marqueurs cytologiques ont été utilisés pour l'évaluation des ressources génétiques des animaux d'élevage sur la base du nombre et de la morphologie des chromosomes animaux. Les marqueurs cytologiques comprennent les caryotypes chromosomiques, les bandes, les répétitions, les délétions, les translocations et les inversions. Les chromosomes sont les porteurs de matériel génétique et les mutations chromosomiques sont des sources cruciales de variation génétique, nous pouvons utiliser ces mutations comme marqueurs pour déterminer l'emplacement spécifique d'un gène sur le chromosome et sa position par rapport aux autres gènes. Par exemple, les chercheurs peuvent retracer les origines et l'histoire évolutive du bétail et évaluer la diversité génétique des animaux domestiques en comparant le nombre et la structure des chromosomes entre les animaux domestiques et leurs ancêtres sauvages.

6.3.Marqueurs biochimiques

Marqueurs biochimiques, par ex. le groupe sanguin et les isoenzymes, représentent des traits biochimiques et peuvent être analysés par électrophorèse des protéines. Les chercheurs ont étudié la variation génétique au sein des espèces et les relations phylogénétiques entre les espèces par des différences dans la composition en acides aminés des isozymes et des protéines solubles. Néanmoins, ni les protéines ni les isozymes ne sont du matériel génétique mais les produits de l'expression génique, et ils sont vulnérables aux impacts environnementaux et aux écarts de croissance individuels, ce qui limite l'étendue de leur application. A l'inverse, l'électrophorèse des protéines est une technique rapide, économique et simple et fournit une représentation plus détaillée des polymorphismes que les marqueurs morphologiques ou cytologiques ; ainsi, il est encore largement utilisé pour élucider l'origine et la classification des espèces.

6.4.Marqueurs moléculaires (marqueurs à base d'ADN)

Avec le développement de la biotechnologie moléculaire, les marqueurs moléculaires ont fait des progrès rapides. Un marqueur moléculaire est basé sur les mutations de la séquence nucléotidique dans le génome de l'individu ; ce sont les marqueurs les plus fiables disponibles. Les marqueurs moléculaires peuvent être utilisés pour étudier les variations génétiques au niveau de l'ADN entre différentes populations et individus ; son avantage est de pouvoir trouver rapidement et directement des variations génétiques. Les marqueurs moléculaires se sont développés rapidement et deviennent de plus en plus informatifs. Jusqu'à présent, divers types de marqueurs moléculaires ont été utilisés pour évaluer les polymorphismes de l'ADN, par ex.

RFLP. La réaction en chaîne par polymérase (PCR) peut amplifier de manière exponentielle un fragment d'ADN in vitro, et depuis son invention, une série de techniques ont émergé en combinaison avec la PCR, par ex. PCR—RFLP, AFLP, répétitions de séquences simples (SSR) et polymorphismes de nucléotide unique (SNP). Dans cette revue, nous nous concentrons principalement sur l'introduction de plusieurs marqueurs importants basés sur l'ADN et leurs diverses applications dans la caractérisation des ressources génétiques animales.

6.4.1. Marqueurs RFLP (Restriction fragment length polymorphism)

RFLP est une méthode établie en 1974, il est utilisé pour identifier les polymorphismes de l'ADN chez différents individus. Son principe de base est le suivant : premièrement, l'ADN génomique de différents individus est digéré en fragments d'ADN de taille variable, à l'aide d'enzymes de restriction connues. Deuxièmement, les fragments digérés sont séparés par analyse électrophorétique. Enfin, les fragments séparés sont hybridés avec des sondes homologues radioactives ou chimioluminescentes et exposés à un film radiographique ; les différents fragments sont visibles par autoradiographie. La base moléculaire du RFLP est que les substitutions, les insertions, les délétions, les duplications et les inversions de bases de nucléotides dans l'ensemble du génome peuvent supprimer ou créer de nouveaux sites de restriction.

RFLP a été le premier marqueur basé sur l'ADN pour la construction de cartes de liaison génétique ; c'est également l'un des marqueurs les plus largement utilisés dans les évaluations des ressources zoogénétiques et le développement de programmes de sélection. En combinant cette méthode avec la PCR (PCR-RFLP), les chercheurs ont détecté quatre nouveaux polymorphismes génétiques dans le gène de la leptine de différentes races porcines. Les principaux avantages des RFLP incluent :

- 1) une grande fiabilité, car ils sont générés à partir de sites spécifiques via des enzymes de restriction connues et les résultats sont constants dans le temps et dans l'espace.
- 2) Co-dominance, ce qui signifie que les chercheurs sont capables de distinguer les hétérozygotes des homozygotes.
- 3) La neutralité sélective fait référence à une situation dans laquelle différents allèles d'un certain gène confèrent une aptitude égale.

Les inconvénients des RFLP sont les suivants :

- 1) travail intensif et chronophage.

- 2) Les RFLP ne peuvent vérifier que des mutations spécifiques au niveau des sites de coupure enzymatique, ce qui limite l'identification de la variation du génome entier chez les animaux.
- 3) Le polymorphisme des marqueurs RFLP est relativement faible et doit être détecté par radio-isotope, ce qui limite son application.

6.4.1.1.Principe du RFLP

En utilisant des endonucléases de restriction, des fragments d'ADN sont obtenus et le fragment souhaité est détecté en utilisant des sondes de restriction. L'hybridation Southern ou le transfert Southern, utilisant des endonucléases de restriction pour l'isolement de la longueur souhaitée des fragments d'ADN, est un exemple de RFLP

6.4.1.2. Les étapes de l'AFLP

- **Résumé des restrictions :** Extraction de fragments d'ADN après digestion de l'ADN génomique par des endonucléases de restriction (RE). RE a des sites de restriction spécifiques sur le brin d'ADN, de sorte qu'il coupe ou coupe l'ADN en fragments. Différentes tailles de fragments sont générées avec les fragments spécifiques souhaités.
- **Électrophorèse sur gel :** L'électrophorèse sur gel de polyacrylamide ou l'électrophorèse sur gel d'agarose peuvent être utilisées pour séparer les fragments sur la base de leur longueur, c'est-à-dire leur taille ou leur poids moléculaire.
- Différents fragments forment différentes bandes en fonction de leur taille.
- **Dénaturation:** Le gel avec les bandes est placé dans une solution d'hydroxyde de sodium (NaOH) pour dénaturation, de sorte que les fragments d'ADN double brin deviennent simple brin.
- **Blotting :** L'ADN simple brin obtenu est transféré dans une membrane de charge, c'est-à-dire du papier de nitrocellulose par électro-transfert ou transfert capillaire.
- **Cuisson et blocage :** Le papier de nitrocellulose avec l'ADN transféré est fixé par autoclavage.
- Cette membrane est ensuite bloquée en utilisant de l'albumine de sérum bovin ou de la caséine pour empêcher la liaison de la sonde marquée de manière non spécifique à la membrane chargée.
- **Hybridation et visualisation :** La sonde RFLP marquée est autorisée à s'apparier avec l'ADN simple brin complémentaire sur le papier de nitrocellulose, le processus appelé hybridation.

- Les sondes RFLP sont marquées avec des isotopes radioactifs afin qu'elles forment des bandes de couleur sous visualisation par autoradiographie.

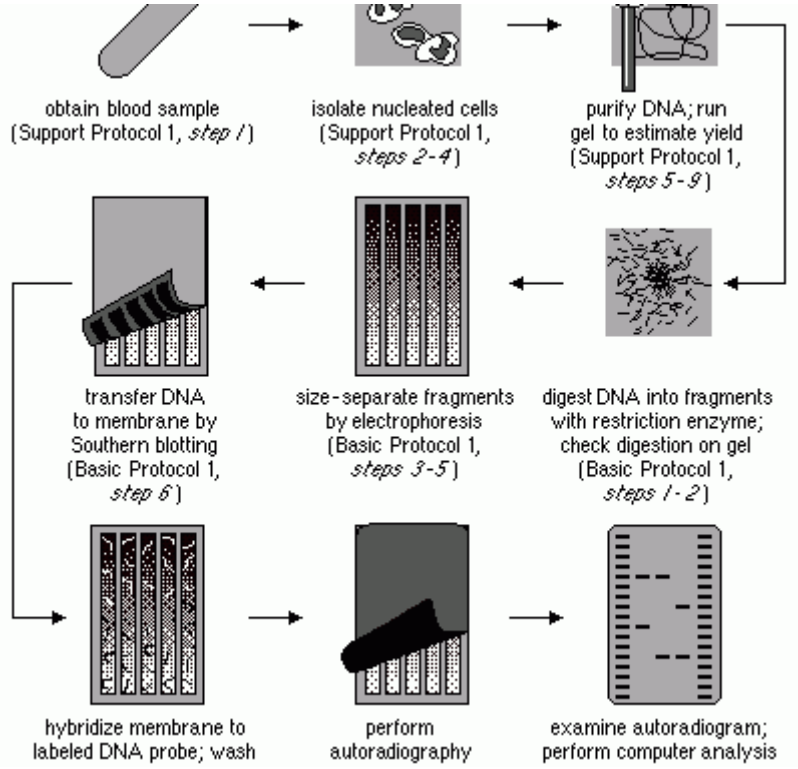


Figure 4. Etapes de la technique RFLP **(29)**

6.4.1.3.Applications du RFLP

Il aide à l'analyse du modèle unique du génome des organismes vivants pour leur identification et leur différenciation.

1. Cartographie du génome : Le taux de recombinaison dans les loci entre les sites de restriction peut être déterminé.

2.Analyse des maladies génétiques : Après identification du gène d'une maladie génétique ou héréditaire particulière, ce gène peut être analysé parmi d'autres membres de la famille.

3. Détection du gène muté

4. Empreintes digitales ADN ou profilage ADN ou typage ADN (test médico-légal) : L'empreinte ADN, l'une des techniques de test de paternité, d'identification criminelle, etc. est basée sur le RFLP.

2.4.2. Marqueurs RAPD (Rapid amplified polymorphic DNA)

Le RAPD a été développé par des scientifiques américains en 1990. Il amplifie l'ADN génomique cible avec des amorces courtes et arbitraires (généralement 10 pb) dans une réaction PCR, et peut être utilisé pour produire des profils d'ADN relativement compliqués pour détecter les polymorphismes de longueur de fragment amplifiés entre les organismes. Étant donné que les amorces arbitraires complètent différentes parties de l'ADN génomique, les produits de PCR différeront en nombre et en taille (polymorphisme).

Les empreintes digitales RAPD-PCR ont été utilisées avec succès pour définir la diversité génétique entre différentes espèces. Par exemple, la méthode RAPD a été utilisée pour générer des empreintes digitales spécifiques de dix espèces différentes : sanglier, porc, cheval, buffle, bœuf, venaison, chien, chat, lapin et kangourou.

Les marqueurs RAPD ont plusieurs caractéristiques évidentes telles que résumées dans la littérature :

- 1) aucune connaissance préalable de la séquence n'est nécessaire pour concevoir les amorces spécifiques, qui peuvent ensuite être utilisées dans différentes matrices.
- 2) La quantité d'ADN nécessaire est très faible car elle sera amplifiée par PCR.
- 3) Les RAPD sont simples, rapides et rentables par rapport aux RFLP

Cependant, les RAPD présentent également certains inconvénients, notamment

- 1) la répétabilité et la fiabilité des profils polymorphes RAPD sont médiocres.
- 2) Une certaine liaison non spécifique et donc non reproductible des amorces se produit.
- 3) Les RAPD sont des marqueurs génétiques dominants qui ne peuvent pas être utilisés pour distinguer les génotypes homozygotes des hétérozygotes dans les populations F2.

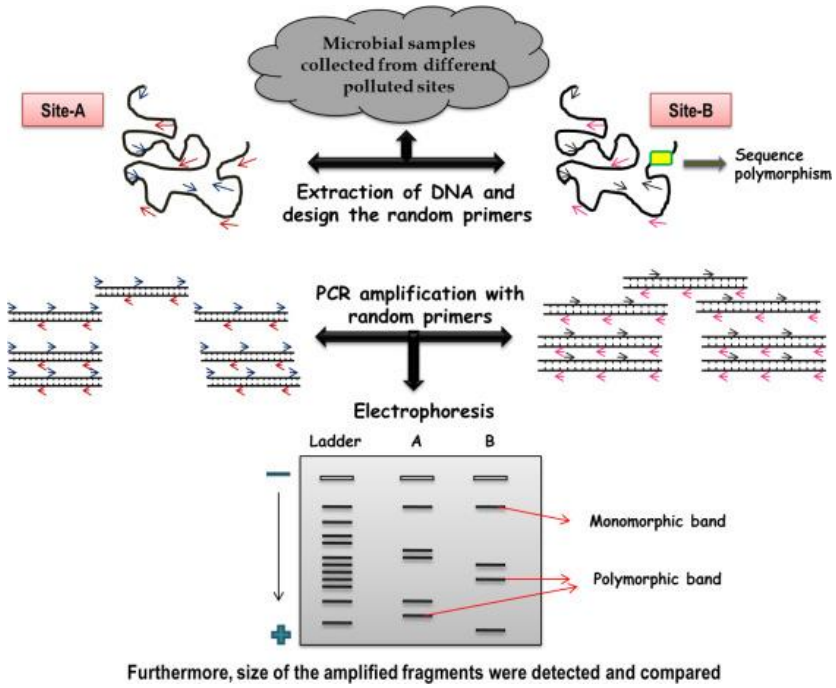


Figure 5. Illustration schématique de l'analyse aléatoire du polymorphisme de longueur amplifié **(27)**

2.4.2.2. Principe du RAPD

L'amplification conduit à des modèles de PCR multi-fragments et permet de cribler le polymorphisme dans plusieurs régions simultanément, résultant de la variation de la séquence d'ADN au niveau et entre les sites de liaison aux amorces.

L'utilisation d'amorces universelles (non spécifiques à une espèce) convient en l'absence d'informations sur le génome de l'espèce étudiée. Cependant, certaines préoccupations concernant la reproductibilité de l'analyse RAPD au sein et entre les laboratoires peuvent rester non résolues, principalement en raison de la faible température de recuit liée à la longueur limitée des amorces. Les marqueurs RAPD peuvent être convertis en SCAR. Les marqueurs SCAR sont de courtes séquences uniques qui identifient un locus d'ADN spécifique amplifié par PCR.

2.4.2.1. Avantages

- Les marqueurs RAPD sont abondamment présents dans le génome et sont répartis dans tout le génome.
- Aucune information sur la séquence précédente n'est nécessaire.
- Il peut amplifier une faible quantité d'ADN.

2.4.2.2. Désavantages

- Le marqueur n'est pas spécifique au locus.
- La sensibilité du RAPD est également plus faible.
- La reproductibilité est également très faible.
- Les homozygotes et les hétérozygotes ne sont pas faciles à distinguer.
- Les résultats RAPD sont difficiles à interpréter car la courte amorce peut amplifier n'importe laquelle des séquences aléatoires présentées dans le génome.

2.4.4. Marqueurs AFLP (Amplified fragment length polymorphism)

La technique AFLP a été développé par Zabeau et Vos en 1993 ; c'est une combinaison des techniques RFLP et PCR. La procédure AFLP est la suivante : d'abord, l'ADN génomique est digéré avec une enzyme de restriction, puis les fragments digérés sont ligaturés à des adaptateurs synthétiques et amplifiés avec des amorces spécifiées qui sont complémentaires d'une séquence sélective sur les adaptateurs. La séparation ultérieure des fragments amplifiés est obtenue par des amorces sélectives et visualisée par autoradiographie. Les AFLP surmontent les inconvénients de la méthode RFLP à forte intensité de main-d'œuvre et de temps et résolvent le problème de fiabilité causé par les amplifications non spécifiques dans les RAPD

Les AFLP se distinguent par leur stabilité génétique, ils fournissent un outil efficace, rapide et économique pour détecter un grand nombre de marqueurs génétiques polymorphes, qui peuvent être génotypés automatiquement. Cependant, les AFLP sont des marqueurs bi-alléliques dominants et sont incapables de distinguer les individus homozygotes dominants des individus hétérozygotes dominants. La méthode AFLP est une approche moléculaire idéale pour la génétique des populations et le typage du génome, elle est par conséquent largement appliquée pour détecter les polymorphismes génétiques, évaluer et caractériser les ressources génétiques animales.

2.4.4.1.Principe de l'AFLP

L'AFLP implique la digestion de l'ADN génomique à l'aide d'endonucléus de restriction, suivie d'une ligature d'adaptateur et d'une amplification par PCR.

Les produits amplifiés sont visualisés sur des gels de polyacrylamide haute résolution ou des séquenceurs automatisés. La variation de la longueur des fragments est analysée, ce qui donne l'estimation des relations génétiques entre les individus ou des variations.

2.4.3.2. Étapes de l'AFLP

1). Extraction de l'ADN et digestion de restriction

La technique AFLP nécessite une très bonne qualité ADN génomique pur, qui doit être exempt de protéines et de contaminants. L'étape suivante est la digestion avec des endonucléases de restriction. Eco R1 et Mse1 sont les enzymes les plus utilisées dans cette technique.

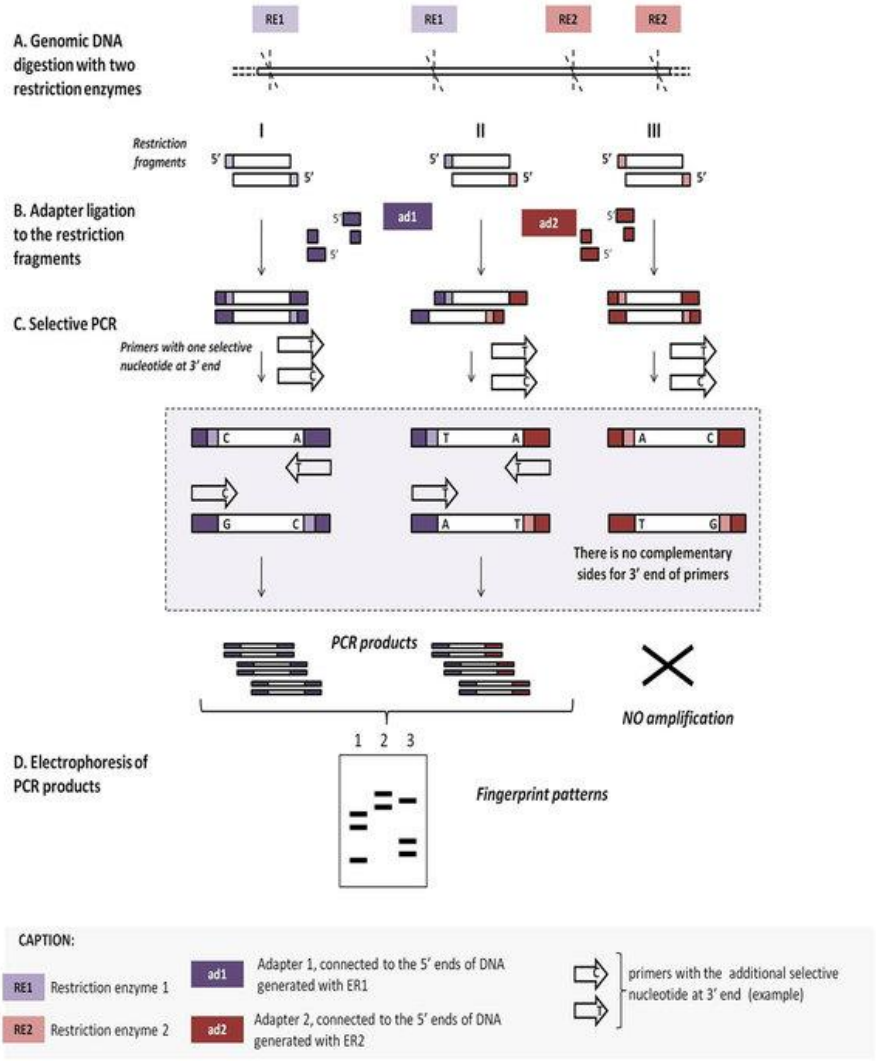


Figure 6. Aperçu de la méthode AFLP **(17)**

2). Ligation des adaptateurs d'oligonucléotides

Les adaptateurs sont de courtes séquences d'oligonucléotides double brin de 14 à 20 paires de bases. Deux adaptateurs différents sont utilisés, un pour Eco R1 et un autre pour Mse1, par exemple. Ces adaptateurs de séquences connues servent de cible pour l'amplification par PCR.

3). Amplification PCR

L'amplification se fait en deux phases,

- La première phase est connue sous le nom d'amplification présélective et

- La deuxième phase est l'amplification sélective

a) Amplification présélective

C'est le premier cycle d'amplification, dans lequel quelques fragments sont sélectivement amplifiés.

Une réaction PCR est définie qui contient

- L'ADN génomique,
- Les dNTP
- Polymérase
- Enzymes de restriction
- Amorces avec un nucléotide supplémentaire

La PCR de 20 cycles est définie en utilisant le produit PCR de l'étape d'amplification pré-sélective.

b). Amplification sélective

Un deuxième cycle d'amplification est connu sous le nom d'amplification sélective. Dans l'amplification sélective, des amorces plus stringentes (les amorces contiennent 3 nucléotides supplémentaires) sont utilisées pour réduire le nombre de fragments amplifiés.

Quelques cycles de Touchdown PCR sont configurés pour effectuer l'amplification.

Dans le touchdown pcr, la température de recuit est abaissée de certains degrés après chaque cycle de PCR pour améliorer l'efficacité d'amplification de l'AFLP.

4). Séparation et analyse

La séparation des fragments peut être effectuée sur un gel de polyacrylamide à 6 % ou un séquenceur automatisé contenant des gels pop. le schéma de flexion des fragments est analysé manuellement ou avec un logiciel d'analyse. La séparation et l'analyse doivent être effectuées sur un séquenceur, auquel cas des amorces marquées par fluorescence sont utilisées lors de l'amplification sélective.

2.6.3.3.Applications

Pour détecter divers polymorphismes dans différentes régions génomiques. Pour l'identification de la variation génétique dans les souches ou les espèces étroitement apparentées de plantes, de champignons, d'animaux et de bactéries.

La technique AFLP a été utilisée dans les tests criminels et de paternité pour déterminer de légères différences au sein des populations et dans les études de liaison pour générer des cartes pour l'analyse QTL (Quantitative trait locus).

2.6.3.4.Avantages du marqueur AFLP

Comme les sites de restriction sont présents dans l'ensemble du génome d'un individu, le marqueur AFLP permet d'analyser plusieurs locus à la fois. Les informations de séquence sur l'organisme ne sont pas essentielles car les amorces complémentaires aux séquences adaptatrices sont conçues. Contrairement au RFLP qui prend plus de temps pour l'hybridation de la sonde et plus de compétences, l'AFLP est relativement simple car l'amplification PCR des fragments est effectuée. L'AFLP est possible avec une quantité moindre de matrice génomique. Les résultats sont hautement reproductibles compte tenu de la qualité élevée de l'ADN en entrée.

2.6.3.5.Inconvénients de l'AFLP

L'AFLP ne peut pas être fait avec un ADN de mauvaise qualité ou un ADN dégradé. Comme les AFLP sont des marqueurs dominants dans la nature, ils ne peuvent pas détecter les individus homozygotes ou hétérozygotes. On ne peut pas déterminer quel fragment appartient à quel locus d'ADN car les AFLP sont de nature multi-locus.

2.7. Marqueurs ADN microsatellites

L'ADN microsatellite, également connu sous le nom de répétitions de séquences simples (SSR) ou de courtes répétitions en tandem (STR), sont des séquences répétées courantes dans les génomes eucaryotes. Généralement, ils consistent en des motifs constitués de 1 à 6 paires de bases (pb) répétées plusieurs fois en tandem (par exemple CACACACACACACACA). Les régions flanquantes des séquences répétées au niveau des loci microsatellites sont pour la plupart conservatrices et les motifs de répétition sont très variables entre différentes espèces et même différents individus de la même espèce. Ainsi, nous pouvons concevoir des amorces spécifiques basées sur les séquences conservées et amplifier les séquences répétées du noyau par PCR, les polymorphismes génétiques peuvent alors être détectés par électrophorèse.

Les SSR présentent les mêmes avantages que les RFLP et évitent l'utilisation de radio-isotopes essentiels aux RFLP; il a une répétabilité et une stabilité plus élevées que les RAPD; par rapport

aux marqueurs AFLP, les SSR sont des marqueurs co-dominants et capables de distinguer les homozygotes des hétérozygotes. Jusqu'à récemment, les microsatellites étaient les marqueurs les plus largement utilisés pour la diversité génétique, cartographiant les locus de traits quantitatifs pour la production et les traits fonctionnels chez les animaux de ferme; ils ont également été utilisés pour des pratiques de sélection assistée par marqueurs.

Les avantages et les inconvénients des marqueurs SSR ont été rapportés par de nombreux auteurs. Ses avantages sont les suivants : faibles quantités d'ADN matrice requises (10 à 100 ng), polymorphisme élevé, marqueurs co-dominants, haute précision, haute reproductibilité, différents microsatellites peuvent être multiplexés en PCR et se prêtent à l'automatisation. Ses inconvénients incluent : temps et coût de développement, les hétérozygotes peuvent être classés à tort comme homozygotes lorsque des allèles nuls se produisent en raison de mutations dans les sites d'hybridation des amorces, les bandes de bégaiement peuvent compliquer la notation précise des polymorphismes, le modèle de mutation sous-jacent largement inconnu et les marqueurs microsatellites aident à identifier la biodiversité neutre mais ne fournissent pas d'informations sur la biodiversité des traits fonctionnels. Malgré ces inconvénients, les marqueurs microsatellites sont encore des marqueurs d'ADN nucléaire populaires pour l'étude de la variation génétique entre et au sein des espèces.

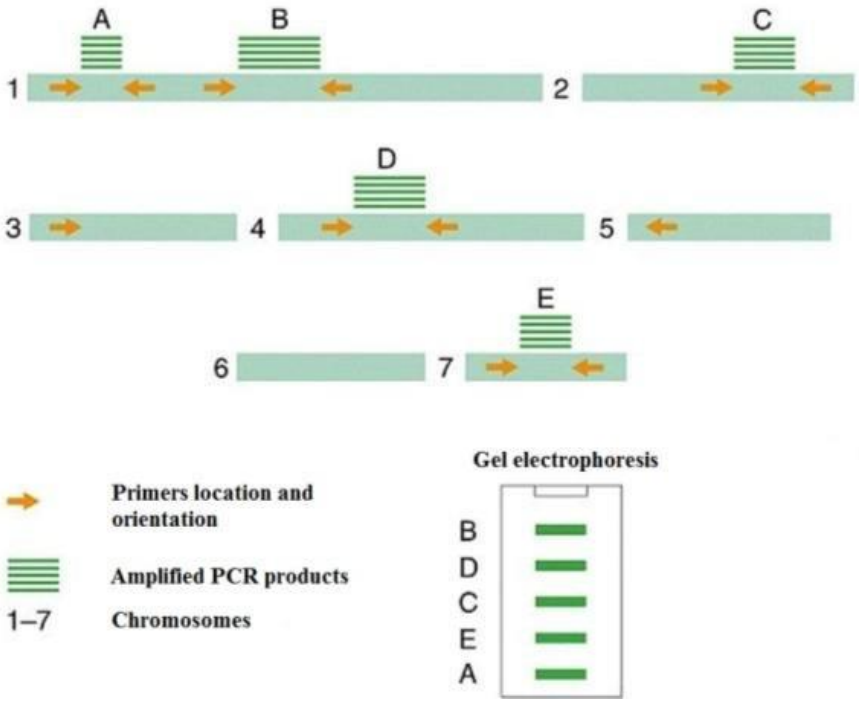


Figure 7. Mise en évidence du polymorphism SSR **(32)**

2.8. Marqueurs SNP (single nucleotide polymorphism)

Le SNP, une nouvelle technologie de marqueur moléculaire, a été proposé pour la première fois par Lander en 1996. Il fait référence à un polymorphisme de séquence causé par une mutation d'un seul nucléotide à un locus spécifique dans la séquence d'ADN. Ce type de polymorphisme comprend des transitions, des transversions, des insertions et des délétions à une seule base, et la fréquence des allèles mineurs doit être de 1 % ou plus. De tous les types de mutations SNP, les transitions sont les plus courantes (environ 2/3). Actuellement, les marqueurs SNP sont l'une des approches de génotypage préférées, car ils sont abondants dans le génome, génétiquement stables et se prêtent à une analyse automatisée à haut débit.

Le principe fondamental des SNP est d'hybrider les fragments d'ADN détectés avec des matrices de sondes d'ADN à haute densité (également appelées puces SNP) ; l'allèle SNP est ensuite nommé en fonction des résultats d'hybridation. Les SNP sont des marqueurs bi-alléliques, indiquant un polymorphisme spécifique dans seulement deux allèles d'une population. Les SNP se distribuent à la fois dans les régions codantes et non codantes des génomes, ils sont des acteurs essentiels dans le processus de variations génétiques des populations et d'évolution des espèces.

Actuellement, la technologie des puces à ADN est généralement utilisée lors des enquêtes SNP. Un groupe de locus SNP associés situés sur une certaine région du chromosome peut former un haplotype SNP. Les SNP sont une technologie de marqueurs moléculaires de troisième génération venant après les RFLP et les SSR; il a été utilisé avec succès pour étudier la variation génétique entre différentes espèces et races.

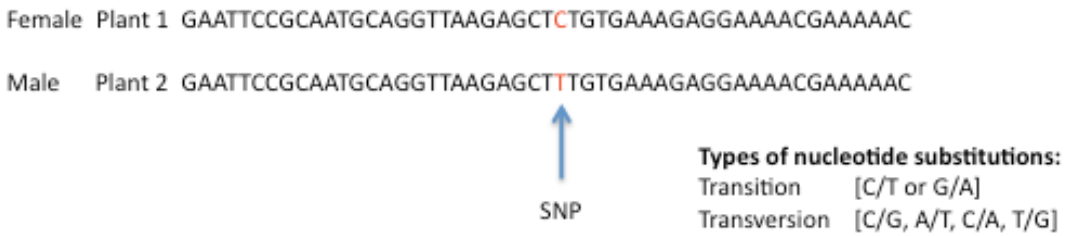


Figure 8. SNP entre deux individus de même espèce **(34)**

2.6.1. Avantages

Par rapport aux marqueurs précédents, les SNP présentent les avantages suivants :

- Ils sont nombreux et largement distribués dans tout le génome.

- Stabilité génétique élevée, excellente répétabilité et grande précision.
- Permettre un génotypage rapide et à haut débit
- Pratique pour distinguer efficacement les allèles hétérozygotes des homozygotes en raison de ses co-dominances.

En raison de leur distribution étendue et de leurs variations abondantes, les SNP jouent un rôle important dans la recherche sur la structure, la différenciation génétique, l'origine et l'évolution des populations d'animaux d'élevage. Par exemple, le déséquilibre de liaison (LD) entre différents SNP peut être utilisé pour l'analyse d'association. De plus, nous pouvons obtenir des informations sur la diversité des populations animales et l'évolution des populations (origines, différenciation et migrations) via les haplotypes SNP parmi différentes populations.

2.6.2. Désavantages

Un inconvénient des marqueurs SNP est le faible niveau d'information obtenu par rapport à celui d'un microsatellite hautement polymorphe, mais cela peut être compensé en utilisant un nombre plus élevé de marqueurs (puces SNP) et le séquençage du génome entier.

Avec l'amélioration de la technologie de séquençage, le séquençage du génome entier/gène est devenu disponible pour caractériser la diversité génétique chez les animaux de ferme. C'est la méthode la plus simple et elle fournit des informations plus complètes sur la variation génétique entre différentes populations car elle peut détecter toutes les variations au sein du génome. Actuellement, le problème avec le séquençage du génome entier est la mise en place d'une plate-forme d'analyse de données de haut niveau pour explorer des informations utiles pour la conservation et l'utilisation des animaux de ferme.

Références

1. Andersson DI, Slechta ES, Roth JR. Evidence that gene amplification underlies adaptive mutability of the bacterial *lac* operon. Science. (1998);282:1133–1135.
2. Ashley CT, Warren ST. Trinucleotide repeat expansion and human disease. Ann. Rev. Genet. (1995);29:703–728.
3. Bishop MD, Hawkins GA, Keeler CL: Use of DNA markers in animal selection. Ther. 1995, 43: 61-70.
4. Blattner FR, Plunkett G, Bloch CA. et al. The complete genome sequence of *Escherichia coli* K-12. Science. (1997);277:1453–1462.
5. Brooks SA, Gabreski N, Miller D, Brisbin A, Brown HE, Streeter C, Mezey J, Cook D, Antczak DF: Whole-genome SNP association in the horse: identification of a deletion in myosin Va responsible for lavender foal syndrome. PLoS Genet. 2010, 6: e1000909-10.
6. Demeke T, Adams RP, Chibbar R: Potential taxonomic use of random amplified polymorphic DNA (RAPD): a case study in Brassica. Theor Appl Genet. 1992, 84: 990-994.
7. Devine SE, Boeke JD. Integration of the yeast retrotransposon *Ty1* is targeted to regions upstream of genes transcribed by RNA polymerase III. Genes Devel. (1996);10:620–633.
8. Drinkwater RD, Hetzel DJS: Application of molecular biology to understanding genotype-environment interactions in livestock production. Proc. of an International Symposium on Nuclear Techniques in Animal Production and Health. 1991, IAEA, FAO, Vienna, 437-452. 15–19
9. Eggleston AK, West SC. Exchanging partners in *E. coli*. Trends Genet. (1996);12:20–26.
10. Foster PL. Mechanism of stationary phase mutation: a decade of adaptive mutation. Ann. Rev. Genet. (1999);33:57–88.
11. Francino MP, Ochman H. Strand asymmetries in DNA evolution. Trends Genet. (1997);13:240–245.
12. Freudenreich CH, Kantrow SM, Zakian VA. Expansion and length-dependent fragility of CTG repeats in yeast. Science. (1998);279:853–856.
13. Gacy AM, Goellner GM, Spiro C. et al. GAA instability in Friedreich's Ataxia shares a common, DNA-directed and intraallelic mechanism with other trinucleotide diseases. Mol. Cell. (1998);1:583–593.
14. Goodman MF. Coping with replication ‘train wrecks’ in *Escherichia coli* using Pol V, Pol II and RecA proteins. Trends Biochem. Sci. (2000);25:189–195.
15. Hassan E., Hoeft E., Zenglu L., Tulsieram L., 2011, Genetic markers for Orobanche resistance in sunflower, US Patent 7,872,170.
16. Koh MC, Lim CH, Chua SB, Chew ST, Phang STW: Random amplified polymorphic DNA (RAPD) fingerprints for identification of red meat animal species. Meat Sci. 1998, 48: 275-285.
17. Krawczyk B, Kur J, Stojowska-Swędrzyńska K., Śpibida M. (2016). Principles and applications of Ligation Mediated PCR methods for DNA-based typing of microbial organisms. Acta biochimica Polonica · 63, No 1/2016 39–52
18. Li J., Schulz B., Stich B. Population structure and genetic diversity in elite sugar beet germplasm investigated with SSR markers. Euphytica, 2010, 175: 35–42.
19. Li Y.C., Korol A.B., Fahima T., Beiles A., Nevo E. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Molecular Ecology, 2002, 11(12): 2453-2465. 14.

20. Lindahl T, Wood RD. Quality control by DNA repair. Science. (1999);286:1897–1905.
21. Mandel J-L. Breaking the rule of three. Nature. (1997);386:767–769.
22. Mellon I, Rajpal DK, Koi M, Boland CR, Champe GN. Transcription-coupled repair deficiency and mutations in human mismatch repair genes. Science. (1996);272:557–560.
23. Meselson M, Radding CM. A general model for genetic recombination. Proc. Natl Acad. Sci. USA. (1975);72:358–361.
24. Miladinović D., Taški-Ajduković K., Nagl N., Kovačević B., Balešević-Tubić S., Dušanić N., Jocić S. DNA polymorphism of wild sunflower accessions highly susceptible or highly tolerant to white rot on stalk. Helia 2011, 34 (55): 91-100.
25. Nadler CF, Hoffmann RS, Woolf A: G-band patterns as chromosomal markers, and the interpretation of chromosomal evolution in wild sheep (Ovis). Cell Mol Life Sci. 1973, 29: 117-119.
26. Nagl N., Taški-Ajduković K., Ćurčić Ţ., Danojević D., Kovačev L. (2012). Development of SSR markers for detection of DNA polymorphism in sugar beet pollinator lines. Proceedings of International Conference on Bioscience: Biotechnology and Biodiversity- Step in the Future, The Forth Joint UNS-PSU Conference, 18-20 June 2012, Novi Sad, Serbia, 117.
27. Panigrahi S, Rao TS. (2019). Functional Microbial Diversity in Contaminated Environment and Application in Bioremediatio. Microbial Diversity in the Genomic Era. 2019, Pages 359-385
28. Perutz MF. Glutamine repeats and neurodegenerative diseases: molecular aspects. Trends Biochem. Sci. (1999);24:58–63.
29. Restriction fragment length polymorphism (RFLP): Principle, Procedure and Applications: https://onlinesciencenotes.com/restriction-fragment-length-polymorphism-rflp-principle-procedure-and-applications/
30. Silversides FG, Crawford RD, Wang HC: The cytogenetics of domestic geese. Heredity. 1988, 79: 6-8.
31. Syvänen AC: Accessing geneic variation: genotyping single nucleotide polymorphisms. Nature Rev Genet. 2001, 2: 930-942.
32. Taški-Ajduković K and Nagl N (2015). Assessment of plant genetic variability by molecular markers. Applications of Molecular Markers in Plant Genome Analysis and Breeding, 2015: 1-18
33. Tautz D: Hypervariability of simple sequences as a generalsource for polymorphic DNA markers. Nucleic Acids Res. 1989, 17: 6463-6471.
34. Terence A Brown TA. (2002). Genome. 2snd edition. Wiley-Liss. Oxford
35. Van Wezel IL, Rodgers RJ: Morphological characterization of bovine primordial follicles and their environment in vivo. Biol Reprod. 1996, 55: 1003-1011.
36. Varshney RK, Chabane K, Hendre PS, Aggarwal RK, Graner A: Comparative assessment of EST-SSR, EST-SNP and AFLP markers for evaluation of genetic diversity and conservation of genetic resources using wild, cultivated and elite barleys. Plant Sci. 2007, 173: 638-649.
37. Vignal A, Milan D, SanCristobal M: A review on SNP and other types of molecular markers and their use in animalgenetics. Genet Selec Evol. 2002, 34: 275-305.
38. Vos P, Hogers R, Bleeker M, Reijans M, Lee TVD, Hornes M, Friters A, Pot J, Paleman J, Kuiper M, Zabeau M: AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res. 1995, 23: 4407-4414.
39. Williams JGK, Kubeilik AR, Livak KJ, Rafalski JA, Tingey SV: DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res. 1990, 18: 6531-6535.

Chapitre 04
Construction d'une banque génomique

Chapitre 4. Construction d’une banque génomique

6. Définition d’une banque génomique

Une banque génomique ou une banque de gènes est une collection complète de fragments d'ADN clonés qui constitue l'intégralité du génome d'un organisme. Il représente tous les gènes exprimés, non exprimés, introns, exons, etc. Les banques génomiques peuvent être conservées pendant de nombreuses années et les copies peuvent être utilisées à des fins de recherche.

7. Avantages d'une bibliothèque génomique

- Les banques génomiques dérivées d'organismes eucaryotes sont essentielles pour étudier la séquence du génome d'un gène d'intérêt particulier.
- Il est utile pour les procaryotes avec de petits génomes d'identifier un clone codant pour un gène d'intérêt spécifique.
- Il aide les chercheurs à en savoir plus sur la structure et la fonction génomique d'un organisme. Il est également utilisé pour étudier les mutations génétiques.
- Des gènes pharmaceutiquement importants peuvent également être identifiés par cette méthode

8. Type de banques génomiques

8.1. Banque d'ADN génomique

L'ADN génomique est l'ADN chromosomique d'une entité représentative de la collection de son contenu génomique. Ils sont différents de celui de l'ADN complémentaire, de l'ADN mitochondrial ou de l'ADN plasmidique bactérien. Ceux-ci sont directement préparés à partir de l'ADN génomique, et représentent le génome complet d'une entité. Pour leur construction, les ligases et les endonucléases de restriction sont indispensables. Comme il porte des introns, ils sont incapables de s'exprimer chez les procaryotes. De plus, les procaryotes manquent de machines pour traiter les introns.

8.2.Banque d'ADNc (banque d'ADN complémentaire)

Il s'agit d'une copie d'ADN d'une molécule d'ARNm générée par la transcriptase inverse, une ADN polymérase qui peut utiliser l'ARN ou l'ADN comme matrice. Ceux-ci sont préparés à l'aide d'ARNm comme matrices, leur matériel de départ. Ils ne sont représentatifs que des gènes du génome qui sont exprimés dans des conditions spécifiques. L'ADNc n'a pas d'introns et peut donc être exprimé dans des cellules procaryotes.

9. Clonage moléculaire

Le clonage moléculaire, un terme qui en est venu à signifier la création de molécules d'ADN recombinant, a stimulé les progrès dans toutes les sciences de la vie. À partir des années 1970, avec la découverte des endonucléases de restriction - des enzymes qui coupent sélectivement et spécifiquement les molécules d'ADN - la technologie de l'ADN recombinant a connu une croissance exponentielle à la fois en termes d'application et de sophistication, produisant des outils de plus en plus puissants pour la manipulation de l'ADN. Le clonage de gènes est maintenant si simple et efficace qu'il est devenu une technique de laboratoire standard. Cela a conduit à une explosion de la compréhension de la fonction des gènes au cours des dernières décennies. Les technologies émergentes promettent des possibilités encore plus grandes, comme permettre aux chercheurs d'assembler de manière transparente plusieurs fragments d'ADN et de transformer les plasmides résultants en bactéries, en moins de deux heures, ou l'utilisation de cassettes de gènes interchangeables, qui peuvent être facilement déplacées entre différentes constructions, pour maximiser rapidité et flexibilité. Dans un avenir proche, le clonage moléculaire verra probablement l'émergence d'un nouveau paradigme, avec des techniques de biologie synthétique qui permettront la synthèse chimique in vitro de toute construction d'ADN spécifiée in silico. Ces avancées devraient permettre une construction et une itération plus rapides des clones d'ADN, accélérant le développement de vecteurs de thérapie génique, de procédés de production de protéines recombinantes et de nouveaux vaccins.

9.1.Technique du clonage

Le clonage moléculaire fait référence à l'isolement d'une séquence d'ADN à partir de n'importe quelle espèce (souvent un gène) et à son insertion dans un vecteur de propagation, sans altération de la séquence d'ADN d'origine. Une fois isolés, les clones moléculaires peuvent être utilisés pour générer de nombreuses copies de l'ADN pour l'analyse de la séquence du gène et/ou pour exprimer la protéine résultante pour l'étude ou l'utilisation de la fonction de la protéine. Les clones peuvent également être manipulés et mutés in vitro pour modifier l'expression et la fonction de la protéine.

Le workflow de clonage de base comprend quatre étapes :

- Isolement de fragments d'ADN cibles (souvent appelés inserts)
- Ligation des inserts dans un vecteur de clonage approprié, créant des molécules recombinantes (par exemple, des plasmides)

- Transformation de plasmides recombinants en bactéries ou autres hôtes appropriés pour la propagation
- Criblage/sélection des hôtes contenant le plasmide recombinant prévu

Ces quatre étapes révolutionnaires ont été soigneusement reconstituées et exécutées par plusieurs laboratoires, à partir de la fin des années 1960 et du début des années 1970. Un résumé des découvertes qui composent le clonage moléculaire traditionnel est décrit dans les pages suivantes.

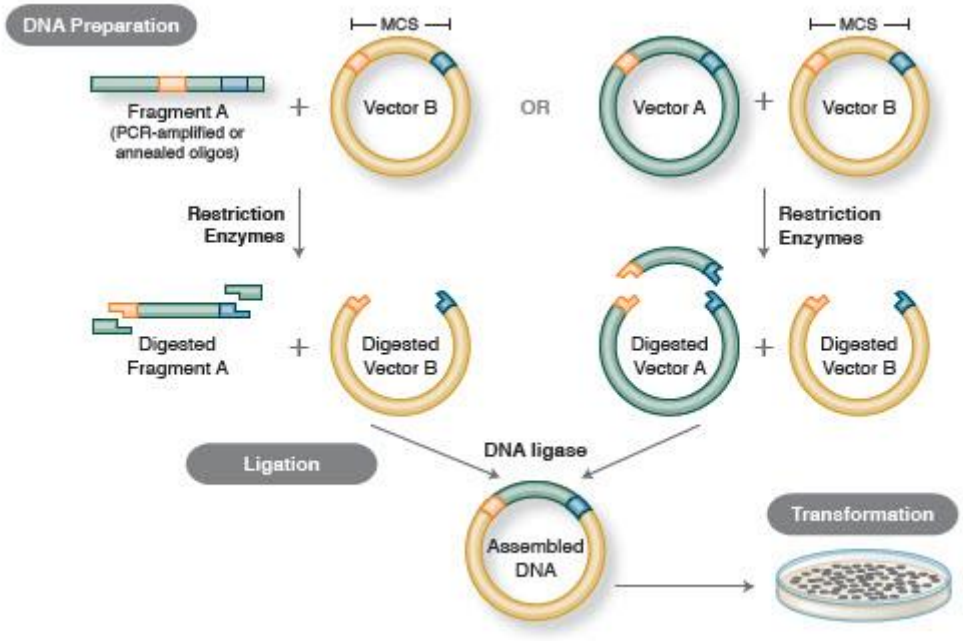


Figure 1. En utilisant la PCR, des sites de restriction sont ajoutés aux deux extrémités d'un ADNdb, qui est ensuite digéré par les enzymes de restriction correspondantes (REases). L'ADN clivé peut ensuite être ligaturé à un vecteur plasmidique possédant des extrémités compatibles. Des fragments d'ADN peuvent également être déplacés d'un vecteur à un autre par digestion avec des REases et ligature avec des extrémités compatibles du vecteur cible. La construction assemblée peut ensuite être transformée en *Escherichia coli* **(11).**

9.2.Protocole du clonage

9.2.1. Préparation du vecteur de clonage

Les vecteurs utilisés dans les méthodes de clonage traditionnelles sont basés sur des plasmides, qui sont des ADN circulaires à double brin qui se répliquent à l'intérieur des bactéries indépendamment de l'ADN génomique. Tous les vecteurs de clonage basés sur des plasmides contiennent un certain nombre d'éléments cruciaux, y compris une origine bactérienne de

réplication pour se propager efficacement dans la cellule hôte bactérienne ; site(s) d'enzymes de restriction unique ou, plus communément, un site de clonage multiple (MCS) qui contient un certain nombre de sites d'enzymes de restriction pour permettre l'ajout facile d'un insert d'intérêt ; et un marqueur (par exemple, la résistance aux antibiotiques) pour sélectionner les bactéries après une absorption réussie du vecteur.

Dans certains vecteurs, le MCS est situé dans un gène qui sert de marqueur et permet le criblage de clones dans lesquels l'insert a été épissé avec succès. Par exemple, le vecteur pUC18 exprime le gène lacZα codant pour le peptide alpha de la bêta-galactosidase qui, en combinaison avec X-gal, permet la sélection de la couleur des colonies bactériennes formées après clonage (en savoir plus sur la sélection bleu/blanc dans le dépistage des colonies).

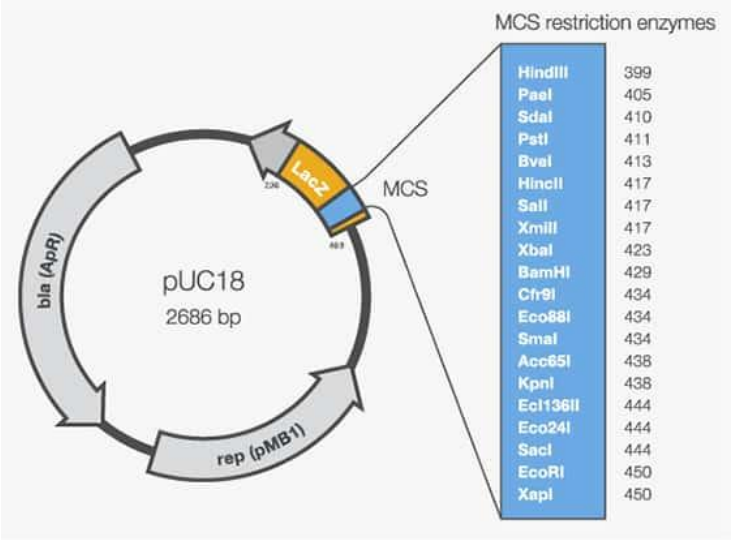


Figure 2. Carte de pUC18 avec son MCS **(24)**

La première étape de la préparation du vecteur pour le clonage traditionnel consiste à créer un site d'insertion par digestion de restriction. Le choix des enzymes de restriction dépend de la présence et de l'emplacement de leurs séquences de reconnaissance sur le vecteur et l'insert, et de leur compatibilité pour la ligature. Le MCS, s'il est disponible, est souvent le premier choix pour l'insertion, car la région est spécifiquement conçue pour le clonage.

Après la digestion de restriction, la déphosphorylation du vecteur peut être nécessaire pour empêcher l'auto-ligation, en particulier si les extrémités résultantes de la digestion du vecteur sont compatibles ou franches. Au cours de la déphosphorylation, l'enzyme phosphatase alcaline élimine les groupes phosphate 5 'aux extrémités. Cela empêche l'auto-ligature du vecteur car l'enzyme ligase nécessite à la fois un 5 'phosphate et un 3' OH pour joindre les deux extrémités lors de la recircularisation du vecteur.

La déphosphorylation du vecteur est importante pour réduire le bruit de fond et favoriser l'insertion du fragment souhaité dans le vecteur. Les molécules vectorielles auto-ligaturées et les molécules vectorielles porteuses d'inserts peuvent être absorbées par les bactéries lors de la transformation et confèreront la même résistance aux antibiotiques à ces cellules. Cela créera un fond plus élevé de colonies indésirables si le vecteur n'est pas déphosphorylé.

L'émoussage des extrémités du vecteur peut être nécessaire, selon les enzymes de restriction utilisées. La purification des fragments souhaités est également recommandée pour une ligature réussie.

9.2.2. Préparation de l'insert

La source de l'insert pour le clonage peut être de l'ADN génomique, une partie d'un autre plasmide ou un fragment d'ADN linéaire. Quel que soit le type d'ADN source, une première étape courante dans la préparation de l'insert consiste à effectuer une digestion de restriction pour générer des extrémités compatibles pour un épissage ultérieur dans le vecteur.

Comme pour la préparation du vecteur, les enzymes de restriction qui conviennent au clonage de l'insert dans le vecteur sont sélectionnées. L'une des stratégies les plus populaires consiste à effectuer une double digestion de l'insert et du vecteur pour le clonage directionnel. Dans l'exemple suivant, deux enzymes qui génèrent des extrémités non compatibles (EcoRI et KpnI) sont utilisées. Étant donné que les extrémités du vecteur et de l'insert ne peuvent se joindre que dans une seule orientation en raison de la compatibilité (EcoRI avec EcoRI, KpnI avec KpnI), cette approche permet de cloner l'insert de manière directionnelle.

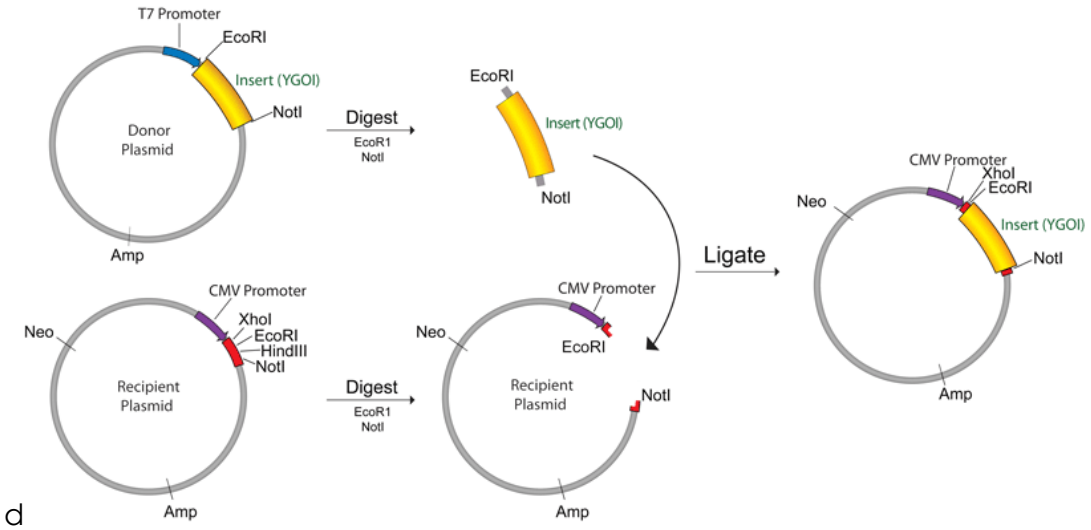


Figure 3. Digestion du vecteur et ligation de l'insert **(21).**

9.2.3. Assemblage (ligature)

Une fois les fragments d'intérêt obtenus, une réaction de ligation peut être mise en place pour joindre l'insert et le vecteur. L'enzyme la plus couramment utilisée pour la ligature est l'ADN ligase T4, qui relie les extrémités de l'ADN entre les groupes 5 'phosphate et 3' OH. La réaction de l'ADN ligase T4 nécessite de l'ATP, du DTT et du Mg2+, qui sont généralement fournis dans le tampon de réaction (Figure 3). Pour améliorer le résultat de la ligature, une recommandation générale consiste à mettre en place plusieurs réactions avec différents rapports molaires insert:vecteur, généralement compris entre 1:1 et 5:1. Pour des ligatures moins efficaces, comme avec des fragments d'ADN à extrémités franches, l'ajout de macromolécules inertes comme le polyéthylène glycol (PEG) est souvent recommandé augmenter la concentration efficace des composants de la réaction et ainsi améliorer l'efficacité de la ligature.

Le mélange ligaturé peut être utilisé directement dans la transformation de cellules chimiquement compétentes mais peut nécessiter une purification avant la transformation de cellules électrocompétentes. Si du PEG a été utilisé dans la réaction de ligature, l'inactivation par la chaleur de la ligase n'est pas recommandée après la réaction, car cela peut réduire l'efficacité de la transformation.

9.2.4. Transformation

La transformation est un processus naturel dans lequel les cellules bactériennes absorbent l'ADN étranger à une faible fréquence. Dans les applications de biologie moléculaire, ce processus est amélioré et exploité pour propager des plasmides à l'intérieur de bactéries qui ont été rendues « compétentes » (poreuses) pour l'absorption d'ADN.

Des cellules compétentes sont disponibles dans le commerce pour une transformation efficace et fiable. L'approche la plus courante pour préparer les bactéries à être compétentes pour la transformation consiste à traiter les cellules bactériennes en phase logarithmique avec du chlorure de calcium. Lorsque les cellules chimiquement compétentes sont mélangées avec l'ADN de la réaction de ligature, puis soumises à un choc thermique à 42°C, une partie de l'ADN est absorbée par les cellules bactériennes, où il commence à se répliquer.

Différentes souches de cellules compétentes sont disponibles, et le choix est basé sur des objectifs expérimentaux et des applications en aval. Par exemple, pour effectuer un dépistage « bleu/blanc », il faut choisir une souche bactérienne porteuse d'une mutation lacZ (lacZΔM15). Si l'expérience nécessite une digestion avec des enzymes de restriction sensibles à la

méthylation, le plasmide doit être propagé dans une souche bactérienne dcm–/dam–. Pour l'expression des protéines, la souche doit s'adapter à la stabilité et à la traduction de l'ARNm, ainsi qu'à une induction élevée de l'expression de la protéine recombinante.

De plus, l'efficacité de transformation des cellules compétentes est une considération importante. Les fabricants fournissent l'efficacité de transformation des cellules compétentes en unités formant colonies par microgramme d'ADN (UFC/µg), allant généralement de 1 x 106 à 1 x 10^9 UFC/µg. Dans des stratégies de ligature et de clonage plus difficiles, le choix de cellules avec les efficacités de transformation les plus élevées peut grandement améliorer la probabilité d'obtenir les clones souhaités.

Une autre méthode pour transformer les cellules bactériennes est l'électroporation. Dans cette technique, les cellules bactériennes électrocompétentes et les plasmides ligaturés sont traités avec un courant électrique qui crée des pores transitoires dans la membrane cellulaire bactérienne pour l'absorption d'ADN.

Les bactéries transformées (après choc thermique ou électroporation) sont ensuite étalées sur une plaque de gélose avec un antibiotique approprié, et criblées (par criblage bleu-blanc ou une autre méthode) pour les colonies qui portent le plasmide souhaité avec insert.

9.3. Criblage (screening) des clones

La réaction de transformation contient un mélange de cellules sans vecteur, le vecteur sans insert, l'insert seul et le vecteur et l'insert ligaturés avec succès. Les bactéries sans le vecteur n'ont pas le gène de résistance aux antibiotiques et ne se développeront pas, tandis que les bactéries transformées avec le vecteur (avec ou sans l'insert) survivent grâce au gène de résistance aux antibiotiques exprimé (Figure 4). Ainsi, la résistance aux antibiotiques permet la sélection pour l'absorption d'un plasmide intact.

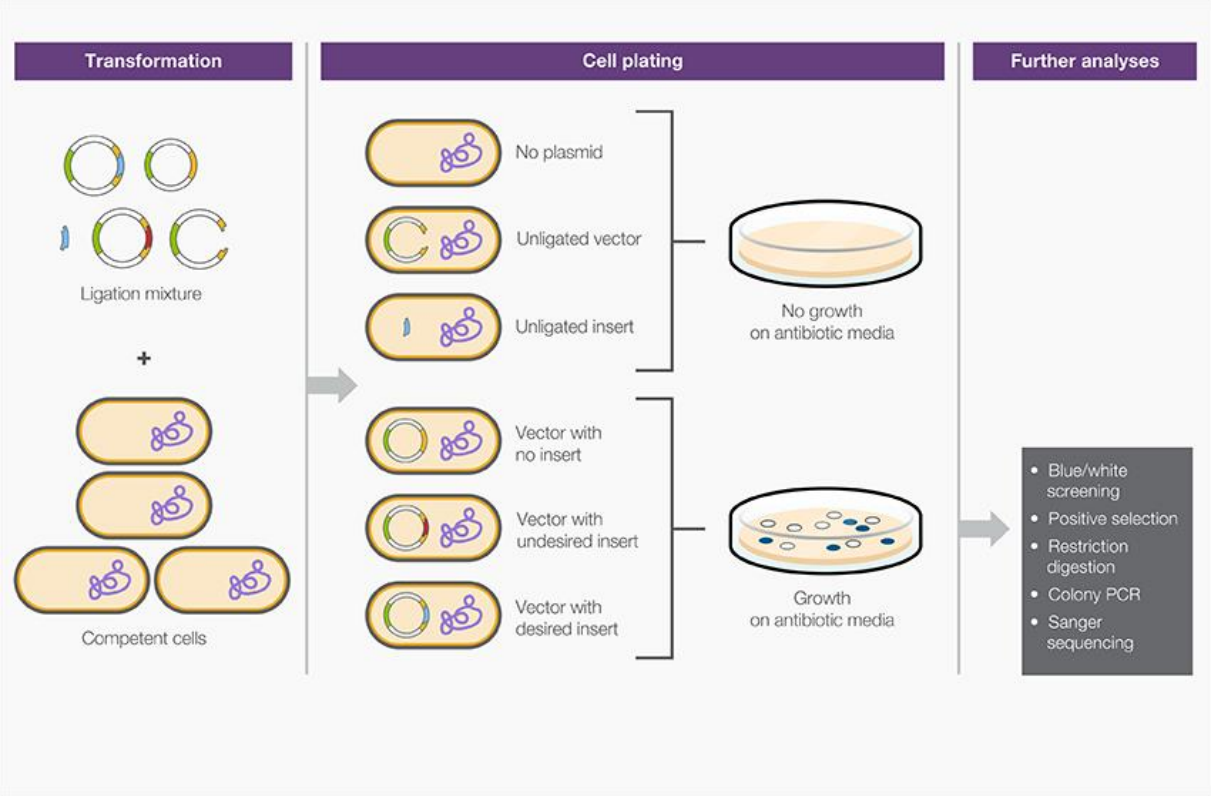


Figure 4. Mélange de bactéries après transformation et leurs phénotypes (croissance/pas de croissance/bleu-blanc) sur une plaque de milieu de sélection antibiotique. Le mélange de ligation peut comprendre des produits défaillants (insert non ligaturé, vecteur non ligaturé et vecteur sans insert/vide), ainsi que des vecteurs porteurs d'insert souhaités/non souhaités **(24)**

Pour identifier si les colonies transformées contiennent un insert, un certain nombre de méthodes peuvent être employées, dont les plus courantes sont le criblage « bleu/blanc » et la sélection positive. Le criblage bleu/blanc repose sur la transformation d'une souche bactérienne exprimant un gène lacZ mutant (lacZΔM15), qui peut être complémenté par le peptide alpha de la bêta-galactosidase, codé sur le vecteur (complémentation alpha). Les cellules transformées sont étalées sur un milieu de croissance qui comprend un inducteur transcriptionnel pour l'expression de lacZ, IPTG, et un substrat chromogénique de LacZ, X-gal (5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside). Dans le criblage bleu/blanc, LacZ hydrolysera le X-gal, produisant un colorant bleu et donc une colonie bleue. Lorsqu'un insert d'ADN perturbe le gène lacZα codé par le vecteur, aucun lacZ fonctionnel n'est formé et les colonies transformées sont blanches.

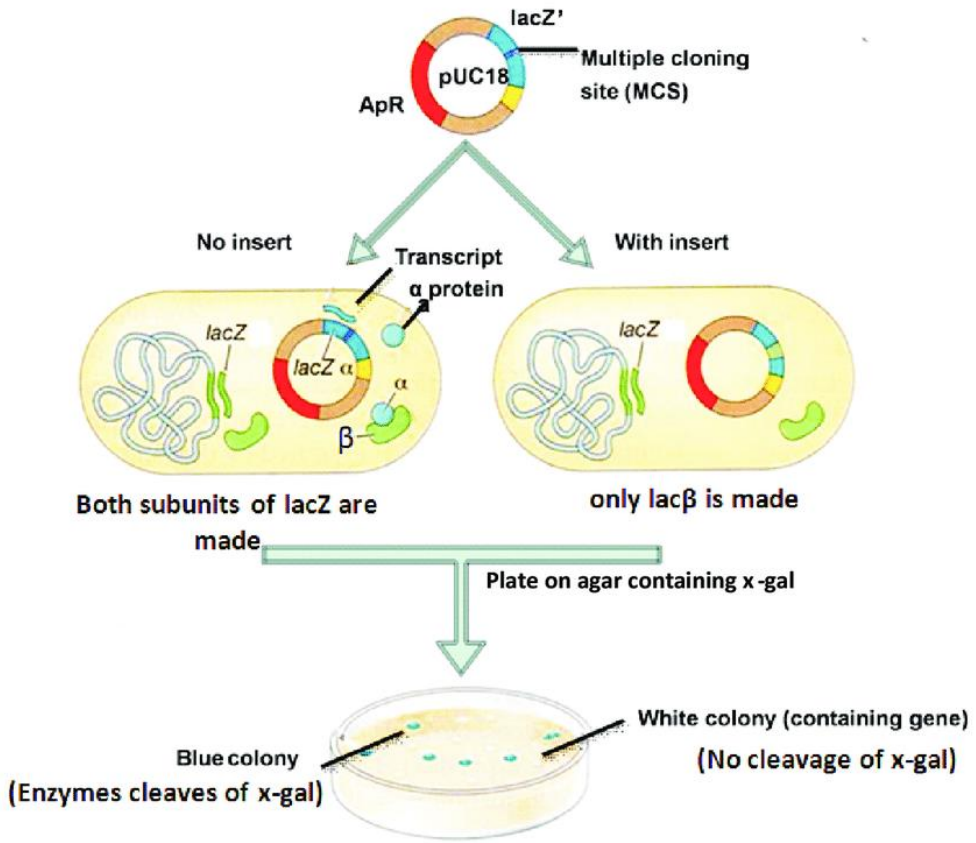


Figure 5. Tramage bleu/blanc pour reconnaître les vecteurs contenant des inserts **(19).**

9.4.Sélection des clones cibles

9.4.1. Hybridation fluorescente in situ (FISH)

L'hybridation in situ est utilisée pour localiser les séquences d'ADN sur les chromosomes. Dans l'hybridation moléculaire, une séquence d'ADN ou d'ARN marquée est utilisée comme sonde pour identifier ou quantifier la contrepartie naturelle de la séquence dans un échantillon biologique.

9.4.1.1.Etape de la méthode FISH

a- Préparation d'une sonde fluorescente

La première étape du processus consiste à faire soit une copie fluorescente de la séquence de la sonde soit une copie modifiée de la séquence de la sonde qui peut être rendue fluorescente plus tard dans la procédure (figure 6).

b- Dénaturation de la cible

Ensuite, avant que toute hybridation puisse se produire, les séquences cible et sonde doivent être dénaturées avec de la chaleur ou des produits chimiques (Figure 6). Cette étape de dénaturation est nécessaire pour que de nouvelles liaisons hydrogène se forment entre la cible et la sonde lors de l'étape d'hybridation ultérieure.

c- Hybridation de la cible avec la sonde

La sonde et les séquences cibles sont ensuite mélangées (Figure 6) et la sonde s'hybride spécifiquement à sa séquence complémentaire sur le chromosome.

d- Détection et selection de la cible

Si la sonde est déjà fluorescente, il sera possible de détecter directement le site d'hybridation. Dans d'autres cas, une étape supplémentaire peut être nécessaire pour visualiser la sonde hybridée. Les hybrides formés entre les sondes et leurs cibles chromosomiques peuvent être détectés à l'aide d'un microscope à fluorescence.

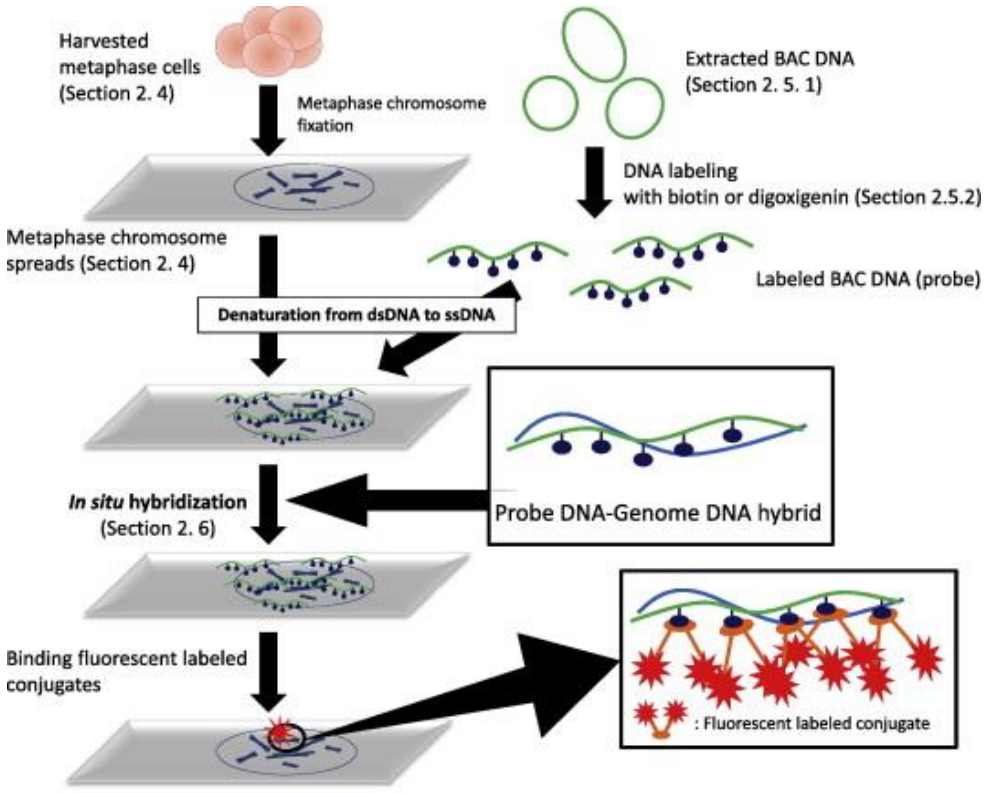


Figure 6. Procédure expérimentale d'hybridation in situ par fluorescence utilisant l'ADN BAC comme sonde (BAC-FISH) **(3).**

Lorsque les chercheurs conçoivent une expérience FISH, ils doivent déterminer si la sensibilité et la résolution nécessaires à l'expérience se situent dans les limites techniques de la microscopie à fluorescence. La sensibilité dépend de la capacité de collecte de lumière du microscope particulier, qui détermine si de petites séquences cibles, qui sont plus difficiles à voir que de grandes séquences cibles, peuvent être détectées. La résolution fait référence à la capacité de distinguer deux points sur la longueur d'un chromosome. En fin de compte, la microscopie optique ne peut pas résoudre les objets séparés par moins de 200 à 250 nm, la limite inférieure du spectre de la lumière visible. Avec ces limites techniques à l'esprit, les chercheurs doivent également tenir compte de la conformation de l'ADN dans le chromosome. Les chromosomes en métaphase sont des milliers de fois plus compacts que les chromosomes en interphase, qui à leur tour sont au moins dix fois plus compacts que l'ADN nu. Lorsque tous ces facteurs sont pris en compte ensemble, les chercheurs s'attendent généralement à obtenir une résolution de l'ordre des mégabases pour les positions sur les chromosomes en métaphase et une résolution de l'ordre de des dizaines de milliers de kilobases pour les chromosomes en interphase.

10. **Génomique fonctionnelle**

Le domaine de la génomique fonctionnelle tente de décrire les fonctions et les interactions des gènes et des protéines en utilisant des approches à l'échelle du génome, contrairement à l'approche gène par gène des techniques classiques de biologie moléculaire. Il combine des données dérivées des divers processus liés à la séquence d'ADN, à l'expression des gènes et à la fonction des protéines, telles que la transcription codante et non codante, la traduction des protéines, les interactions protéine-ADN, protéine-ARN et protéine-protéine. Ensemble, ces données sont utilisées pour modéliser des réseaux interactifs et dynamiques qui régulent l'expression des gènes, la différenciation cellulaire et la progression du cycle cellulaire.

L'étude des cellules au niveau des systèmes a été facilitée par les progrès technologiques récents, ainsi que par la disponibilité de séquences génomiques complètes. Depuis la publication historique de la première ébauche du génome humain en 2001, les génomes de centaines d'organismes de toutes les branches de l'arbre de la vie ont été séquencés. Cela a conduit à des annotations améliorées des gènes et de leurs produits, et a permis des études à l'échelle du génome visant à comprendre les interactions et les processus moléculaires dans la cellule.

10.1. Techniques de la génomique fonctionnelle

10.1.1. Puces à ADN

10.1.1.1. Définition d’une puce à ADN

Également appelées puces à ADN, puces à gènes et biopuces. Les biopuces sont des biocapteurs de dernière génération développés par l'utilisation de sondes ADN. Les puces à ADN sont des supports solides généralement constitués de verre ou de silicium sur lesquels l'ADN est attaché dans une conception de grille organisée et pré-arrangée. Chaque point d'ADN, appelé sonde, signifie un seul gène. Les puces à ADN peuvent examiner simultanément l'expression de dizaines de milliers de gènes. Des fragments d'ADN simple brin marqués ou d'ARN antisens provenant d'un échantillon d'intérêt sont hybrides au microréseau d'ADN dans des conditions très strictes. Chaque sonde est identifiée par son emplacement sur le microréseau d'ADN, et la quantité d'hybridation détectée pour une sonde spécifique est proportionnelle au niveau d'acides nucléiques à partir de l'emplacement génomique correspondant dans l'échantillon d'origine.Il existe 2 types de puces à ADN, à savoir les puces à base d'ADNc et les puces à base d'oligonucléotides.

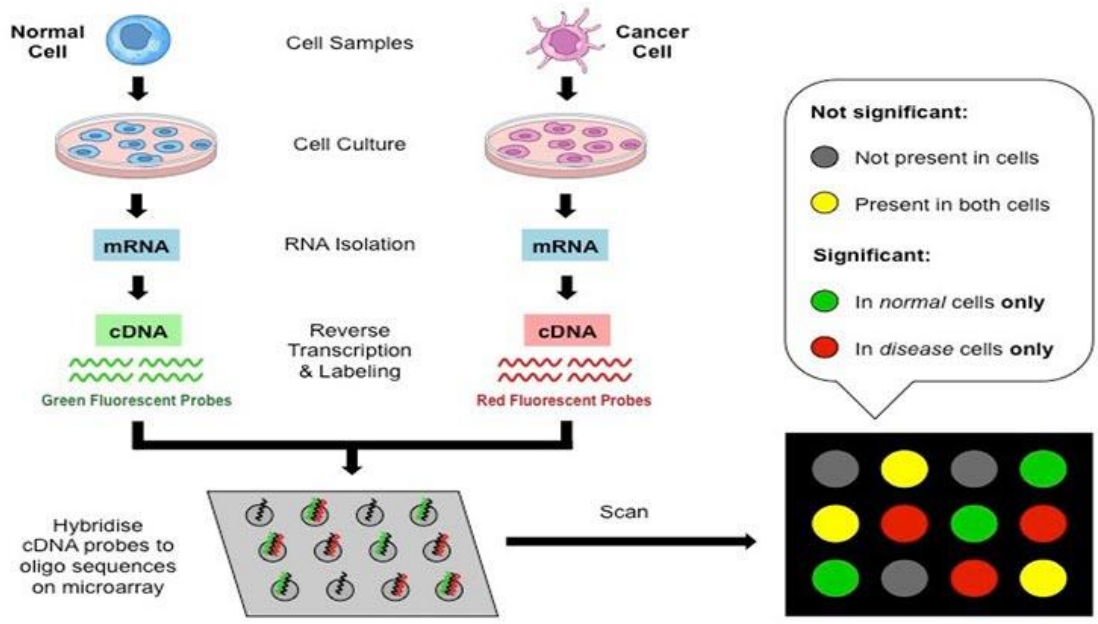


Figure 7. Principe de la puce à ADN **(8)**

10.1.1.2. Principe de la puce à ADN

La technologie des puces à ADN est issue du transfert de Southern, dans lequel l'ADN fragmenté est attaché à un substrat puis sondé avec une séquence d'ADN connue. La puce à ADN est basée sur le principe de l'hybridation entre les brins d'acide nucléique.

Les séquences d'acides nucléiques complémentaires ont la particularité de s'apparier spécifiquement par la formation de liaisons hydrogène entre des paires de bases nucléotidiques complémentaires. Un échantillon inconnu de séquence d'ADN est appelé échantillon ou cible et la séquence connue de molécule d'ADN est appelée sonde. Des colorants fluorescents sont utilisés pour marquer les échantillons et au moins 2 échantillons sont hybridés sur la puce. Un grand nombre de paires de bases complémentaires dans la séquence nucléotidique suggère une liaison non covalente plus étroite entre les deux brins.

Après le lavage des séquences de liaison non spécifiques, seuls les brins fortement appariés resteront hybridés. Ainsi, les séquences cibles marquées par fluorescence qui s'apparient à la sonde libèrent un signal qui repose sur la force de l'hybridation détectée par le nombre de bases appariées.

Les puces à ADN utilisent une quantification relative dans laquelle la comparaison d'un même caractère est effectuée dans deux conditions différentes et l'identification de ce caractère est connue par sa position. Une fois l'hybridation terminée, la surface de la puce peut être examinée à la fois qualitativement et quantitativement à l'aide d'une autoradiographie, d'un balayage laser, d'un dispositif de détection de fluorescence, d'un système de détection d'enzymes. La présence d'une séquence génomique ou d'ADNc sur 1 00 000 ou plus peut être détectée en une seule hybridation en utilisant une micropuce à ADN.

10.1.1.3. Étapes impliquées dans les biopuces à base d'ADNc :

a- Collecte d'échantillons

Un échantillon peut être n'importe quelle cellule/tissu sur lequel nous souhaitons mener notre étude. Généralement, 2 types d'échantillons sont prélevés, c'est-à-dire des cellules saines et infectées, pour comparer et obtenir les résultats.

b- Isolement de l'ARNm

L'extraction de l'ARN d'un échantillon est réalisée à l'aide d'une colonne ou d'un solvant comme le phénol-chloroforme. L'ARNm est isolé de l'ARN extrait en laissant derrière lui l'ARNr et l'ARNt. Comme l'ARNm a une queue poly-A, des billes de colonne avec des queues poly-T sont utilisées pour lier l'ARNm. Suite à l'extraction, un tampon est utilisé pour rincer la colonne afin d'isoler l'ARNm des billes.

c- Création d'ADNc marqué

La transcription inverse de l'ARNm donne l'ADNc. Les deux échantillons sont ensuite intégrés avec différents colorants fluorescents pour la production de brins d'ADNc fluorescents, ce qui permet de différencier la catégorie d'échantillon des ADNc.

d- Hybridation

Les ADNc marqués des deux échantillons sont placés sur la puce à ADN qui permet l'hybridation de chaque ADNc à son brin complémentaire. Ensuite, ils sont soigneusement lavés pour éliminer les séquences non appariées.

e- Collecte et analyse

Un scanner de puces à ADN est utilisé pour collecter les données. Le scanner contient un laser, un ordinateur et une caméra. Le laser est chargé d'exciter la fluorescence de l'ADNc, générant des signaux. La caméra enregistre les images produites au moment où le laser balaye le réseau. Ensuite, l'ordinateur stocke les données et donne des résultats instantanément. Les données sont maintenant analysées. L'intensité distincte des couleurs pour chaque endroit détermine le caractère du gène dans cet endroit particulier.

10.1.1.4. Applications de la technique des puces à ADN

- Découverte de médicament
- Étude de la génomique fonctionnelle
- Séquençage ADN
- Profilage de l'expression génique
- Étude de la protéomique
- Diagnostic et génie génétique
- Recherches toxicologiques
- Pharmacogénomique

5.1.3. Technologies de séquençage de nouvelle génération

Trois principales plateformes de séquençage de nouvelle génération (NGS) sont largement utilisées : la plateforme Roche 454 (Roche Life Sciences), la plateforme Applied Biosystems SOLiD (Applied Biosystems), et les plateformes Illumina (anciennement Solexa) Genome Analyzer et HiSeq (Illumina). Pour ces trois plates-formes NGS, l'ADN matrice est fragmenté, lié à des adaptateurs, amplifié par réaction en chaîne par polymérase, puis immobilisé sur des billes ou sur un réseau où des grappes constituées de fragments d'ADN identiques sont formées. Ces grappes sont lues par des cycles séquentiels d'incorporation, de lavage et de détection de

nucléotides, où le nombre de cycles détermine finalement la longueur de lecture (Fig. 1). Une quatrième technologie de séquençage de l'ADN a été récemment développée par Ion Torrent. La technologie Ion Torrent tire parti de l'ion hydrogène qui est libéré comme sous-produit de l'incorporation d'un nucléotide dans un brin d'ADN par la polymérase. Le séquenceur détecte directement les ions produits par la synthèse d'ADN polymérase dirigée par la matrice sur un dispositif de détection à semi-conducteur parallèle massif qui transforme directement ce signal chimique en information numérique.

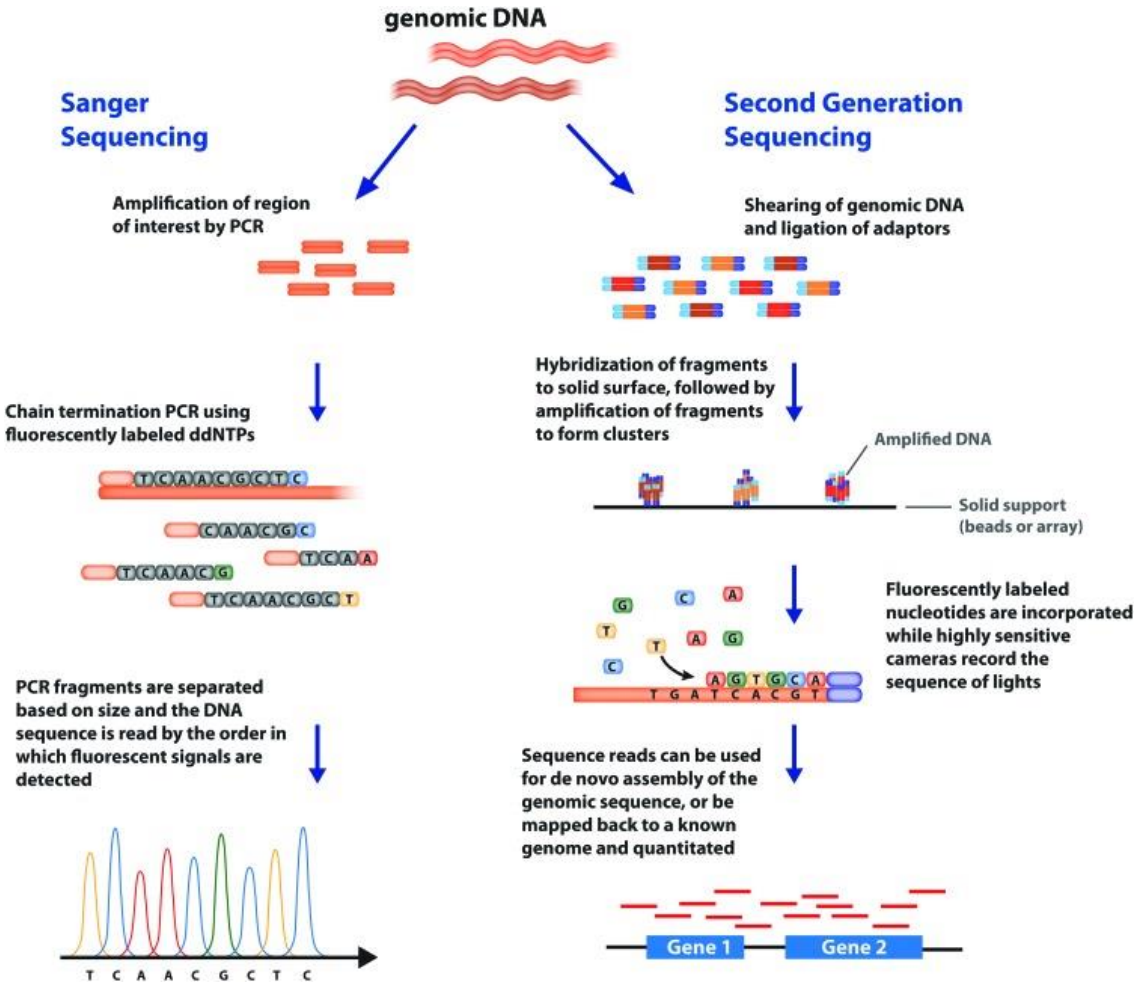


Figure 8. Comparaison entre les technologies de séquençage Sanger et de séquençage de nouvelle génération (NGS). Le séquençage de Sanger se limite à déterminer l'ordre d'un fragment d'ADN par réaction, jusqu'à une longueur maximale d'environ 700 bases. Les plateformes NGS peuvent séquencer des millions de fragments d'ADN en parallèle en une seule réaction, produisant d'énormes quantités de données **(3).**

Au fil des ans, les pipelines de séquençage ont considérablement amélioré le débit et les coûts des instruments et des réactifs, ainsi que des améliorations de la puissance de calcul, du

stockage des données et des outils bioinformatiques qui facilitent l'analyse des quantités croissantes de lectures de séquences. Ensemble, ces avancées ont entraîné une baisse spectaculaire des coûts de séquençage, qui sont tombés à environ 0,09 $ (US) par mégabase au début de 2012. Plusieurs nouvelles sociétés, telles que Helicos Biosciences, Pacific Biosciences et Oxford Nanopore Technologies, développent actuellement de nouvelles techniques de séquençage de troisième génération qui ne nécessitent pas d'amplification de l'ADN matrice, mais sont capables de lire la séquence de molécules d'ADN uniques. Ces technologies pourraient faire progresser considérablement le domaine du séquençage en réduisant considérablement le coût des réactifs et en améliorant le débit, tout en éliminant tout biais introduit lors de l'étape d'amplification du modèle du protocole NGS.

Références

1. AlMalki, F. A., Flemming, C. S., Zhang, J., Feng, M., Sedelnikova, S. E., Ceska, T., et al. (2016). Direct observation of DNA threading in flap endonuclease complexes. Nat. Struct. Mol. Biol. 23, 640–646.
2. Benoit, R. M., Ostermeier, C., Geiser, M., Li, J. S., Widmer, H., and Auer, M. (2016). Seamless insert-plasmid assembly at high efficiency and low cost. PLoS ONE 11:e0153158.
3. Bunnik EMa nd Roch GL. (2013). An Introduction to Functional Genomics and Systems Biology. Adv Wound Care (New Rochelle). 2(9): 490–498.
4. Cao Y, Kimura S, Itoi T, Honda K, Ohtake H, Omasa T. (2002). Fluorescence in situ hybridization using bacterial artificial chromosome (BAC) clones for the analysis of chromosome rearrangement in Chinese hamster ovary cells. Methods. 56, 418-423
5. Casini, A., MacDonald, J. T., De Jonghe, J., Christodoulou, G., Freemont, P. S., Baldwin, G. S., et al. (2014). One-pot DNA construction for synthetic biology: the modular overlap-directed assembly with linkers (MODAL) strategy. Nucleic Acids Res. 42:e7.
6. Citovsky, V., Lee, L. Y., Vyas, S., Glick, E., Chen, M. H., Vainstein, A., et al. (2006). Subcellular localization of interacting proteins by bimolecular fluorescence complementation in planta. J. Mol. Biol. 362, 1120–1131.
7. Cohen, S. N., Chang, A. C., Boyer, H. W., and Helling, R. B. (1973). Construction of biologically functional bacterial plasmids in vitro. Proc. Natl. Acad. Sci. U.S.A. 70, 3240–3244.
8. DNA Microarray: Principle, Types and steps involved in cDNA microarrays: https://www.onlinebiologynotes.com/dna-microarray-principle-types-and-steps-involved-in-cdna-microarrays/
9. Engler, C., Gruetzner, R., Kandzia, R., and Marillonnet, S. (2009). Golden gate shuffling: a one-pot DNA shuffling method based on type IIs restriction enzymes. PLoS ONE 4:e5553.
10. Engler, C., Kandzia, R., and Marillonnet, S. (2008). A one pot, one step, precision cloning method with high throughput capability. PLoS ONE 3:e3647.
11. Foundations of Molecular Cloning - Past, Present and Future https://international.neb.com/tools-and-resources/feature-articles/foundations-of-molecular-cloning-past-present-and-future

12. Fu, C., Donovan, W. P., Shikapwashya-Hasser, O., Ye, X., and Cole, R. H. (2014). Hot fusion: an efficient method to clone multiple DNA fragments as well as inverted repeats without ligase. PLoS ONE 9:e115318.
13. Gibson, D. G., Young, L., Chuang, R. Y., Venter, J. C., Hutchison, C. A. III., and Smith, H. O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat. Methods 6, 343–345.
14. Grefen, C., and Blatt, M. R. (2012). A 2in1 cloning system enables ratiometric bimolecular fluorescence complementation (rBiFC). Biotechniques 53, 311–314.
15. Halleran, A. D., Swaminathan, A., and Murray, R. M. (2018). Single day construction of multigene circuits with 3G assembly. ACS Synth. Biol. 7, 1477–1480.
16. Hartley, J. L., Temple, G. F., and Brasch, M. A. (2000). DNA cloning using in vitro site-specific recombination. Genome Res. 10, 1788–1795.
17. Hellens, R. P., Edwards, E. A., Leyland, N. R., Bean, S., and Mullineaux, P. M. (2000). pGreen: a versatile and flexible binary Ti vector for Agrobacterium-mediated plant transformation. Plant Mol. Biol. 42, 819–832.
18. Hsu, S. Y., and Smanski, M. J. (2018). Designing and implementing algorithmic DNA assembly pipelines for multi-gene systems. Methods Mol. Biol. 1671, 131–147.
19. Kumar A, Gupta AK & Gupta SM (2012) "Recombinant DNA technology". In: Biotechnology in medicine and agriculture: principles and practices. (eds. Kumar A, Pareek A & Gupta SM) I. K. International. publishing house Pvt. Ltd., New Delhi, India, pp. 31-59
20. Lampropoulos, A., Sutikovic, Z., Wenzl, C., Maegele, I., Lohmann, J. U., and Forner, J. (2013). GreenGate—a novel, versatile, and efficient cloning system for plant transgenesis. PLoS ONE 8:e83043.
21. Plasmid Cloning by Restriction Enzyme Digest (aka Subcloning): https://www.addgene.org/protocols/subcloning/
22. Sparkes, I. A., Runions, J., Kearns, A., and Hawes, C. (2006). Rapid, transient expression of fluorescent fusion proteins in tobacco plants and generation of stably transformed plants. Nat. Protoc. 1, 2019–2025.
23. Torella, J. P., Lienert, F., Boehm, C. R., Chen, J. H., Way, J. C., and Silver, P. A. (2014b). Unique nucleotide sequence-guided assembly of repetitive DNA parts for synthetic biology applications. Nat. Protoc. 9, 2075–2089.
24. Traditional Cloning Basics : https://www.thermofisher.com/dz/en/home/life-science/cloning/cloning-learning-center/invitrogen-school-of-molecular-biology/molecular-cloning/cloning/traditional-cloning-basics.html#vector
25. Tuo, D., Fu, L., Shen, W., Li, X., Zhou, P., and Yan, P. (2017). Generation of stable infectious clones of plant viruses by using Rhizobium radiobacter for both cloning and inoculation. Virology 510, 99–103.
26. Vazquez-Vilar, M., Quijano-Rubio, A., Fernandez-Del-Carmen, A., Sarrion-Perdigones, A., Ochoa-Fernandez, R., Ziarsolo, P., et al. (2017). GB3.0: a platform for plant bio-design that connects functional DNA elements with associated biological data. Nucleic Acids Res. 45, 2196–2209.
27. Wang, J., Xu, R., and Liu, A. (2014). IRDL cloning: a one-tube, zero-background, easy-to-use, directional cloning method improves throughput in recombinant DNA preparation. PLoS ONE 9:e107907.
28. Weber, E., Engler, C., Gruetzner, R., Werner, S., and Marillonnet, S. (2011). A modular cloning system for standardized assembly of multigene constructs. PLoS ONE 6:e16765.
29. Wold B. Myers RM. Sequence census methods for functional genomics. *Nat Methods.* 2008;5:19

Chapitre 05
Bioinformatique fonctionnelle

Chapitre 5. Bioinformatique fonctionnelle

5. Introduction

Avec l'afflux important de données de séquences brutes provenant de projets de séquençage du génome, il existe un besoin de méthodes automatiques fiables pour l'analyse et la classification des séquences de protéines. Les outils les plus utiles utilisent diverses méthodes pour identifier des motifs ou des domaines trouvés dans des familles de protéines précédemment caractérisées. Cet article passe en revue les outils et les ressources disponibles sur le Web pour identifier les signatures dans les protéines et explique comment ils peuvent être utilisés dans l'analyse de séquences de protéines nouvelles ou inconnues.

6. Analyse d'une famille de séquence protéique

En juin 2000, la première ébauche de la séquence humaine a été annoncée et a été considérée comme une réalisation égale à celle de mettre le premier homme sur la lune. L'annonce a apporté des promesses de percées dans le traitement des maladies humaines, mais en fait tout cela signifiait était un flot de données à convertir en informations biologiques utiles. Pour tenir la promesse de la séquence, les premiers obstacles sont de classer les gènes qu'elle contient et d'attribuer des fonctions aux produits des gènes. Les séquences de protéines peuvent être classées en identifiant le type de protéine, mais elles doivent ensuite être caractérisées davantage pour attribuer une fonction biologique. Le défi réside dans cette application de connaissances biologiques utiles à des séquences protéiques particulières.

6.1. Raisons de choisir l'étude de séquences protéiques

Il existe plusieurs raisons de choisir de caractériser des protéines plutôt que des séquences d'ADN. Ceux-ci incluent : le plus grand alphabet (21 acides aminés contre 4 bases) ; le rapport signal/bruit plus faible dans les recherches de séquences protéiques ; la proximité entre la séquence protéique et la fonction ; et la disponibilité de bonnes bases de données bien annotées de séquences protéiques et de signatures de séquences protéiques. Les protéines peuvent être caractérisées à différents niveaux : elles exercent une fonction dans une cellule, mais cette fonction est également exercée dans un contexte particulier, par exemple dans le cadre d'une voie complexe, ainsi qu'à un emplacement cellulaire défini. Au niveau fonctionnel, cela peut se résumer à l'analyse de la séquence protéique sur toute sa longueur, au niveau des domaines ou motifs uniques, ou au niveau le plus fin, des résidus d'acides aminés importants uniques. Avec

la disponibilité accrue de génomes entièrement séquencés et l'utilisation des outils et des ressources appropriés, il est possible de caractériser les protéines à tous ces niveaux.

La première étape de l'analyse de séquences de protéines nouvelles ou non caractérisées consiste traditionnellement à rechercher dans les bases de données de protéines des séquences similaires.

6.2. Principales bases de données de séquences de protéines

Les principales bases de données de séquences de protéines disponibles sont SWISS-PROT et TrEMBL, la Protein Information Resource (PIR) et GenPept, qui est une traduction de GenBank.

6.2.1. Modèles et profils PROSITE

PROSITE est une base de données de modèles et de profils. Les motifs PROSITE sont construits à partir d'alignements de séquences apparentées, qui proviennent de diverses sources : d'une famille de protéines bien caractérisée ; tiré de la littérature; à partir des résultats des recherches de séquences contre SWISS-PROT et TrEMBL ; ou du regroupement de séquences. Les alignements sont vérifiés pour les régions conservées, qui, en particulier pour les familles de protéines caractérisées, peuvent s'être avérées expérimentalement impliquées dans l'activité catalytique ou se lier à un substrat. Un motif central est créé sous la forme d'une expression régulière qui spécifie quel(s) acide(s) aminé(s) peut(vent) ou non apparaître à chaque position. Les expressions régulières sont des chaînes de texte qui décrivent des modèles, utilisées pour représenter un ensemble de chaînes. Ils peuvent être considérés comme similaires aux outils de correspondance de modèles génériques utilisés traditionnellement sous Unix et les utilitaires de système d'exploitation de type Unix. Les expressions régulières sont beaucoup plus élaborées et puissantes que les expressions génériques standard, mais elles sont aussi beaucoup plus complexes. Une fois le modèle de base réalisé, il est testé par rapport aux séquences de SWISS-PROT. Si le bon ensemble de protéines correspond à ce modèle, il est conservé; s'il ne parvient pas à capter certains membres de la famille ou capte trop de protéines non apparentées, le modèle est affiné et re-testé jusqu'à ce qu'il soit optimisé.

Les modèles ont de nombreux avantages, mais ils ont aussi leurs limites sur des séquences entières, c'est pourquoi PROSITE crée également des profils, pour compléter les modèles. Pour ceux-ci, le processus commence également par des alignements de séquences multiples ; il utilise ensuite une table de comparaison de symboles pour convertir les distributions de

fréquences résiduelles en poids, ce qui donne un tableau de poids spécifiques à la position. Une table de comparaison de symboles comprend des valeurs décrivant la comparaison entre des paires d'acides aminés. Le tableau a une valeur pour la qualité de correspondance de chaque paire possible d'acides aminés et est utilisé pour fournir des scores pour la probabilité qu'un acide aminé soit remplacé par un autre à une position particulière dans l'alignement de séquence. Ces chiffres sont utilisés pour calculer un score de similarité pour l'alignement entre le profil et les séquences dans SWISS-PROT ; un alignement avec un score de similarité égal ou supérieur à une valeur seuil donnée constitue un vrai succès. Le profil est ensuite affiné jusqu'à ce que seul l'ensemble prévu de séquences protéiques dépasse le seuil du profil.

6.2.2. Pfam

Pfam est une collection d'alignements de séquences de protéines multiples et de HMM, et fournit un bon référentiel de modèles pour identifier les familles, les domaines et les répétitions de protéines. La base de données Pfam comprend deux parties : PfamA, un ensemble de modèles sélectionnés et annotés manuellement ; et PfamB, qui a une couverture plus élevée mais est entièrement automatisé (sans curation manuelle). Les HMM PfamB sont créés à partir d'alignements générés par ProDom dans leur regroupement automatique des séquences protéiques dans SWISS-PROT et TrEMBL.

6.2.3. SMART

La base de données SMART ("simple modular architecture research tool") produit des HMM qui facilitent l'identification et l'annotation des domaines génétiquement mobiles et l'analyse des architectures de domaine. La base de données est très peuplée de modèles pour les domaines trouvés dans les protéines de signalisation, extracellulaires et associées à la chromatine. Les modèles reposent sur des alignements de séquences multiples sélectionnés à la main de membres représentatifs de la famille, basés sur des structures tertiaires lorsque cela est possible, mais autrement trouvés par PSI-BLAST. Une fois les modèles créés, ils sont utilisés pour rechercher dans la base de données des membres supplémentaires à inclure dans l'alignement de séquence. Ce processus itératif est répété jusqu'à ce qu'aucun autre homologue ne soit détecté.

6.2.4. TIGRFAM

Les TIGRFAM créent des HMM qui regroupent des protéines homologues qui sont conservées en termes de fonction. Les modèles sont produits de la même manière que ceux de Pfam et

SMART, mais ne doivent toucher que des équivalogues, des protéines dont il a été démontré qu'elles ont la même fonction

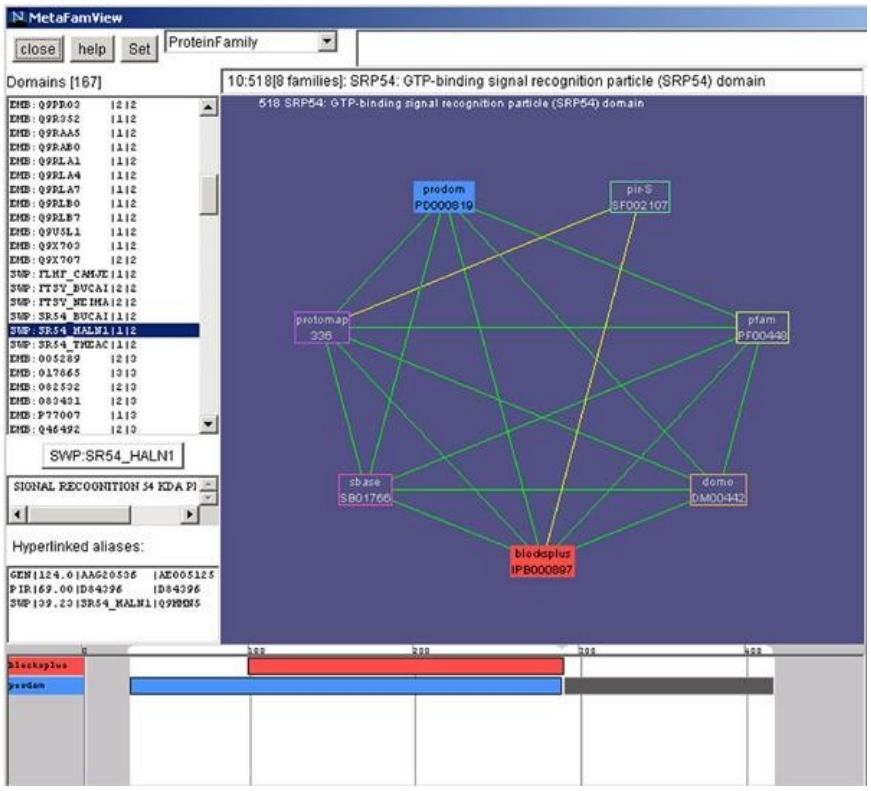
6.2.5. FingerPRINTS

La base de données PRINTS-S utilise des « empreintes digitales » comme signatures de diagnostic, dans une variante des méthodes décrites ci-dessus. Une empreinte digitale est un groupe de motifs conservés utilisés pour caractériser une famille de protéines. Plutôt que de se concentrer uniquement sur de petites zones conservées, l'occurrence de ces zones conservées sur l'ensemble de la séquence est prise en compte. Une fois de plus, le point de départ est un alignement de séquences multiples organisé. Des profils sont construits pour de petites régions conservées dans la séquence et, ensemble, ils constituent une empreinte digitale. Les « doigts », ou motifs, doivent être présents dans la séquence dans le bon ordre pour que l'empreinte digitale soit comptée comme une correspondance dans une séquence cible. Lors de la création d'empreintes digitales, chaque motif est utilisé pour scanner la base de données de séquences de protéines et les listes de résultats résultantes sont corrélées, pour ajouter des séquences à l'alignement d'origine. De nouveaux motifs sont alors générés et le processus est répété jusqu'à convergence. La reconnaissance d'éléments individuels dans l'empreinte digitale est mutuellement conditionnelle et les vrais membres correspondent à tous les éléments dans le bon ordre, tandis que les membres d'une sous-famille peuvent ne correspondre qu'à une partie de l'empreinte digitale. De nombreuses empreintes digitales ont été créées pour identifier les protéines au niveau de la superfamille ainsi qu'aux niveaux de la famille et de la sous-famille; pour cette raison, de nombreuses empreintes digitales sont liées les unes aux autres dans une structure hiérarchique ordonnée.

6.3.Exemples d’étude

- **Exemple 1**

Un exemple d'entrée de famille MetaFam. Il s'agit de l'entrée pour SRP54, le domaine des particules de reconnaissance de signal de liaison GTP, et montre les liens entre les entrées associées dans ProDom, les superfamilles PIR, Pfam, DOMO, Blocks+, SBASE et ProtoMap. La structure du domaine de la protéine SWISS-PROT SR54_HALN1 sélectionnée est indiquée au bas de l'entrée.



- **Exemple 2**

Un exemple d'entrée InterPro. Il s'agit de l'entrée IPR000402, qui décrit la sous-unité β Na+, K+ ATPase. L'entrée intègre deux modèles PROSITE et un HMM Pfam, qui sont diagnostiques pour la même famille. Il comprend un résumé décrivant cette famille et des listes de correspondance de toutes les protéines SWISS-PROT et TrEMBL appartenant à la famille. (b) La sortie du résultat de la recherche de séquence d'InterProScan. La protéine biosynthétique flagellaire d'*Escherichia coli* FliP a été scannée et s'est avérée correspondre à l'entrée InterPro IPR002039, qui décrit la famille de protéines FlIP. Les résultats sont affichés à la fois sous forme graphique et sous forme de tableau, et incluent les positions des acides aminés pour les correspondances de signature.

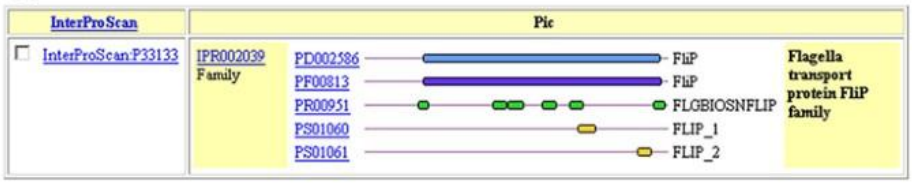
(a)

InterPro Entry IPR000402

Na+,K+ ATPase beta subunit

Database	InterPro
Accession	IPR000402; Na_K_beta (matches 71 proteins)
Name	Na+,K+ ATPase beta subunit
Type	Family
Dates	08-OCT-1999 (created) 12-MAR-2001 (last modified)
Signatures	PS00390; ATPASE_NA_K_BETA_1 (49 proteins) PS00391; ATPASE_NA_K_BETA_2 (42 proteins) PF00287; Na_K-ATPase (71 proteins)
Process	potassium transport (GO:0006813) sodium transport (GO:0006814)
Function	sodium/potassium-exchanging ATPase (GO:0005391)
Component	membrane (GO:0016020)
Abstract	The sodium pump (Na^+,K^+ ATPase), located in the plasma membrane of all animal cells [1], is an heterotrimer of a catalytic subunit (alpha chain), a glycoprotein subunit of about 34 Kd (beta chain) and a small hydrophobic protein of about 6 Kd. The beta subunit seems [2] to regulate, through the assembly of alpha/beta heterodimers, the number of sodium pumps transported to the plasma membrane. Structurally the beta subunit is composed of a charged cytoplasmic domain of about 35 residues, followed by a transmembrane region, and a large extracellular domain that contains three disulfide bonds and glycosylation sites. This structure is schematically represented in the figure below. `+-----+ +--+ +----------+` `|				|` `xxxxxxxxxxxxxxxxxxxxxxxxCxxxxCxCxxCxxxxxxxCxxxxxxxxxxCxxxx` `|-Cyt-||TM||----------Extracellular----------|`  `'C': conserved cysteine involved in a disulfide bond.`
Examples	• P05026 ATNB_HUMAN: beta-1 isoform • P14415 ATNC_HUMAN: beta-2 isoform • P51164 ATHB_HUMAN: Gastric (K+, H+) ATPase (proton pump) responsible for acid production in the stomach consist of two subunits [3] View examples
References	1. Horisberger J.D., Lemas V., Krahenbul J.P., Rossier B.C. *Structure-function relationship of Na,K-ATPase.* Annu. Rev. Physiol. 53: 565-584(1991). [MEDLINE:91254051] 2. McDonough A.A., Gerring K., Farley R.A. *The sodium pump needs its beta subunit.* FASEB J. 4: 1598-1605(1990). [MEDLINE:90201633] 3. Toh B.-H., Gleeson P.A., Simpson R.J., Moritz R.L., Callaghan J.M., Goldkorn I., Jones C.M., Martinelli T.M., Mu F.-T., Humphris D.C., Pettitt J.M., Mori Y., Masuda T., Sobieszczuk P., Weinstock J., Mantamadiotis T., Baldwin G.S. *The 60- to 90-kDa parietal cell autoantigen associated with autoimmune gastritis is a beta subunit of the gastric H+/K(+)-ATPase (proton pump).* Proc. Natl. Acad. Sci. U.S.A. 87: 6418-6422(1990). [MEDLINE:90349627]
Database links	Blocks; IPB000402 PROSITE doc; PDOC00328
Matches	Table all Graphical all Condensed graphical view

(b)



Protein	P33133 **FLIP_ECOLI** length: 245 crc64: C8E84B7B10FD1D26	
InterPro IPR002039	***Flagella transport protein FliP family*** *(Family)*	
	PRODOM: PD002586 *FliP*	[47-239]T
	PFAM: PF00813 *FliP*	[47-240]T
	PRINTS: PR00951 *FLGBIOSNFLIP*	[43-52]T [103-115]T [115-129]T [143-155]T [165-177]T [233-243]T
	PROSITE: PS01060 *FLIP_1*	[172-187]T
	PROSITE: PS01061 *FLIP_2*	[220-232]T

7. Banques de données

7.1. Définition d'une base de données

Une base de données est une archive informatisée utilisée pour stocker et organiser des données de telle manière que les informations peuvent être récupérées facilement via une variété de critères de recherche.

Les bases de données sont toutes organisées en fonction d'un modèle de données (data model), qui peut être de différents types. L'un des modèles les plus utilisés aujourd'hui est le modèle de bases de données relationnelles (SGBDR) qui a été inventé un 1970 par Edgar Frank Codd. Ces SGBDR permettent d'accéder à la base de données directement via Internet afin d'en assurer la diffusion de l'information**.**

7.2. Types des bases de données

Il existe plus de 1000 bases de données biologiques qui varient selon la taille, la qualité, la couverture, le niveau d'intérêt. Bon nombre des plus importantes sont couverts par le numéro de base de données annuelle de la revue *Nucleic Acids Research* (NAR)**.**

Les bases de données biologiques peuvent être classées de différentes manières selon la propriété prise en considération**:**

a. Le type de données

- Séquences nucléotidiques : GenBank, EMBL, DDBJ
- Séquences protéiques : SwissProt, PIR (Protein Information Ressource), trEMBL (traduction directe de EMBL), GenPept (Traduction directe de GenBank)
- Domaines et motifs protéiques : Pfam, Prosit
- Structure 3D macromoléculaire : PDB (Protein Data Bank), NDB (Nucleic Acid Data Bank)
- Données d'expression génique : Array Express (EBI) and Geo (NCBI)
- Voies métaboliques: KEGG (The Kyoto Encyclopedia of Genes and Genomes), WIT (What Is There?)
- Séquences d'intérêt immunologique: The Immunogenetics database (IMGT), The Kabat Database of Sequences of Proteins of Immunological Interest.
- Bibliographie: PubMed, OMIM (Online Mendelian Inheritance in man), Medline, HighWire.
- Génome: Ensembl (EMBL), UCSC (University of California Santa Cruz), GDB (spécialisée dans le génome de l'homme), SGD (spécialisée dans le génome de *Saccharomyces cerevisiae*), AtDB (spécialisée dans le génome de Arabidopsis), TIGR (The Institute for Genomic Research).
- Protéome : SWISS 2D PAGE
- Nomenclature: Taxonomies, Mendel

b. Les données primaires ou dérivées

- Bases de données primaires: résultats expérimentaux directement dans la base de données
- Bases de données secondaires: résultats de l'analyse des bases de données primaires
- Agrégat de nombreuses bases de données : combinaison de données

c. La conception technique

- Fichier plat (flat files) : de simples fichiers texte, aucune organisation pour faciliter la récupération des données.
- Base de données relationnelle (SQL) : données organisées sous forme de tableaux «relations». Caractéristiques communes entre les tables permet une recherche rapide.
- Base de données orientée (objet) (par exemple CORBA, XML) : données organisées comme des «objets». Objets associés hiérarchiquement.

d. Le statut de manageur

- Etablissement public (par exemple EMBL, NCBI)
- Institut quasi-académique (par exemple l'Institut Suisse de Bioinformatique, TIGR)
- Groupe académique ou scientifique
- Société commerciale

e. La disponibilité

- Accessibles au public, pas de restrictions
- Disponible, mais avec le droit d'auteur
- Accessible, mais pas téléchargeable
- Académique, mais pas librement disponible
- Commerciale; éventuellement gratuite pour les universitaires

Dans ce cours, le type de classification retenu pour caractériser les bases de données biologiques est basé sur les données primaires et secondaires à partir desquelles les bases de données sont classées en :

- bases de données primaires
- bases de données secondaires

7.2.1. Les bases de données primaires

Les bases de données primaires contiennent des données expérimentales telles que des séquences nucléotidiques, des séquences de protéines ou des structures macromoléculaires. Les résultats expérimentaux sont présentés directement dans la base de données par les chercheurs **(9).**

- **Séquences nucléiques**

- GenBank
- European Molecular Biology Lab (EMBL)
- DNA Data Bank of Japan (DDBJ)

- **Structures (proteins, ADN, ARN)**

- Protein Data Bank **(**PDB)
- Nucleic Acid Data Bank (NDB)

- **Données d'expression génique**

- Gene Expression Omnibus (GEO de NCBI)

- ArrayExpress (EBI)

- **Voies biochimiques**

- The Kyoto Encyclopedia of Genes and Genomes (KEGG)
- What Is There? (WIT)

7.2.2. Les bases de données secondaires

Les bases de données secondaires comprennent des données dérivées à partir des résultats de l'analyse de données primaires. Les bases de données secondaires tirent souvent leurs informations de nombreuses sources, y compris d'autres bases de données (primaires et secondaires). Elles sont très organisées, souvent en utilisant une combinaison complexe d'algorithmes de calcul et d'analyse manuelle, et elles sont renforcées avec une annotation plus complète des séquences**.**

- **Séquences protéiques**

- Swiss-Prot, TrEMBL et PIR: combinés en UniProt

- **Projets du séquençage génomique**

- Ensembl

Tableau 1. Les aspects essentiels des bases de données primaires et secondaire**s**

	Base de données primaire	**Base de données secondaire**
Synonymes	Base de données d'archive	Base de données organisée; base de connaissances
Sources des données	Soumission directe des données expérimentales dérivées de chercheurs	Résultats de l'analyse, la recherche et l'interprétation de la littérature, souvent des données dans les bases de données primaires
Exemples	- ENA, GenBank et DDBJ - ArrayExpress Archive - GEO - Protein Data Bank (PDB)	- InterPro (familles de protéines, des motifs et des domaines) - UniProt Knowledgebase (la séquence et les informations fonctionnelles des protéines) -Ensembl (variation, fonction, régulation des séquences du génome entier)

7.3.GenBank

7.3.1. Définition

GenBank® est une base de données publique de toutes les séquences nucléotidiques et protéiques connues avec annotation bibliographique et biologique à l'appui, construite et distribuée par le National Center for Biotechnology Information (NCBI), une division de la National Library of Medicine (NLM), située sur le campus des National Institutes of Health (NIH) des États-Unis. Le NCBI a été créé par le Congrès en 1988 pour développer des systèmes d'information, tels que GenBank, afin de soutenir la communauté de la recherche biomédicale. Le NCBI a également été mandaté pour mener des recherches fondamentales et appliquées et, dans le cadre du programme intra-muros des NIH, les scientifiques du NCBI travaillent dans les domaines de l'analyse des gènes et du génome, de la biologie structurale computationnelle et des méthodes mathématiques pour l'analyse des séquences.

Le NCBI construit GenBank principalement à partir de la soumission directe des données de séquence des auteurs et secondairement à partir de la numérisation de la littérature des revues. Une source majeure de données sont les soumissions en masse d'EST et d'autres données à haut débit des centres de séquençage. Les données sont complétées par des séquences provenant d'autres bases de données publiques. Grâce à une collaboration internationale avec la bibliothèque de données EMBL au Royaume-Uni et la banque de données ADN du Japon (DDBJ), les données sont échangées quotidiennement pour garantir que les trois sites conservent des ensembles complets d'informations sur les séquences. Les données sont mises à disposition gratuitement sur Internet, soit en téléchargeant des fichiers de base de données, soit par des services de recherche de similarité de texte et de séquence.

7.3.2. Organisation de la base de données GenBank

GenBank a connu une nouvelle année de croissance sans précédent. Au cours des 12 derniers mois, 420 000 nouvelles séquences ont été ajoutées. Depuis la version 96 d'août 1996, GenBank contenait 602 072 354 bases nucléotidiques de 920 588 séquences différentes. Notamment, 1996 est l'année où le génome complet d'un organisme eucaryote, la levure Saccharomyces cerevisiae a été complété et ajouté à GenBank (2). Les séquences génomiques complètes d'un archéon, Methanococcus jannaschii (3) sont également entrées dans la base de données cette année (voir section Génomes ci-dessous). Historiquement, la taille de la base de données doublait environ tous les 18 mois, mais ce rythme s'est rapidement accéléré en raison de l'énorme croissance des données provenant des étiquettes de séquence exprimées (EST). Plus de 65% des séquences de la version actuelle sont des EST et la majeure partie de la croissance en termes d'enregistrements de séquences au cours des 2 dernières années provient du projet collaboratif entre Merck & Co. et l'Université de Washington (4,5). Cette croissance devrait se

poursuivre car l'Université de Washington et le Howard Hughes Medical Institute poursuivent un projet EST de souris de la même envergure que le séquençage EST humain. En outre, le projet du génome humain est entré dans sa phase pilote de séquençage à grande échelle et 100 millions de nucléotides de données de séquence d'ADN génomique humain sont attendus au cours des 2 prochaines années. Les enregistrements de séquences de plusieurs des six centres américains financés pour faire ce travail commencent déjà à apparaître dans GenBank.

8. Notion d'analyse phylogénétique

8.1. Introduction à la phylogénétique moléculaire

La phylogénétique est la science qui étudie la relation évolutive entre les espèces. Afin de faire des prédictions sur ces relations, des arbres phylogénétiques reliant les espèces sont construits.La relation établie entre deux espèces est classée comme une phylogénie. Un arbre phylogénétique est une représentation arborescente binaire de la relation résultante.

Traditionnellement, les caractères morphologiques ont été utilisés pour déduire la phylogénie, et cela en utilisant souvent des fossiles, qui contiennent des informations morphologiques sur les ancêtres des espèces actuelles. Actuellement, les séquences moléculaires fournissent des ensembles de caractères qui portent une grande quantité d'informations. Si nous avons un ensemble de séquences de différentes espèces, par conséquent, nous pouvons être en mesure de les utiliser pour déduire une phylogénie probable des espèces en question.

La phylogénétique moléculaire peut être définie comme l'étude des relations évolutives de gènes et d'autres macromolécules biologiques en analysant les mutations à différentes positions dans leurs séquences et en élaborant des hypothèses au sujet de la parenté évolutive des biomolécules. Sur la base de la similarité de séquence des molécules, les relations évolutives entre les organismes peuvent souvent être déduites.

Mais un arbre phylogénétique d'un groupe de séquences ne reflète pas nécessairement l'arbre phylogénétique des espèces hôtes, car la duplication de gènes est un autre mécanisme, en plus de la spéciation, par lequel deux séquences divergent à partir d'un ancêtre commun. Les gènes qui ont divergé en raison de la spéciation sont appelés orthologues. Les gènes qui divergent par la duplication de gènes sont appelés paralogues. Si nous sommes intéressés à inférer l'arbre phylogénétique des espèces, nous devons utiliser des séquences orthologues.

Mais bien sûr, nous pourrions être intéressés par la phylogénie des événements de duplication, dans ce cas, nous pourrions construire une phylogénie de paralogues.

8.2.Terminologie

Les relations évolutives entre les séquences étudiées sont représentées par des arbres phylogénétiques. Un arbre phylogénétique typique est représenté dans la **figure 1.**

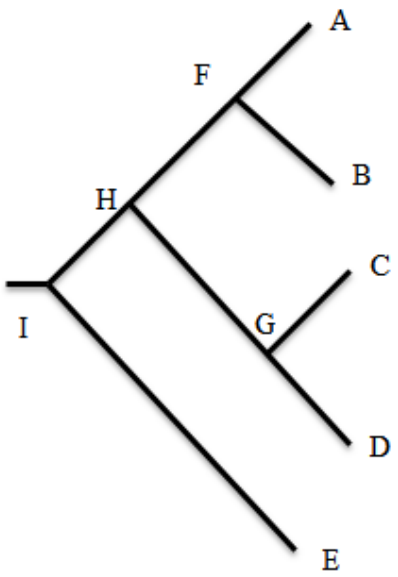


Figure 1. Arbre phylogénétique typique

Les arbres sont des graphes composés de :

- **Les taxa (singulier : taxon) ou les unités taxonomiques opérationnelles (OTU):** situés à l'extrémité des branches et ce sont des espèces ou des séquences actuelles (A, B, C, D et E)
- **Les nœuds ou les unités taxonomiques hypothétiques (HTU):** qui sont les points de jonction où deux branches adjacentes se joignent ce qui représente un ancêtre inférées de taxa actuels (F et G). Le point bas de l'arbre bifurquant est le nœud racine, qui représente l'ancêtre commun de tous les membres de l'arbre (I).
- **Les branches** : qui sont les lignes de l'arbre ; et représentent les relations de parentés (ancêtre /descendants).
- **Clade ou groupe monophylétique**: c'est un groupe de taxons descendant d'un ancêtre commun unique (ex : A/B et C/D).

- **Groupe paraphylétiques** : quand un certain nombre de taxa partagent plus d'un seul ancêtre commun proche, ils ne correspondent pas à la définition d'un clade. Dans ce cas, ils sont appelés groupes paraphylétiques (par exemple, les taxa C, D et E).
- **La topologie de l'arbre** : est l'ensemble des branchements de cet arbre.

8.3.Arbres enracinés et Arbres non enracinés

Un arbre phylogénétique peut être soit enraciné ou sans racines. Un arbre phylogénétique non enraciné ne suppose pas la présence d'un ancêtre commun, mais seulement positionne les taxa et montre leurs relations relatives. Dans un arbre non enraciné, il n'y a pas de direction d'un chemin d'évolution parce qu'il n'y a aucun ancêtre commun. Pour définir la direction d'un chemin d'évolution, un arbre doit être enraciné. Dans un arbre enraciné, toutes les OTU étudiés ont un ancêtre commun (Cenancestor = MRCA (Most Recent Common Ancestor)) à partir duquel un chemin unique d'évolution conduit à chaque OTU. De toute évidence, un arbre enraciné est plus instructif qu'un arbre sans racine (**Figure 2**).

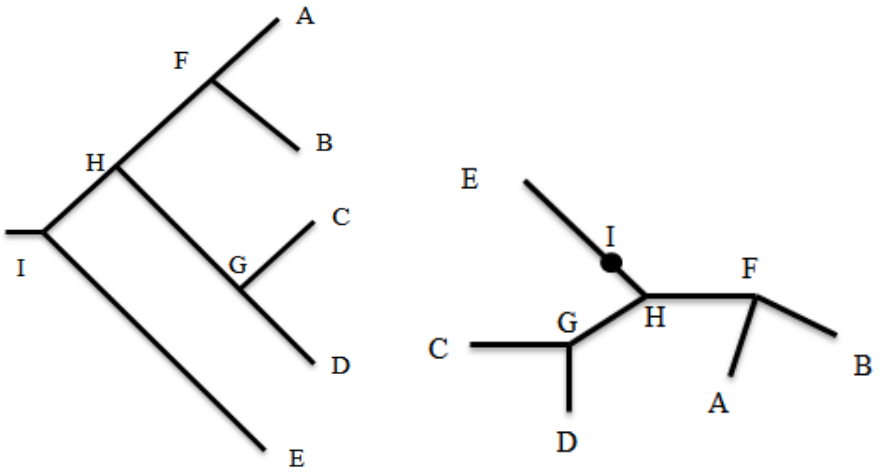


Figure 2. Représentation d'un arbre enraciné et un arbre non enraciné

8.4.Enracinement d'un arbre phylogénétique par un outgroup

Pour convertir un arbre non enraciné à un arbre enraciné, on peut se servir d'un outgroup qui est une séquence homologue aux séquences étudiées, mais qui s'est séparés de ces séquences à un moment de l'évolution précoce. Par exemple, une séquence d'oiseaux peut être utilisée comme une racine pour l'analyse phylogénétique des mammifères (**Figure 3**).

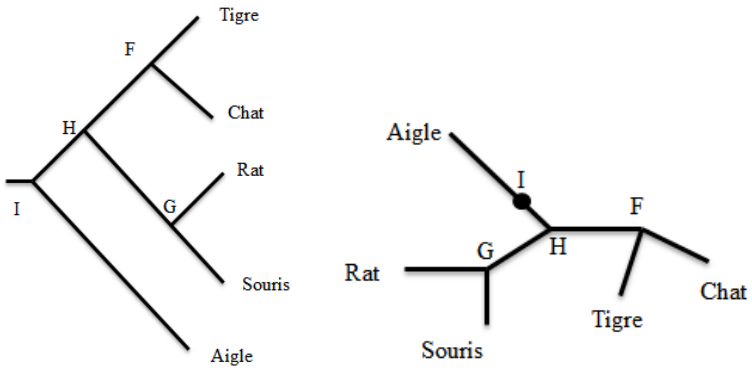


Figure 3. Enracinement d'un arbre phylogénétique de mammifères en utilisant un oiseau comme outgroup

8.5. Enracinement d'un arbre phylogénétique par l'approche de l'enracinement au milieu « midpoint rooting »

En l'absence d'une bon outgroup, un arbre peut être enraciné en utilisant l'approche d'enracinement au milieu « midpoint rooting », dans lequel le milieu des deux groupes les plus divergents jugés par des longueurs globales des branches est affecté comme la racine. Ce type d'enracinement suppose que la divergence de la racine jusqu'au OTU des deux branches est égale et suit l'hypothèse de l'horloge moléculaire.

8.6. Types topologiques des arbres phylogénétiques

Les arbres peuvent être dessinés de deux manières, soit comme un cladogramme ou un phylogramme (**Figure 4**).

a- **Le phylogramme** : c'est une topologie où les longueurs de branches représentent la quantité de divergence évolutive. Ces arbres sont dits être mis à l'échelle. Les arbres mis à l'échelle ont l'avantage de montrer à la fois les relations et les informations sur le temps de divergence des taxa.

b- **Le cladogramme :** dans ce type d'arbre les taxa extérieurs alignent parfaitement dans une ligne ou une colonne. Leurs longueurs de branches ne sont pas proportionnelles au nombre de changements évolutifs et ont donc aucune signification phylogénétique. Dans ce type topologique sans échelle, seule l'ordre relatif des taxa est représenté **(3).**

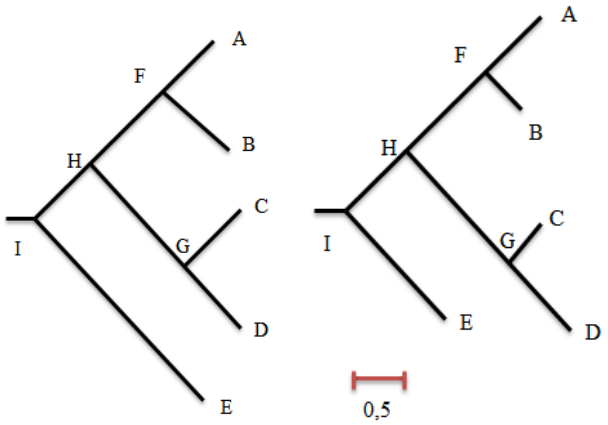


Figure 4. A gauche : représentation sans échelle (cladogramme) ; A droit : représentation avec échelle (phylogramme)

8.7.Procédure de l'établissement d'un arbre phylogénétique

La construction d'un arbre phylogénétique peut être divisée en quatre étapes:

- réalisation d'alignement de séquences multiples
- calcul de la distance entre les OTUs (séquences)
- détermination d'une méthode de construction de l'arbre
- évaluation de la fiabilité de l'arbre.

8.7.1. Calcul de la distance entre les OTUs (séquences)

b- Cas de séquences nucléiques

Une mesure simple de la divergence entre les deux séquences est de compter le nombre de substitutions dans un alignement. La proportion des substitutions définit la distance observée entre les deux séquences. Cependant, le nombre observé de substitutions ne peut pas représenter les véritables événements évolutifs qui ont effectivement eu lieu. Ces scénarios peuvent etre en cause :

- dans une mutation comme A remplacé par C, le nucléotide peut avoir fait l'objet d'un certain nombre d'étapes intermédiaires pour devenir C, comme A → T → G → C.
- une mutation de retour aurait pu se produire quand un nucléotide muté revenue au nucléotide d'origine. Par exemple si le nucléotide G identique est observé, des mutations de types G → C → G peuvent avoir réellement eu lieu.

- un nucléotide identique observé dans l'alignement pourrait être due à des mutations parallèles. Par exemple lorsque les deux résidus des séquences se mutent en T.

Ces multiples scénarios obscurcissent l'estimation des véritables distances évolutives entre les séquences. Cet effet est connu comme l'homoplasie, qui, non corrigée, peut conduire à la génération d'arbres incorrects. Pour corriger l'homoplasie, des modèles statistiques sont nécessaires pour déduire les vraies distances évolutives entre les séquences **(3).**

➢ **Calcul de distance par le modèle le de Jukes-Cantor**

Le modèle de substitution de nucléotides le plus simple possible, introduit par Jukes et Cantor en 1969, précis que les fréquences d'équilibre des quatre nucléotides sont de 25% chacune et que pendant l'évolution, tous les nucléotides ont la même probabilité d'être remplacé l'un par un autre. Ces hypothèses correspondent à une matrice avec πA = πC = πG = πT = ¼. La formule de calcul des distances évolutives est**:**

$$\boldsymbol{d_{AB} = -(3/4)\, ln[1 - \left(\frac{4}{3}\right) pAB]}$$

où :

- d_{AB} est la distance évolutive entre les séquences A et B
- p_{AB} est la distance de séquence observée mesurée par la proportion de substitutions sur toute la longueur de l'alignement.

➢ **Exemple :** soit l'alignement de deux séquences A et B

SeqA :	A	C	C	T	T	G	C	T	T	A	C	C	G	A	C
SeqB :	A	T	C	T	C	A	C	A	T	A	C	T	C	A	C

$$p_{AB} = 6 / 15 = 0{,}4 \ (40\%)$$

La correction de la distance évolutive entre A et B par le modèle Jukes-Cantor donne : $d_{AB} = -(3/4)\, ln[1 - \left(\frac{4}{3}\right) 0{,}4] = 0{,}5$.

Le modèle Jukes-Cantor ne peut traiter que des séquences étroitement liées. Si par exemple deux séquences d'ADN ont seulement 25% de similarité, P_{AB} est 0,75. Cela conduit à une valeur (d) égale à l'infiniment**.**

➢ **Calcul de la distance par le modèle de Kimura à deux paramètres**

Un autre modèle pour corriger les distances évolutives est le modèle de Kimura à deux paramètres. Dans ce modèle les taux de mutation pour les transitions et les transversions sont supposés être différents. Selon ce modèle, les transitions se produisent plus fréquemment que les transversions. Le modèle Kimura utilise la formule suivante **(10):**

$$\boldsymbol{d_{AB} = -(1/2)\, ln(1 - 2p_{ti} - p_{tv}) - (1/4)\, ln(1 - 2p_{tv})}$$

où :

- d_{AB} est la distance évolutive entre les séquences A et B,
- p_{TI} est la fréquence observée pour la transition,
- p_{TV} est la fréquence observée de transversion.

Dans l'exemple précédent de l'alignement des séquences A et B :

p_{TI} = 4/15 = 0.26 (26%) et p_{TV} = 2/15 = 0,13 (13%) donc :

$$d_{AB} = -\left(\frac{1}{2}\right) ln\left(1 - 2\,(0{,}26) - (0{,}13)\right) - \left(\frac{1}{4}\right) ln\left(1 - 2(0{,}26)\right) = 0{,}7$$

8.8.Détermination d'une méthode de construction de l'arbre phylogénétique

Les méthodes de constructions d'arbres phylogénétiques peuvent être divisées en deux groupes :

a. les méthodes basées sur les mesures de distances
b. les méthodes basées sur les caractères

8.8.1. Méthodes basées sur les distances

Pour l'analyse phylogénétique, le score de distance entre deux séquences est utilisé. Ce score entre les deux séquences est le nombre de positions de mismatch dans l'alignement. Les gaps peuvent être ignorés dans ce calcul ou traités comme des substitutions. Les vraies distances évolutives entre les séquences peuvent être calculées à partir des distances observées après correction en utilisant les modèles évolutifs. Les distances évolutives calculées peuvent être utilisées pour construire une matrice de distances entre toutes les paires individuelles de taxa. Sur la base des scores de distance par paires dans la matrice, un arbre phylogénétique peut être construit pour tous les taxa concernés

- **Unweight Pair Group Method with Arithmetic mean (UPGMA)**

Cette méthode est utilisée pour reconstruire des arbres phylogénétiques si les séquences ne sont pas trop divergentes. L'UPGMA utilise un algorithme de clustérisation séquentiel dans

lequel les relations sont identifiées dans l'ordre de leur similarité et la reconstruction de l'arbre se fait pas à pas grâce à cet ordre. Il y a d'abord identification des deux séquences les plus proches et ce groupe est ensuite traité comme un tout, puis on recherche la séquence la plus proche et ainsi de suite jusqu'à ce qu'il n'y ait plus que deux groupes (**13**).

La méthode UPGMA commence par calculer les longueurs de branches entre les séquences les plus étroitement liées, puis fait la moyenne de la distance entre cette paire ou la séquence ou le cluster de séquence suivant, et continue jusqu'à ce que toutes les séquences soient incluses dans l'arbre. Enfin, le procédé prédit une position de la racine de l'arbre (**11**).

- **Exemple :** soit les cinq séquences (taxa) A, B, C, D et E

Seq A : ACGCGTTGGGCGATGGCAAC
Seq B : ACGCGTTGGGCGACGGTAAT
Seq C : ACGCATTGAATGATGATAAT
Seq D: ACACATTGAGTGATAATAAT
Seq E: AGGTCATGGATCAGAACTAC

<u>Etape 01</u> : réalisation de l'alignement multiple

SeqA	A	C	G	C	G	T	T	G	G	G	C	G	A	T	G	G	C	A	A	C
SeqB	A	C	G	C	G	T	T	G	G	G	C	G	A	C	G	G	T	A	A	T
SeqC	A	C	G	C	A	T	T	G	A	A	T	G	A	T	G	A	T	A	A	T
SeqD	A	C	A	C	A	T	T	G	A	G	T	G	A	T	A	A	T	A	A	T
SeqE	A	C	A	T	A	A	T	G	A	A	T	G	A	C	A	A	C	A	A	T

<u>Etape 02</u> : établissement de la matrice des distances

On commence par calculer les distances entre toutes les paires de séquences en utilisant l'un des modèle évolutif (ex : Jukes-Cantor) selon l'équation :

$$d_{AB} = -(3/4)\, ln[1 - \left(\frac{4}{3}\right) pAB]$$

- p_{AB}= 0,15 donc d_{AB} = 0,16
- p_{AC}= 0,4 donc d_{AC} = 0,57
- p_{AD}= 0,4 donc d_{AD} = 0,57
- p_{AE}= 0,55 donc d_{AE} = 0,99
- p_{BC}= 0,3 donc d_{BC} = 0,38
- p_{BD}= 0,35 donc d_{BD} = 0,47
- p_{BE}= 0,5 donc d_{BE} = 0,82
- p_{CD}= 0,2 donc d_{CD} = 0,12

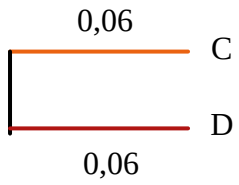
- p_{CE}= 0,3 donc d_{CE} = 0,38
- p_{DE}= 0,25 donc d_{AB} = 0,30

On remplit la matrice de distance :

	A	B	C	D
A	00			
B	0,16	00		
C	0,57	0,38	00	
D	0,57	0,47	0,12	00
E	0,99	0,82	0,38	0,30

<u>Etape 03</u> : établissement de l'arbre UPGMA

En utilisant la matrice de distance impliquant les cinq taxa, A, B, C, D et E le procédé UPGMA joint les deux taxa qui sont les plus rapprochées C et D (0,12 en gris). Parce que tous les taxa sont équidistants à partir du nœud, la longueur des branches C et D au nœud est d_{CD} / 2 = 0,12 / 2 = 0,06.



Les taxa C et D sont réunis en un groupe, ils sont traités comme un nouveau taxon unique, qui est utilisé pour créer une matrice réduite. La distance du cluster CD à tous les autres taxa est la moitié de la somme de distances de chaque taxon à C et D. Cela signifie que la distance de A à CD est (AC + AD) / 2; et celle de B à CD est (BC + BD) / 2 ; et de E à CD est (EC+ ED) / 2

- $d_{A-(C-D)=}$ 0,57+ 0,57/2 = 0,57
- $d_{B-(C-D)=}$ 0,38+ 0,47/2 = 0,42
- $d_{E-(C-D)=}$ 0,38+ 0,30/2 = 0,22

	C-D	A	B
A	0,57	00	
B	0,42	0,16	00

E	0,22	0,99	0,82

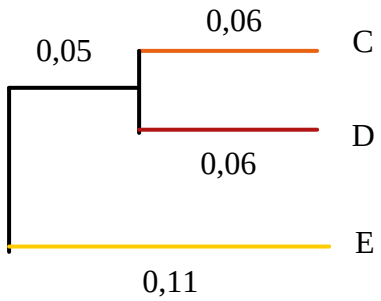
Dans la matrice réduite nouvellement établie, la distance la plus petite est entre B et A (en gris), ce qui permet le regroupement de A et B pour créer un cluster de AB. La longueur de la branche A au nœud est $d_{AB} / 2 = 0,16 / 2 = 0,08$.

Lorsque A et B sont regroupés et traités comme un seul taxon, cela permet à la matrice de se réduire davantage dans seulement trois taxa, DC, E et AB. Les nouvelles distances seront :

- AB à CD : $d_{A-CD} + d_{B-CD} / 2 = 0,57+0,42/ 2 = 0,50$
- AB à E : $d_{A-E} + d_{B-E} / 2 = 0,99 + 0,82 /2 = 0,90$

	CD	AB
AB	0,50	00
E	0,22	0,90

La plus petite distance est entre le taxon CD et le taxon E, donc les deux taxa seront réunis dans un nouveau cluster où la longueur des branches du nœud est : $d_{CD-E}/2 = 0,22 /2 = 0,11$



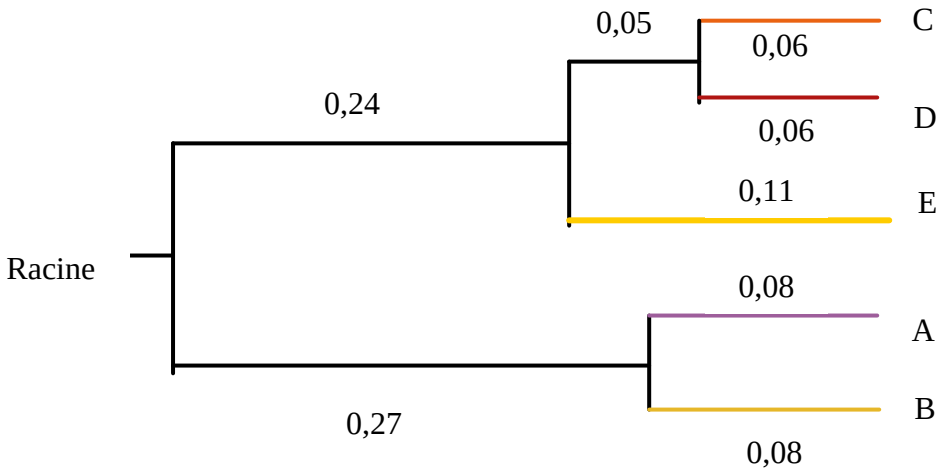
CD et E sont regroupés et traités comme un seul taxon, cela conduit à une nouvelle réduction de la matrice en seulement deux taxa, DCE et AB. Les nouvelles distances seront :

- AB à CD-E : $d_{AB-CD} + d_{AB-E} / 2 = 0,50+0,90/ 2 = 0,70$

	AB
CD-E	0,70

La distance entre les deux taxa restant par apport à leur nœud est :

- $d_{AB-CDE}/ 2 = 0,7/2 = 0,35$



➢ **Neighbor-Joining (NJ)**

Cette méthode développée par Saitou et Nei (1987) tente de corriger la méthode UPGMA afin d'autoriser un taux de mutation différent sur les branches.

La méthode NJ construit un arbre en trouvant séquentiellement des paires de voisins, qui sont les paires d'OTU connectées par un seul nœud intérieur. Le mode de regroupement utilisé par cet algorithme est très différent de celui décrit dans la méthode UPGMA, car il ne tente pas de regrouper les OTUs les plus proches, mais plutôt de minimiser la longueur de toutes les branches internes et donc la longueur de l'arbre entier. L'algorithme NJ commence par l'hypothèse d'un arbre semblable à une étoile qui n'a pas de branches internes. Dans la première étape, il introduit la première branche interne et calcule la longueur de l'arbre résultant. L'algorithme connecte séquentiellement toutes les paires d'OTUs possibles et finalement joint la paire d'OTUs qui produit l'arbre le plus court. La longueur d'une branche joignant une paire de voisins, X et Y à leur nœud adjacent est basée sur la distance moyenne entre tous les OTUs et X pour la branche à X, et tous les OTU et Y pour la branche à Y, en soustrayant les distances moyennes de toutes les paires OTU restantes. Ce processus est ensuite répété, en joignant toujours deux OTUs (voisins) en introduisant la branche interne la plus courte possible.

Le calcul des taux d'évolution inégales entre les séquences se fait en utilisant une étape de conversion. Cette conversion nécessite le calcul des valeurs de divergence « r ».

La valeur **r** est calculée sur la base de la formule suivante :

$$r_i = \sum d_{ij}$$

- où i et j sont deux taxa différents

La distance convertie est calculée par la formule :

$$M_{ij} = d_{ij} - \frac{r_i + r_j}{N - 2}$$

où :

- M_{ij} est la distance convertie entre i et j
- d_{ij} est la distance évolutive réelle entre i et j
- $\mathbf{r_i}$ **(ou** $\mathbf{r_j}$**)** est la somme des distances de i (ou j) pour tous les autres taxa.
- **N** est le nombre de taxa.

La longueur de la branche entre le nouveau nœud U et le taxon i, ainsi qu'entre U et le taxon j est calculé selon la formule suivante :

$$S_{iU} = d_{ij} / 2 + (r_i - r_j) / 2\,(N - 2)$$

La distance entre le nœud U est les autres OTUs est calculée par la formule suivante :

$$d_{kU} = (d_{jk} + d_{jk} - d_{ij}) / 2$$

- **Exemple :** soit la matrice de distances précédente

	A	B	C	D
A	00			
B	0,16	00		
C	0,57	0,38	00	
D	0,57	0,47	0,12	00
E	0,99	0,82	0,38	0,30

Cycle 01 :

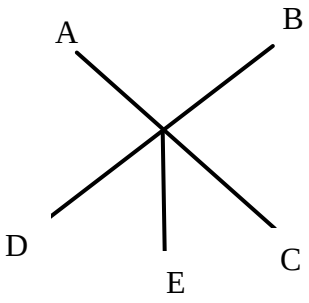
Etape 01 : En utilisant la même matrice de distance utilisée pour la construction de l'arbre UPGMA, la première étape de la méthode NJ est le calcul les valeurs de divergence « r ».

- $r_A = d_{AB}+d_{AC}+d_{AD}+d_{AE} = 0,16+0,57+0,57+0,99 = 2,29$
- $r_B = d_{BA}+d_{BC}+d_{BD}+d_{BE} = 0,16+0,38+0,47+0,82 = 1,82$
- $r_C = d_{CA}+d_{CB}+d_{CD}+d_{CE} = 0,57+0,38+0,12+0,38 = 1,45$
- $r_D = d_{DA}+d_{DB}+d_{DC}+d_{DE} = 0,57+0,47+0,12+0,38 = 1,54$
- $r_E = d_{EA}+d_{EB}+d_{EC}+d_{ED} =0,99+0,82+0,38+0,30= 2,49$

Etape 02: calcul des taux de distances corrigés et construction de la nouvelle matrice de distances

- $M_{AB} = d_{AB} - (r_A + r_B) / (N - 2) = 0,16 – (2,29+1,82) / 3 = - 1,21$
- $M_{AC} = d_{AC} - (r_A + r_C) / (N - 2) = 0,57 – (2,29+ 1,45) / 3 = - 0,67$
- $M_{AD} = d_{AD} - (r_A + r_D) / (N - 2) = 0,57 – (2,29+ 1,54) / 3 = - 0,70$
- $M_{AE} = d_{AE} - (r_A + r_E) / (N - 2) = 0,99 – (2,29+ 2,49) / 3 = - 0,60$
- $M_{BC} = d_{BC} - (r_B + r_C) / (N - 2) = 0,38 – (1,82+ 1,45) / 3 = - 0,71$
- $M_{BD} = d_{BD} - (r_B + r_D) / (N - 2) = 0,47 – (1,82+ 1,54) / 3 = - 0,62$
- $M_{BE}= d_{BE} - (r_B + r_E) / (N - 2) = 0,82 – (1,82+ 2,49) / 3 = - 0,61$
- $M_{CD} = d_{CD} - (r_C + r_D) / (N - 2) = 0,12 – (1,45+ 1,54) / 3 = - 0,87$
- $M_{CE} = d_{CE} - (r_C + r_E) / (N - 2) = 0,38– (1,45+ 2,49) / 3 = - 0,93$
- $M_{DE} = d_{DE} - (r_D + r_E) / (N - 2) = 0,30 – (1,54+ 2,49) / 3 = - 1,04$

	A	B	C	D
A	00			
B	-1,21	00		
C	-0,67	-0,71	00	
D	- 0,70	-0,62	-0,87	00
E	-0,60	-0,61	-0,93	-1,04



Avant la construction de l'arbre, tous les nœuds possibles sont effondrés dans un arbre en étoile. La paire de taxa avec les distances les plus courtes dans la nouvelle matrice est séparée en première de l'arbre en étoile, et cela selon les distances corrigées. Dans ce cas, A et B (-1,21, en gris) ensuite D et E sont les plus courtes. Par conséquent, le premier nœud à construire est A-B.

Etape 03 : calcul des longueurs de branches du nœud U à i et j

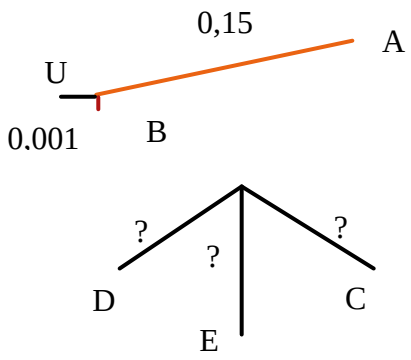
Choisir les taxa voisins les lus proches, c'est à dire des deux OTUs ayant le M(i,j) le plus petit, donc A et B. On prend A et B et on forme un nouveau nœud U et on calcule la longueur de la branche entre U et A ainsi qu'entre U et B :

- $S_{AU} = d_{AB}/2 + (r_A - r_B)/2(N-2) = 0,16/2 + (2,29\text{-}1,82)/\ 2(3) = 0,15$
- $S_{BU} = d_{AB} - S_{AU} = 0,16\text{-}\ 0,15 = 0,001$

Etape 04 : calcul des nouvelles distances entre U est les autres OTUs

- $d_{CU} = (d_{AC} + d_{BC} - d_{AB}) / 2 = \ 0,57+0,38\text{-}0,16/2 = 0,40$
- $d_{DU} = (d_{AD} + d_{BD} - d_{AB})/\ 2 = \ 0,57+0,47\text{-}0,16/2 = 0,44$
- $d_{EU} = (d_{AE} + d_{BE} - d_{AB})/\ 2 = 0,99+0,82\text{-}0,16/2 = 0,82$

	U	C	D
C	0,40	00	
D	0,44	0,12	00
E	0,82	0 ,38	0,30



Cycle 02 : Dans le deuxième cycle de calcul, les étapes de 1 à 4 sont répétées.

Etape 01 : calcul des valeurs de divergence « r »

- $r_U = d_{UC} + d_{UD} + d_{UE} = 0,40+0,44+0,82 = 1,66$
- $r_C = d_{UC} + d_{CD} + d_{CE} = 0,44+0,12+0,38 = 0,94$
- $r_D = d_{UD} + d_{CD} + d_{DE} = 0,44+0,12+0,30 = 0,86$
- $r_E = d_{UE} + d_{CE} + d_{DE} = 0,82+0,38+0,30 = 1,5$

Etape 02: calcul des taux de distances corrigés et construction de la nouvelle matrice de distances

- $M_{UC} = d_{UC} - (r_U + r_C) / (N-2) = 0,40 - (1,66+0,94) / 2 = \text{-}\ 0,90$
- $M_{UD} = d_{UD} - (r_U + r_D) / (N-2) = 0,44 - (1,66+0,86) / 2 = \text{-}\ 0,82$
- $M_{UE} = d_{UE} - (r_U + r_E) / (N-2) = 0,82 - (1,66+1,5) / 2 = \text{-}\ 0,76$
- $M_{CD} = d_{CD} - (r_C + r_D) / (N-2) = 0,12 - (0,94+0,86) / 2 = \text{-}\ 0,78$

- $M_{CE} = d_{CE} - (r_C + r_E) / (N - 2) = 0,38 - (0,94+1,5) / 2 = - 0,84$
- $M_{DE} = d_{DE} - (r_D + r_E) / (N - 2) = 0,30 - (0,86+1,5) / 2 = - 0,88$

	U	C	D
C	-0,90	00	
D	- 0,82	- 0,78	00
E	- 0,76	- 0,84	- 0,88

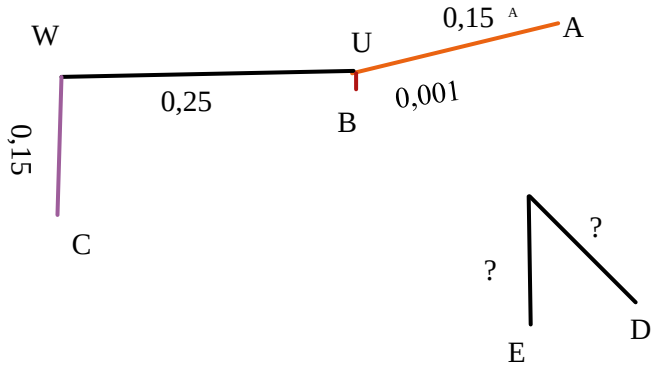
Etape 03 : calcul des longueurs de branche de nœud W à U et à C

- $S_{UW} = d_{UC}/2 + (r_U - r_C)/2(N - 2) = 0,40/2 + (1,16-0,94)/ 2(2) = 0,25$
- $S_{CW} = d_{UC} - S_{UW} = 0,40- 0,25 = 0,15$

Etape 04 : calcul des nouvelles distances entre W(CU) est les autres OTUs

- $d_{DW} = (d_{DC} + d_{DU} - d_{CU}) / 2 =$ 0,12+0,44-0,40/2 = 0,08
- $d_{EW} = (d_{EC} + d_{EU} - d_{CU})/ 2 =$ 0,38+0,82-0,40/2 = 0,4

	W	D
D	0,08	00
E	0,4	0,30



Cycle 03 :

Etape 01 : calcul des valeurs de divergence « r »

- $r_W = d_{WD} + d_{WE} = 0,08+0,4 = 0,48$
- $r_D = d_{DW} + d_{DE} = 0,08+0,30 = 0,38$
- $r_E = d_{EW} + d_{ED} = 0,4+0,30 = 0,70$

Etape 02: calcul des taux de distances corrigés et construction de la nouvelle matrice de distances

- $M_{WD} = d_{WD} - (r_W + r_D) / (N - 2) = 0,08 – (0,48+0,38) / 1 = - 0,78$
- $M_{WE} = d_{WE} - (r_W + r_E) / (N - 2) = 0,4 – (0,48+0,70) / 1 = - 0,78$
- $M_{DE} = d_{DE} - (r_D + r_E) / (N - 2) = 0,30 – (0,38+0,70) / 1= - 0,78$

	W	D
D	-0,78	00
E	-0.78	-0,78

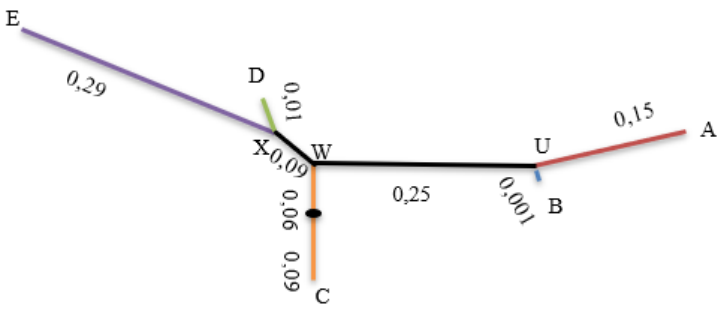
Etape 03 : Définir un nouveau nœud: par exemple E et D sont des voisins et forment un nouveau nœud X. Alternativement, W et D ou W et E pourrait être rejoint.

- $S_{DX} = d_{ED}/2 + (r_D – r_E)/2(N - 2) = 0,30/2 + (0,38-0,70)/ 2(1) = 0,01$
- $S_{EX} = d_{ED} – S_{DX} = 0,30- 0,01 = 0,29$

Etape 04 : calcul des nouvelles distances entre X(ED) et les autres OTUs

- $d_{WX} = (d_{WE} + d_{WD} – d_{ED}) / 2 =$ 0,4+0,08-0,30/2 = 0,09

	W
X	0,09



Si on considère C comme outgroup et on place la racine entre C et (A, B, D et E), la topologie de l'arbre enraciné est obtenue.

Références bibliographiques

1. Bairoch A, Apweiler R: The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. 2000, 28: 45-48. 10.1093/nar/28.1.45.
2. Barker WC, Garavelli JS, Hou Z, Huang H, Ledley RS, McGarvey PB, Mewes HW, Orcutt BC, Pfeiffer F, Tsugita A, et al: Protein Information Resource: a community resource for expert annotation of protein data. Nucleic Acids Res. 2001, 29: 29-32. 10.1093/nar/29.1.29.
3. Barton N.H., Briggs D.E.G., Eisen A.G., Goldstein D.B., Patel N.H. Hardcover. (2007). Phylogenetic reconstruction. (Chap 27). In: Evolution. Cold Spring Harbor Laboratory Press: http://www.evolution-textbook.org/content/free/contents/ch27.html#ch27-4-1
4. Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat N.T, Weissig H., Shindyalov I.N., Bourne P.H. (2000). The Protein Data Bank. Nucleic Acids Research. 28: 235-242.
5. Bioinformatics and data base http://halfonlab.ccr.buffalo.edu/courses/Nutrigenomics/slides/Bioinformatics&Databases.ppt.
6. Bromham L., Penny D. (2003). The modern molecular clock. Nature Reviews Genetics. 4: 216-224
7. Chapter 8: Phylogenetic Tree Construction: http://www.cbrg.ethz.ch/education/CompBiol/slides/Chapter8.pdf
8. Corpet F, Servant F, Gouzy J, Kahn D: ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. Nucleic Acids Res. 2000, 28: 267-269. 10.1093/nar/28.1.267.
9. Durbin R., Eddy, S. Krogh, A., Mitchison, G. (1998). Building phylogenetics trees (Chap 7) In: Biological sequence analysis, UK, Cambridge Univ. Press, pp. 160-191.
10. Galperin M. Y. (2007). The Protein Data Bank. The molecular biology database collection. Nucleic Acids Res 35, D3–D4.
11. Gibas C., Jambeck P. (2001). Biological Research on the Web. (Chap 6). In: Developing Bioinformatics Computer Skills. O'Reilly, Sebastopol, Califonia, pp 130-153.
12. Hofmann K, Bucher P, Falquet L, Bairoch A: The PROSITE database, its status in 1999. Nucleic Acids Res. 1999, 27: 215-219. 10.1093/nar/27.1.215.
13. InterPro. [http://www.ebi.ac.uk/interpro/]
14. Introduction aux bases de données : www.ai.univ-paris8.fr/~lysop/bd/seance1-Introduction.pd
15. Jukes T.H., Cantor C.R. (1969). Evolution of protein molecules. In Mammalian Protein Metabolism, edited by Munro H.H, Vol. III, New York, Academic Press. pp. 21–132.
16. Kimura, M. (1981). Estimation of evolutionary distances between homologous nucleotide sequences. Proceedings of the National Academy of Sciences, USA, 78(1), 454–458.
17. Kraulis P. (2001). Databases in bioinformatics. The different types of databases. Stockholm Bioinformatics Center, SBC : http://www.avatar.se/lectures/strbio2001/databases/types.html
18. La construction d'arbres phylogénétiques – Notes et concepts: http://mon.univ-montp2.fr/claroline/backends/download.php?url=L1REcGh5bG9nZW5pZV8yMDE2LnBkZg%3D%3D&cidReset=true&cidReq=FMOE221

19. Morgane T.C. Introduction à la phylogénie . Computafional systems biology IBENS:http://www.biologie.ens.fr/~mthomas/L3/phylogenie/Thomas-Chollier_phylogenie.pdf.
20. Mount D.W. (2004). Phylogenetic prediction. (Chap 6). Bioinformatics: Sequence and Genome Analysis. New York, USA. CSHL Press, pp. 238-279.
21. NCBI Resource Coordinators· (2016). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. Nucleic Acids Res. 44 (Database issue): D7–D19.
22. Phylogeny-based methods for analysing uncultured microbial communities: http://bioinformatics.org.au/ws14/wp-content/uploads/ws14/sites/5/2014/07/Aaron-Darling_presentation.pdf.
23. Primary and secondary databases - European Bioinformatics Institute: http://www.ebi.ac.uk/training/online/course/bioinformatics-terrified/what-database/relational-databases/primary-and-secondary-databases.
24. PRINTS-S. [http://bioinf.man.ac.uk/dbbrowser/sprint/printss_lis.html]
25. PROSITE. [http://www.expasy.ch/prosite/]
26. Protein Information Resource. [http://pir.georgetown.edu/]
27. Rizzo J., Rouchka E.C. (2007). Review of Phylogenetic Tree Construction. Bioinformatics Review. University of Louisville Bioinformatics Laboratory Technical Report Series number TR-ULBL-2007-01.
28. Sequence-Tagged Sites (STS): http://www.ncbi.nlm.nih.gov/probe/docs/techsts/
29. SMART. [http://smart.embl-heidelberg.de/]
30. SWISS-PROT and TrEMBL. [http://www.expasy.ch/sprot/sprot-top.html]
31. TIGRFAMs. [http://www.tigr.org/TIGRFAMs/]
32. Vandamme A.M. (2009). Basic concepts of molecular evolution. In: The Phylogenetic Handbook: a Practical Approach to Phylogenetic Analysis and Hypothesis Testing, edited by Lemey L, Salemi M, Vandamme A.N. Published by Cambridge University Press. C Cambridge University Press, pp 3-29.
33. Wu CH, Xiao C, Hou Z, Huang H, Barker WC: iProClass: an integrated, comprehensive and annotated protein classification database. Nucleic Acids Res. 2001, 29: 52-54. 10.1093/nar/29.1.52.
34. Xiong J. (2006). Introduction to Biological Databases (Chap 02). In: Essential Bioinformatics. Cambridge University Press, New York, USA, pp 10-27.
35. Yang Z., Rannala B. (2006) Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. Molecular Biology Evolution 23: 212–226.