Cours introductif à la génomique : Approche et intérêt

Introduction

La génomique est une branche de la biologie qui porte sur l'étude des génomes, qui constituent le support moléculaire des caractères héréditaires des êtres vivants. L'étude de ces caractères héréditaires a été initiée par Gregor Mendel sur une plante, le petit pois. Ces travaux ont montré que les caractères héréditaires sont transmis en descendance selon des lois statistiques (lois de Mendel). Ces lois, oubliées puis redécouvertes par Morgan et ses élèves étudiant la mouche drosophile se sont révélées valables aussi bien pour les plantes que les animaux. De plus, les travaux de Morgan ont montré que ces caractères héréditaires appelés gènes sont disposés comme un chapelet de perles sur les chromosomes. Les lois de Mendel permettent de mesurer les distances qui séparent ces caractères héréditaires sur les chromosomes et de proche en proche, en analysant un grand nombre de gènes, de constituer une carte génétique de l'espèce étudiée. Les chromosomes, supports de ces caractères, sont constitués d'un fil d'ADN « emballé » dans une matrice protéique, la chromatine. Les travaux des biologistes moléculaires ont démontré que l'ADN est le support moléculaire des gènes qui sont constitués d'un enchaînement de quatre molécules appelées bases (ATGC). Les techniques de visualisation des chromosomes montrent que leur structure n'est pas homogène. Certaines zones sont très compactes et accumulent les colorants (hétérochromatine) et d'autres zones à coloration régulière sont appelées euchromatiques. On observe aussi une grande hétérogénéité de taille des chromosomes eux-mêmes, selon les espèces. Cette différence reflète-t-elle une différence dans le nombre des gènes présents dans différentes espèces ? La génomique a permis d'élucider cette question et d'expliquer la base de ces hétérogénéités.

Volets de la génomique

- La **génomique structurale** consiste en l'analyse de la structure des gènes et autres parties du génome. Elle contribue à l'annotation des génomes et à l'identification des séquences informatives (les gènes avec ou sans introns codant des protéines ou des ARN fonctionnels, les séquences régulatrices, les séquences répétées, les éléments transposables, ...).

- La **génomique fonctionnelle** analyse la fonction des gènes et autres parties du génome. Elle inclue l'analyse transcriptomique (ARN messagers). Elle contribue aussi

très largement à l'annotation des génomes et à l'identification des séquences informatives.

- Afin de déterminer la fonction des ARN et des protéines associés à ces gènes, la **génomique fonctionnelle** analyse aussi :

- le transcriptome : les produits de la transcription des gènes, les ARN

- le protéome : l'ensemble des protéines synthétisées (traduites) à partir des ARN messagers

- La **protéomique** est l'analyse des protéines, permettant une meilleure compréhension des processus biologiques complexes.

Intérêts et évolution de la génomique

- La génomique a permis d'accroître nos connaissances en cancérologie moléculaire et dans le domaine de la prédisposition génétique. L'utilisation de la protéomique permet de compléter ces acquis, en particulier pour une meilleure compréhension des mécanismes physiopathologiques de la fonction des gènes et pour la recherche des marqueurs diagnostiques ou des cibles thérapeutiques.

- L 'outil de base de la génomique est l'ensemble des séquences d'acides nucléiques et de séquences polypeptidiques obtenues par différentes méthodes de séquençage. Ces séquences et d'autres types d'informations qui découlent de leur analyse sont stockées dans des bases de données. L'accès aux bases de données s'effectue via le Web et Internet. L'analyse de l'ensemble de ces données nécessite des outils bioinformatiques.

- La génomique s'intéressait surtout à l'étude de la **structure**, du **contenu** et de l'**évolution** des génomes, en s'appuyant sur les résultats du **séquençage** de séquences nucléotidiques. Actuellement ce domaine initial d'application de la génomique s'est élargi du fait de l'apparition de nouvelles technologies et d'outils bioinformatiques de plus en plus précis et puissants.

Schématiquement, la génomique fonctionnelle a pour principaux buts de

- ❖ Déterminer le moment dans le cycle cellulaire où un gène est transcrit (appelé expression d'un gène) et donc d'étudier les différences de transcription des gènes dans le temps et pour chaque type de tissus et cellules.
- ❖ D'identifier les éléments constitutifs d'un gène (introns, exons, séquences de régulation de la transcription, ...) et d'identifier les régions des génomes dont on ignore encore le rôle et élucider ce rôle.
- ❖ Les conditions liées à la transcription ou la non-transcription d'un gène; l'intensité (nombre de copies du ou des transcrits) avec laquelle ce gène est transcrit ainsi que le compartiment où est adressé le (ou les) produit(s) de la transcription d'un gène
- ❖ Les interactions que le produit d'un gène peut établir avec d'autres produits de gènes et/ou d'autres types de molécules (interactomique). Ce type d'analyse débouche sur la construction de **réseaux d'interactions** ("*interaction networks*").
- ❖ D'étudier les différences d'activité biologique des produits des gènes dans le temps et pour chaque type de tissus, de cellules, de compartiments sub-cellulaires
- ❖ D'apporter des éléments qui contribuent à déterminer la fonction des ARN et des protéines pour lesquelles les gènes codent:

- Si le produit est un ARN, le rôle qu'il peut avoir dans la régulation post-transcriptionnelle (interférence ARN).

- Si le produit intermédiaire est un ARN messager et le produit final est donc une protéine, le rôle que cette protéine peut avoir dans une voie métabolique et/ou dans la régulation de cette voie métabolique. Intégrer toutes ces informations dans un ensemble plus vaste, celui du métabolisme (métabolome)

- Les nouvelles techniques de séquençage en masse (ou à très haut débit) ont encore élargi le champ d'investigation de la génomique fonctionnelle. On peut citer :
- le séquençage *de novo* ou le reséquençage d'un génome connu
- l'étude de la variabilité génétique et du polymorphisme de nucléotide simple (SNP)
- le séquençage d'haplotypes particuliers lors du clonage positionnel d'un gène d'intérêt

- l'étude de plus en plus fine du transcriptome : telle que l'étude des phénomènes d'épissage alternatif, identification de transcrits rares, identification des séquences frontières intron/exon, analyse quantitative du niveau de transcription des gènes
- étude du profil en petits ARN non codants.
- l'étude des interactions ADN/protéines (régulation de la transcription, facteurs de transcription, ...)
- la génomique médicale
- la génomique comparative qui compare la structure et les fonctions des génomes de différentes espèces (organisation et évolution des génomes).
- la métagénomique : étude du génome d'un organisme prélevé directement dans un environnement complexe (intestin, océan, sols, ...). Le but est d'obtenir des informations sur l'incidence de cet environnement.
- l'épigénétique et l'épigénomique : étude de l'influence de l'environnement et de l'histoire individuelle sur les modifications de l'expression des gènes d'une génération à l'autre.
- l'étude du profil de méthylation (processus épigénétique)

- **Domaines d'application de la génomique**
- **Apport du séquençage en génomique**

L'avènement des technologies de séquençage à très haut débit a bouleversé la portée des résultats que l'on peut obtenir en génomique.

La plus grande partie d'un génome eucaryote étant constitué d'ADN non codant, le séquençage porte aussi sur des clones d'ADNc qui contiennent des séquences issues de la transcription inverse d'ARN messagers.

Le séquençage requiert des logiciels bioinformatiques adaptés à la quantité phénoménale d'information qu'il génère.

- **Génomique et outils bioinformatiques**

- L'étude des génomes, des transcriptomes et des protéomes, des interactomes, nécessite le développement de technologies informatiques, de logiciels ou d'ensemble de logiciels et de théories informatiques, afin :
- D'automatiser l'obtention des données
- De permettre l'analyse de ces données
- stocker et organiser ces données dans des bases de données consultables via Internet et des interfaces Web
- Développer des modèles mathématiques - logiques pour l'analyse des données
- Développer des langages informatiques spécifiques au traitement des données
- Des logiciels bioinformatiques sont nécessaires pour l'étude de la structure des gènes, Par exemple :
- BLAST qui permet d'aligner la séquence du génome avec les séquences d'ADNc ou rechercher des similarités entre ce génome et d'autres génomes déjà connus et annotés.
- "*ORF Finder*" (NCBI)

- **Génomique et expression génique**
- Les méthodes actuelles de la génomique sont basées sur la détection de marqueurs spécifique de chaque gène dans une banque contenant des centaines de milliers de fragments séquencés. on peut alors utiliser des puces à ADN sur lesquelles sont hybridés des ADNc marqués par des molécules fluorescentes.
- Ces ADNc sont extraits à partir de tissus, de cellules auxquels on a appliqué un traitement particulier. La comparaison de l'expression des gènes par rapport à un témoin non traité permet théoriquement d'évaluer le niveau relatif de transcription de l'ensemble des gènes d'un génome dans des centaines de conditions.
- **La collecte de données fonctionnelles sur les aspects biochimiques et sur l'action phénotypique des gènes**
- La génomique fonctionnelle inclut des approches qui permettent de vérifier les propriétés biochimiques et les rôles cellulaires du produit de chaque gène.

- La génétique inverse (qui va du gène vers le phénotype) consiste à inactiver de manière ciblée et systématique des gènes particuliers. Parmi les stratégies employées, on peut citer

- La mutagénèse systématique qui correspond à l'extinction ("*knock-out*") d'une série de gènes un par un; ou encore on peut induire la perte transitoire de la fonction d'un gène en se servant d'ARN interférent ("*RNAi*")…etc

- **Evaluation de la variabilité de la séquence d'ADN au sein d'une même espèce**

Les génomes sont polymorphes. Le polymorphisme d'un seul nucléotide ou SNP ("*Single nucléotide polymorphism*") ou le polymorphisme d'insertion ou de délétion de nucléotides ont une part essentielle dans la variabilité génétique.

Identification et Annotation de l'ensemble des gènes des génomes et de leurs produits d'expression

- L'annotation d'un génome, d'un transcriptome, d'un protéome, d'un métabolome ... consiste à documenter de la manière la plus exhaustive tous les composants de cette information brute. En effet, une fois un gène identifié, il faut l'annoter: le relier aux maximum de données biologiques (par exemple : données de génétique concernant sa fonction, son expression et les variations phénotypiques…etc).

En d'autres termes, on tente d'assigner aux molécules pour lesquelles les gènes codent :

- une fonction biologique / biochimique ; une localisation sub-cellulaire ; leur implication dans des processus de régulation ; leur interactions avec d'autres molécules biologiques ; un profil de transcription

- L'annotation peut être:

a. automatique : s'appuie essentiellement sur des comparaisons des séquences à annoter avec les séquences présentes dans les banques de données. Les algorithmes recherchent des similarités / homologies de séquence, de structure, de motifs, permettant de prédire la fonction d'une molécule et de transférer automatiquement l'annotation entre les molécules homologues.

b. manuelle (ou curation) par des experts (des curateurs) qui valident ou invalident la prédiction en fonction de leurs connaissances ou de résultats expérimentaux. L'annotation

manuelle est donc tout à fait indispensable. Mais, en regard de la quantité phénoménale de données acquises quotidiennement, il est illusoire d'envisager une curation manuelle de l'ensemble des données en temps réel.

c. Structurale: tente de prédire : Le contenu en gènes et leur localisation dans le génome ainsi que l'organisation des gènes.

d. Fonctionnelle tente de prédire la fonction potentielle des gènes.

e. Relationnelle tente de décrire les relations (interactions) entre les produits des gènes (familles de gènes, réseaux de régulation, réseaux métaboliques, ...).

La protéomique

- La protéomique a pour but d'identifier et quantifier l'ensemble des protéines synthétisées (protéome), à un moment donné et dans des conditions données au sein d'un tissu, d'une cellule ou d'un compartiment cellulaire.
- Le protéome est extrêmement complexe, compte-tenu de l'épissage alternatif des transcrits primaires (plusieurs ARNm pour un gène) et compte-tenu des modifications post-traductionnelles des protéines.
- Pour chaque condition environnementale (condition physiologique normale *vs.* conditions de stress) une cellule est caractérisée par un protéome adapté à cette condition alors qu'elle a toujours le même génome.
- La protéomique apporte des réponses quant aux :
- modalités d'expression des gènes pour les organismes dont le génome n'a pas encore été séquencé ou pour lesquels les programmes de prédiction de séquences codantes sont moins fiables.
- estimation quantitative des concentrations des protéines synthétisées
- obtention de données sur la fonction des protéines et les interactions entre protéines ou entre protéines et autres molécules biologiques.

Les approches génomiques ont connu un grand essor avec le développement de nouveaux outils d'analyse des génomes qui viennent compléter chaque année les approches (par exemple actuellement les outils d'analyse du protéome, les techniques de séquençage à très

haut débit). Cet éventail de recherches, s'appuyant aussi sur les techniques d'analyse de la biodiversité des séquences nous apporte des informations sur l'évolution des génomes et sur les processus de domestication. La génomique apporte des outils nouveaux pour étudier la biodiversité des espèces par exemple, et pour constituer des ressources génétiques organisées et exploitables, une étape essentielle au travail d'amélioration génétique.

Structure et fonctionnement des génomes procaryotes et eucaryotes

Les gènes sont organisés de façon assez stéréotypée avec une séquence promotrice non codante en amont immédiat du début de la séquence codante et qui représente le point d'ancrage initial de l'ARN polymérase, puis les séquences codantes proprement dites ou exons, souvent séparées par des séquences non codantes ou introns qui peuvent être très longues.

Les ARN messagers ne sont bien sûr produits qu'en regard des gènes et incluent initialement une copie des introns qui doit ensuite être éliminée (« épissage ») pour ne plus conserver qu'une copie des exons dans l'ARNm terminal, mature, prêt à être traduit en protéine.

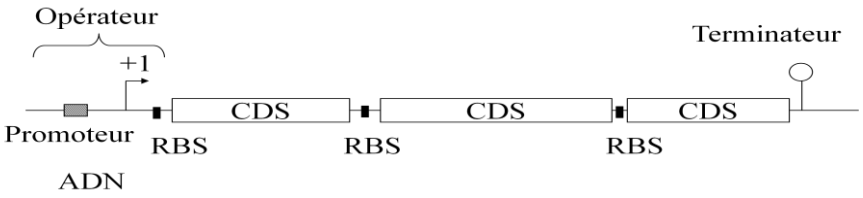
En dehors du gène, la molécule d'ADN est constituée de séquences non codantes dont la fonction est variable : séquences régulatrices, cibles de protéines qui se fixent sur l'ADN et qui régulent la transcription, séquences hautement répétées telles les microsatellites, ou séquences muettes de fonction discutée.

I- Structure et organisation du génome procaryotes

1- Organisation du génome procaryotes

Deux types de génomes coexistent dans la cellule procaryote. Le nucléoïde et les plasmides.

Le nucléoïde : le matériel génétique des procaryotes n'est pas protégé dans un noyau comme chez les eucaryotes. Ceci implique un couplage entre la transcription et la traduction et élimine une étape potentielle de régulation de l'expression de l'information génétique. Associées à l'ADN, des protéines forment une structure qui ressemble à la chromatine (en particulier chez certaines espèces d'Archaebactéries, il existe des protéines similaires aux histones qui forment des nucléosomes). Néanmoins, la structure des chromosomes bactériens n'est en rien, par sa complexité, comparable aux chromosomes des eucaryotes. En particulier la division cellulaire possède des modalités simples (fusion binaire). La taille de leurs génomes varie de 600kb à 10Mb, et ils ne comportent pas d'introns ou de séquences répétées. Les gènes sont regroupés en opéron (unité d'ADN fonctionnelle regroupant des gènes sous contrôle d'un signal moléculaire régulateur), ce qui facilite la génération de messagers polycistroniques.



Structure d'un opéron simple

Les plasmides

Un plasmide est une molécule d'ADN surnuméraire distincte de l'ADN chromosomique, capable de réplication autonome et non essentielle à la survie de la cellule. Les gènes de plasmides sont donc non-nécessaires mais avantageux pour la bactérie. Ils confèrent à cette dernière, des caractéristiques sélectives telles que des résistances aux antibiotiques, aux métaux lourds ou des facteurs de virulence. Les plasmides sont généralement circulaires (E coli). Ils se trouvent quasi-exclusivement dans les bactéries. Certains plasmides sont capables de s'intégrer aux chromosomes ; on les appelle des épisomes.

2- Transcription de l'ADN procaryote

L'ARN polymérase est une protéine ADN dépendante, multimérique possédant les sous-unités α, β, β' et σ. Elle est présente sous deux formes l'enzyme-cœur (α2ββ') et l'holoenzyme (α2ββ'σ). Les ARN polymérases ne nécessitent pas d'amorce et ne possèdent pas d'activité exo-nucléasique.

Les nucléotides triphosphates sont additionnés à l'extrémité 3' de la chaîne en cours de synthèse par complémentarité de la matrice d'ADN. L'hydrolyse de la liaison anhydride fournit l'énergie pour la synthèse de la liaison phosphodiester. La réaction est réalisée en milieu tamponné à pH neutre, contenant du sel de Mg^{2+}, un agent de protection des groupements SH de l'enzyme (agent réducteur comme le β-mercaptoéthanol), les 4 ribonucléotides (rNTP) et de l'ADN bicaténaire contenant un promoteur comme matrice.

La molécule d'ADN est composée d'un brin matriciel (3' vers 5') et servant comme son nom l'indique de matrice à l'ARN-polymérase, ainsi que d'un brin codant (5' vers 3') et ayant une séquence identique à l'ARN transcrit mise à part le fait que la thymine est changée par l'uridine. L'ARN messager est lui monocaténaire et dirigé tout comme le brin codant de 5' vers 3'.

ADN	5' TTACCTG 3' (brin codant) 3' AATGGAC 5' (brin matriciel)	➡	ARN 5' UUACCUG 3'

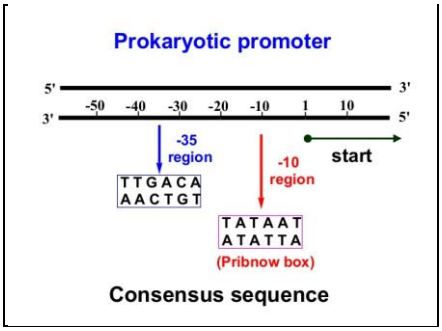
Chez E-coli, une seule ARN-polymérase catalyse la synthèse de tous les ARN de la cellule.

La transcription est divisée en plusieurs étapes : la pré-initiation, l'initiation, l'élongation et la terminaison.

- **A- <u>Pré-initiation</u>**

  Le promoteur, situé en amont du site d'initiation et porte des éléments de séquence reconnus par l'ARN-polymérase et déterminant le sens de la transcription. Il est constitué de courtes séquences conservées d'une unité de transcription à l'autre et appelées séquences consensus :

- En -10 du site d'initiation on trouve la TATA box (boîte de Pribnow) : « TATAAT »
- En -35 du site d'initiation on trouve : « TTGACA »



Promoteur procaryotique

Le promoteur agit sur la transcription du segment d'ADN qui lui est adjacent sur le même chromosome, on dit que le promoteur est actif en « cis ». Le promoteur n'est pas actif sur les séquences codantes situées ailleurs sur le chromosome, dans ce cas, il est actif en trans.

L'affinité de l'ARN-polymérase pour l'ADN dépend de la forme de l'enzyme : l'enzyme-cœur a une affinité faible et non spécifique, l'holoenzyme a une affinité très forte et spécifique pour le promoteur. La sous-unité sigma σ à l'état libre ne se fixe pas sur l'ADN. La sous-unité β' étant basique et l'ADN étant acide, ce sera elle qui facilitera l'interaction du complexe avec le promoteur. La sous-unité sigma σ permet donc une reconnaissance spécifique du promoteur par l'ARN-polymérase et diminue l'affinité de l'enzyme pour les régions non promotrices. Il agit de manière cyclique, après l'initiation faite, le facteur σ se détache pour être recyclé et réutilisé pour d'autres initiations de gènes.

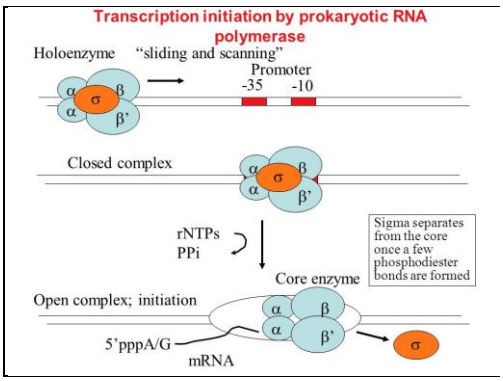
L'ARN-polymérase entraîne la dénaturation des deux brins d'ADN sur 14 paires de nucléotides, on parle de complexe ouvert qui augmente encore l'affinité de l'enzyme pour la double hélice.

B- Initiation

L'initiation correspond à la synthèse de la première liaison phosphodiester réalisée par la sous-unité β qui correspond à la sous-unité catalytique de l'ARN-polymérase.

Le déroulement des premières étapes de la transcription est donc :

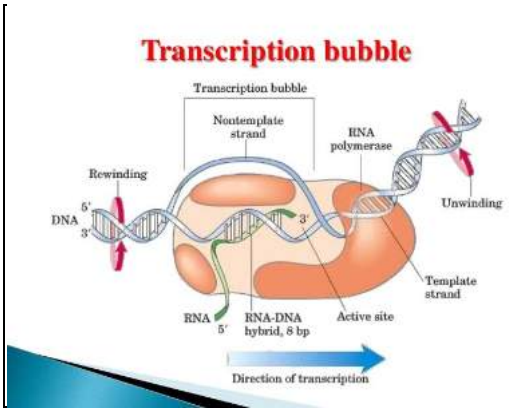
- liaison non spécifique de l'holoenzyme; - formation d'un complexe fermé au niveau du promoteur; - formation du complexe ouvert (déroulement sur 14 nucléotides); - Mise en place du premier nucléotide (très souvent A ou G);- Allongement de 4 à 5 nucléotides; - Détachement du facteur σ, après la transcription des 4-5 premiers nucléotides.



Initiation de la transcription chez les procaryotes

C- Elongation

Correspond au déplacement de la bulle de transcription le long de la molécule d'ADN. Pendant la transcription, l'ARN forme un court appariement avec le brin matriciel de l'ADN formant une hélice hybride ADN-ARN sur une dizaine de paires de bases.



Elongation chez les procaryotes

- **d) Terminaison**

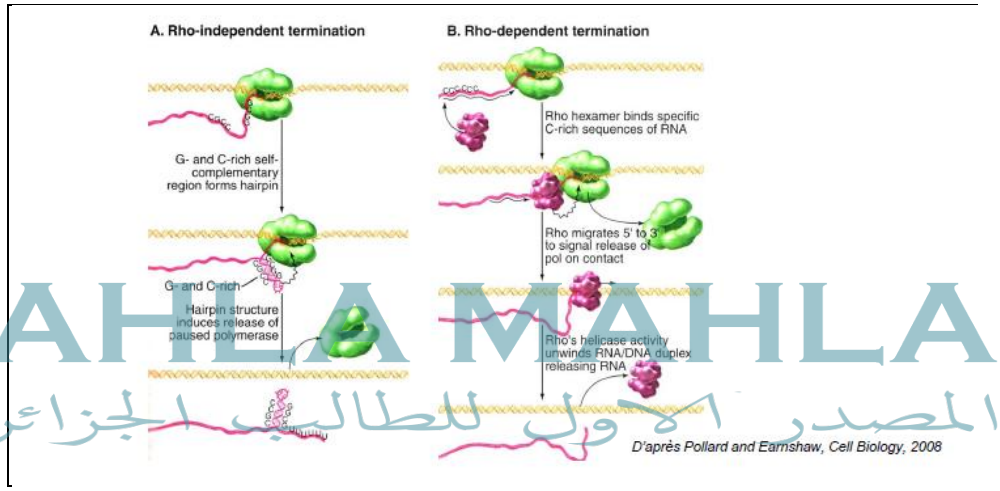
Est réalisée lorsque l'enzyme arrive au niveau de la séquence terminateur, se présentant sous la forme d'un palindrome qui peut être parfait ou imparfait. Ce palindrome entraîne une

complémentarité de séquence au niveau de l'ARNm qui permet la mise en place d'une structure en épingle à cheveux qui est un appariement intra-chaîne qui déstabilise l'ARN-polymérase jusqu'à dissociation.

Elle peut être facilitée par un facteur rho ρ suivant la séquence du terminateur, on met ainsi en évidence des terminateurs rho indépendant (environ les 2/3) et des terminateurs rho dépendant (environ 1/3) :

Pour les terminateurs rho indépendant on trouve une structure en épingle à cheveux riche en paires de bases G-C, suivie d'une séquence poly-U d'environ 6 nucléotides permettant une dissociation plus facile de l'hybride ADN-ARN.

Pour les terminateurs rho dépendant on trouve une structure en épingle à cheveux plus courte et qui n'est pas riche en paires de bases G-C et qui est non-suivie d'une séquence poly-U. Il y a donc nécessité du facteur rho qui a une affinité pour les ARN en court de synthèse, le parcourant de 5' vers 3' jusqu'à trouver l'ARN-polymérase. Le facteur rho est ATP dépendante, dont l'hydrolyse permettra la dissociation du complexe.



Terminaison chez les procaryotes

e) <u>Maturation des transcrits primaires</u>

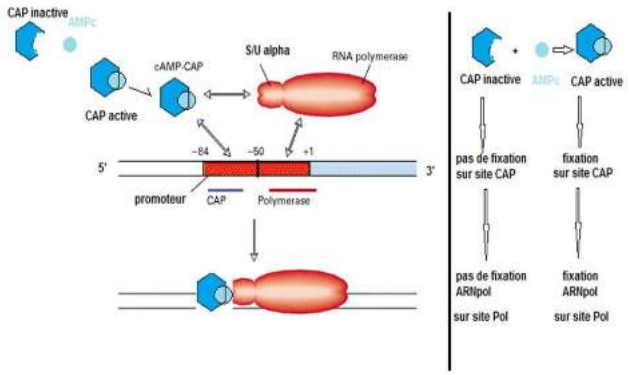
Le transcrit primaire correspond à l'ARN non mature qui nécessite une maturation sous forme de clivages ou de modifications de bases. Cette maturation n'est pas obligatoire. Le transcrit primaire code soit pour un produit (ARN monocistronique), soit plusieurs produits, (ARN polycistronique).

II- Structure et organisation du génome eucaryote

Le génome eucaryote est composé de la succession de séquences codantes et non codantes. Le premier et dernier exon renferment une séquence non traduite mais transcrite dans l'ARN : les séquences UTR (untranslated region), portant des séquences signal. L'UTR du premier exon

renferme la séquence signal de la «CAP» protéine d'activation du catabolisme. l'UTR du dernier exon referme le signal de «poly-adenylation».

Relation entre protéine CAP et ARN pol



Deux types de génome au moins coexistent dans les cellules eucaryotes.

a- Le génome nucléaire

Le matériel génétique des eucaryotes est entouré par une membrane, le tout formant le noyau. A l'intérieur de ce dernier, il adopte une structure chromatinienne et supra-chromatinienne très complexe. Il présente un découplage entre transcription et traduction, qui se passent respectivement dans le noyau et dans le cytoplasme. Ce découplage offre la possibilité de réguler plus finement l'expression des gènes. Le noyau peut subir des modifications de contenue ou de structure. Des pertes ou des amplifications d'ADN sont observées dans certaines cellules d'organismes pluricellulaires, qui ne sont plus destinées à servir pour la reproduction (cellules somatiques), alors que les cellules germinales conservent la totalité du matériel génétique. Il est possible que la séquestration du génome dans le noyau ait permis d'accroître fortement sa taille et donc de complexifier encore plus la cellule eucaryote. Le génome eucaryote est souvent non compact avec la présence de nombreux introns. Les gènes ne sont pas groupés en opérons et les messagers polycistroniques sont très rares.

b- Le génome mitochondriale

La théorie endosymbiotique a permit de démontrer que le génome mitochondrial dérive du génome bactérien. L'ADN mitochondrial est localisé à proximité de la membrane interne mitochondriale. Se concentre au niveau de structures appelées nucléoïdes. Ces structures apparaissent avant la phase S de la méiose et possèdent des points d'ancrage au niveau de la membrane interne mitochondriale. L'ADN mitochondrial existe en plusieurs exemplaires dans chaque cellule et ce nombre de copies peut varier d'un organisme à un autre. En effet, l'homme possèderait entre 200 et 1700 copies d'ADN mitochondrial par cellule tandis que la levure S.cerevisiae quant à elle, en possèderait entre 50 et 100 copies. Chez l'homme, ce nombre peut varier en cas de maladies telles que des maladies du foie ou des cancers du sein. En ce qui concerne la taille du génome mitochondrial, elle varie selon les espèces.

Chez les eucaryotes les unités de transcription sont monocistronique (exception chez certains vertébrés). Les gènes eucaryotes sont départagés en 3 classes :

Les gènes de classe 1 ont comme produits des ARNr. Ils sont répétés en tandem et séparés par des espaces inter-géniques. La transcription des ARNr se fait sous forme d'un précurseur l'ARN 45S qui sera clivé en 3 ARNr.

Les gènes de classe 2 ont comme produits des protéines. Les gènes de classe 2 sont constitués d'exons et d' introns.

Les gènes de classe 3 ont comme produits des ARNt, les ARNr 5S et les petits ARN qui sont également répétés en tandem comme les gènes de classe 1.

1- Les éléments de régulation

Bien que toutes les cellules d'un organisme possèdent la même information génétique, le taux d'expression d'un gène donné n'est pas forcément similaire dans chaque tissu différencié. Ainsi, les gènes possèdent dans leur promoteur des séquences régulatrices qui leur sont propres et qui permettent le contrôle de leur expression par des protéines régulatrices. Ces séquences permettent d'adapter le niveau d'activité d'un gène aux besoins physiologiques et à l'état de différenciation de la cellule. Il existe deux types d'éléments de régulation, classés selon leur éloignement du site d'initiation de la transcription : les séquences proximales et les séquences distales.

a) Les séquences régulatrices proximales

Généralement situées entre 40 et 110 pb en amont du site d'initiation et parfois présentes en plusieurs exemplaires, les séquences proximales peuvent avoir un effet activateur (UAS, pour upstream activating sequence) ou répresseur (URS, pour upstream repressing sequence) en fonction de l'activité de la protéine qui les reconnaît. Les deux séquences les mieux étudiées sont les boîtes CAAT et les motifs riches en GC.

b) Les séquences régulatrices distales

Les éléments distaux du promoteur peuvent être situés jusqu'à quelques milliers de paires de bases en amont ou en aval du site d'initiation. Ces éléments peuvent activer la transcription, 'enhancers', ou la réprimer, 'silencers'. Le mécanisme d'action par lequel des séquences éloignées régulent la transcription est encore assez mal connu. Une hypothèse séduisante propose qu'un repliement de l'ADN permettrait le rapprochement de ces séquences régulatrices du site d'initiation.

2- Les ARN-polymérases eucaryotes

Trois ARN-polymérases eucaryotes ont été mis en évidence; qui diffèrent par leur localisation dans le noyau, par la nature des ARN formés et de par leur sensibilité à des inhibiteurs tels que l'α-amanitine.

ARN-polymérase I dans le nucléole pour les ARNr 5,8 ; 18 et 28 S, et est insensible à l'α-amanitine

ARN-polymérase II dans le nucléoplasme pour les ARNm et est sensible à l'α-amanitine

ARN-polymérase III dans le nucléoplasme pour les ARNt, ARNr 5 S et pour les petits ARN, elle est également sensible à l'α-amanitine mais à hautes doses.

L'α-amanitine se fixe sur certaine sous-unité de l'ARN-polymérase et inhibe l'élongation de la transcription.

L'actinomycine D inhibe la transcription eucaryote et procaryote en s'intercalant entre certaine base de l'ADN pendant l'élongation.

3- Les séquences régulatrices

Le promoteur : il détermine le début et l'orientation de la transcription, c'est le site de fixation de l'ARN polymérases.

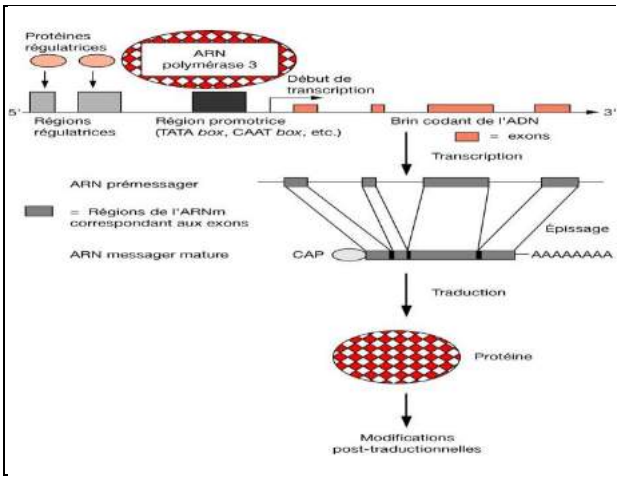
Le Silencer: inhibiteur, situé entre le promoteur et le gène de structure, il permet de ralentir ou arrêter la transcription.

L'Enhancer : c'est un activateur de la transcription (trans-activateur), il agit à distance et peut se trouver en amont ou en aval du gène de structure.

Des régions situées en amont du site d'initiation sont importantes pour la transcription et sont:

–La boite TATA: elle est située à environ – 30 paires de bases de l'origine de la transcription; dite séquence consensus. Cette boite fixe un facteur de transcription qui est absolument nécessaire pour l'initiation. La boîte TATA, est localisée à environ une trentaine de paires de bases (jusqu'à 120 pb chez la levure) en amont du site d'initiation. Cette séquence est reconnue par la machinerie transcriptionnelle via la sous-unité TBP (TATA Binding Protein) du facteur de transcription TFIID, et pourrait également jouer un rôle dans la détermination du sens de transcription. L'élément initiateur (Inr) est situé autour du site d'initiation et est constitué d'une séquence riche en pyrimidines, moins conservée que la TATA. Il permet la transcription des gènes qui ne contiennent pas de boîte TATA.

–La boite CCAAT : souvent située à environ – 70 paires de bases du site d'initiation



Structure du génome eucaryote

4- Transcription de l'ADN eucaryote

- l'initiation de la transcription

L'ARN Pol II n'est pas capable de démarrer seule la synthèse d'ARN au niveau d'un promoteur. L'initiation de la transcription *in vitro* nécessite la présence de facteurs auxiliaires, appelés facteurs généraux de transcription: TFIIA, TFIIB, TFIID, TFIIE, TFIIF et TFIIH. Ces facteurs et L'ARN constituent la machinerie transcriptionnelle de base, qui est la cible d'activateurs ou de répresseurs qui modulent le taux d'expression de chaque gène en réponse à divers signaux.

- L'assemblage du complexe de pré-initiation

L'étape initiale de la formation du complexe de pré-initiation (PIC : pre-initiation complex) est la reconnaissance de la boîte TATA par TFIID, et plus particulièrement par la TBP. TBP existe dans la cellule sous forme de dimères et sa fixation sur le promoteur nécessite une dissociation de ces dimères.

En se fixant sur le promoteur, TFIID crée une courbure de l'ADN d'environ 90° qui permet d'une part le rapprochement des séquences situées en amont avec celles situées en aval de la TATA. Deux voies sont alors possibles : la voie de l'assemblage séquentiel, où les facteurs se fixent dans un ordre établi, et la voie de l'holoenzyme où tous les facteurs (à l'exception de TFIID) sont pré-assemblés au sein d'un gros complexe protéique et viennent se fixer ensemble au promoteur.

a) Assemblage séquentiel

Une fois TFIID associé au promoteur, il recrute les autres facteurs de transcription. Ainsi, TFIIA, composé de trois sous-unités de poids moléculaires de 13, 19 et 37 KDa, vient stabiliser le complexe TFIID/TATA en interagissant directement avec l'ADN en amont de la boite TATA. Le rôle de ce facteur est multiple puisqu'il stimule également la liaison de TFIID

à l'ADN sur des promoteurs contenant un Inr en plus de la boîte TATA. En effet, TFIIA est responsable de la séparation des dimères de TBP, accélérant ainsi la cinétique d'association de celui-ci sur l'ADN.

La structure formée par l'association ADN/TFIID/TFIIA est reconnue par une protéine de 35 KDa : TFIIB. TFIIB ne peut se fixer au niveau du promoteur que si TBP est présent. Il interagit aussi bien avec le domaine carboxy-terminal de TBP qu'avec des séquences d'ADN situées en aval et en amont de la boîte TATA.

L'ARN Pol II s'associe alors au complexe conjointement à TFIIF. En effet, ces deux complexes s'associent l'un à l'autre pour rejoindre le PIC en formation. TFIIF; serait responsable du positionnement correct de la polymérase en réduisant les interactions non-spécifiques de celle-ci avec l'ADN. TFIIF favorise l'enroulement de l'ADN, sur près d'un tour, autour de l'ARN Pol II. TFIIF interagit avec l'ADN, entre la boîte TATA et le site d'initiation. Il interagit également avec de nombreux partenaires dont TFIIB, TFIIE, TBP. Cette dernière interaction semble indispensable à l'initiation de la transcription. Outre son rôle dans l'initiation de la transcription, TFIIF est impliqué dans la transition initiation-élongation en inhibant un arrêt prématuré de la polymérase ainsi que dans l'élongation de la transcription où il stabiliserait la polymérase.

La liaison directe de TFIIE avec la polymérase, la grande sous-unité de TFIIF et TBP/TFIID, constitue l'étape suivante de la formation du PIC. Cet hétérotétramère, recrute également le complexe TFIIH. La petite sous-unité de TFIIE, présente légèrement en amont du site d'initiation de la transcription, possède un motif en doigt de zinc indispensable à l'association de TFIIE avec l'ADN. Elle est supposée jouer un rôle dans l'ouverture de la double hélice d'ADN, en partenariat avec TFIIH.

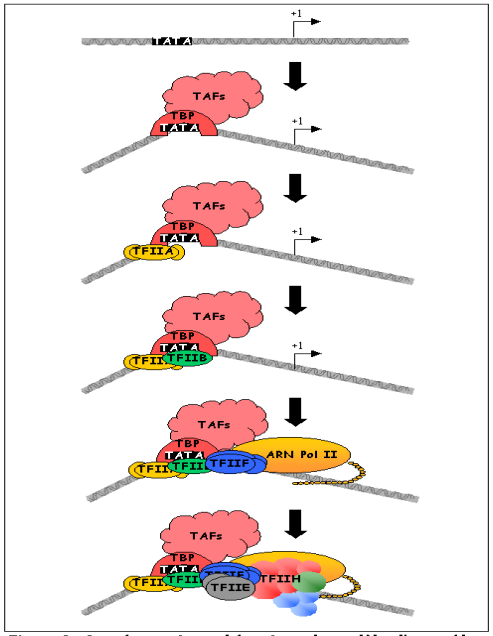


Figure 2 : **Représentation schématique du modèle d'assemblage séquentiel du complexe de pré-initiation sur le promoteur.**

b) L'holoenzyme

Une fois TFIID mis en place, un complexe macromoléculaire appelé holoenzyme contenant non seulement la machinerie transcriptionnelle (ARN Pol II et facteurs de transcription) mais aussi des protéines régulatrices de la transcription (co-activateurs, complexe médiateur), peut se positionner autour du promoteur.

La composition de l'holoenzyme varie selon l'organisme étudié. Par ailleurs, si l'existence d'un complexe contenant toutes les activités enzymatiques nécessaires à la transcription et permettant le recrutement de celles-ci en une seule étape facilite vraisemblablement la rapidité de la réponse de la machinerie transcriptionnelle.

Tous les facteurs d'initiation de la transcription étant correctement assemblés au niveau du promoteur, le complexe de pré-initiation est alors prêt à entrer dans la phase d'initiation de la transcription proprement dite. Il est maintenant établi que l'initiation se déroule en trois étapes distinctes : l'ouverture de l'ADN au niveau du site d'initiation (appelée 'bulle' de transcription), l'allongement de la bulle et le relargage des facteurs d'initiation, simultanément au recrutement des facteurs d'élongation.

Au moment de l'association des facteurs d'initiation à l'ADN, celui-ci se trouve dans une conformation double-brin. Après l'assemblage des différents facteurs, l'ADN est ouvert entre les positions -9 et +2. Cette ouverture fait intervenir les activités hélicases de TFIIH.

L'initiation de la synthèse de la chaîne d'ARN, par la formation de la première liaison phosphodiester, aboutit à la formation d'un dinucléotide. Un complexe de transcription stable n'est obtenu qu'après la formation du produit de 4 nucléotides.

Il s'en suit alors un allongement de la bulle, au fur et à mesure de l'allongement de l'ARN, jusqu'à la position +9. La troisième transition prend place lorsque la bulle se déplace pour atteindre le nucléotide situé en position +11. Ce passage de l'initiation à élongation est appelé 'échappée du promoteur'. La phosphorylation du CTD par TFIIH semble être l'évènement clef de cette étape de la transcription. Cette phosphorylation pourrait déstabiliser les interactions entre l'ARN Pol II et les facteurs d'initiation.

La réaction d'initiation s'achève par le départ des facteurs d'initiation de la bulle de transcription. TFIIB, se détacherait le premier, suivit de TFIIE, de TFIIF et TFIIH. Certains facteurs de transcription (en particulier TFIID, TFIIH et TFIIE) peuvent cependant être maintenus sur le promoteur par un activateur transcriptionnel.

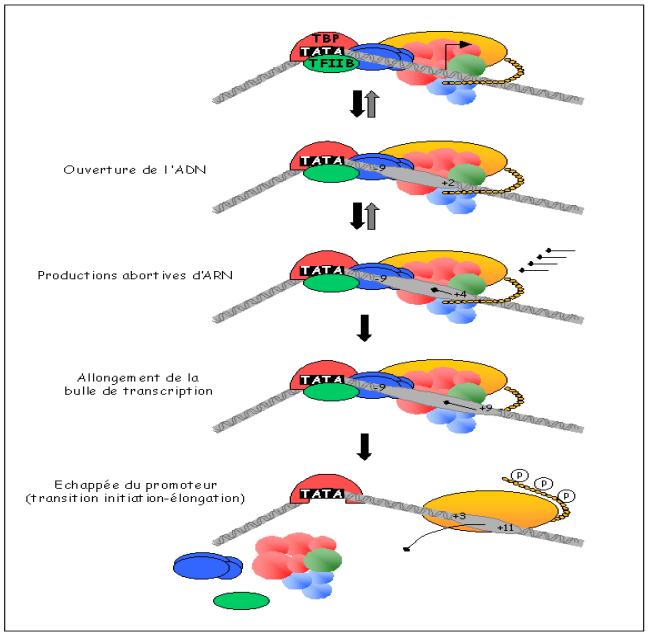


Figure 4 : **Les différentes étapes de l'initiation de la transcription.**

- Elongation de la transcription

Débute quand les facteurs d'élongation prennent place aux côtés de la polymérase. On peut classer les facteurs d'élongation en plusieurs catégories : 1) TFIIF, l'élongine et ELL qui augmentent le taux global d'élongation de la chaîne d'ARN; 2) TFIIS permet à l'ARN Pol II de repartir lorsqu'elle est arrêtée; 3) P-TEFb stimule l'élongation par phosphorylation du CTD; 4) DSIF, NELF et le Facteur 2 favorisent l'arrêt de la polymérase; et enfin, 5) FACT régule le taux d'élongation à travers les nucléosomes.

TFIIF s'associe directement à l'ARN Pol II. Cette association induit un changement de conformation de l'enzyme, empêchant l'arrêt de la polymérase au niveau de sites de pauses transitoires. L'élongine (ou facteur SIII), constituée de trois protéines de 110 (sous-unité A), 18 (B) et 15 KDa (C), augmente la vitesse de polymérisation de l'ARN. ELL est un polypeptide de 80 KDa qui augmente également la vitesse de synthèse de l'ARN.

La protéine de 38 KDa TFIIS (ou SII) interagit avec la grande sous-unité de l'ARN Pol II et permet à celle-ci de reprendre l'élongation, par un mécanisme impliquant le clivage de l'ARN synthétisé puis une resynthèse de l'ARN.

L'hétérodimère P-TEFb (Positive Transcription Elongation Factor b), formé de la kinase cdk9 et de la cycline T, permet l'allongement des transcrits grâce à la phosphorylation du domaine CTD de l'ARN Pol II. En l'absence de P-TEFb, deux facteurs négatifs d'élongation, DSIF (DRB Sensitivity Inducing Factor) et NELF (Negative ELongation Factor), s'associent à l'ARN Pol II et s'opposent à sa progression au delà de quelques dizaines de nucléotides après le site d'initiation. La phosphorylation du CTD par P-TEFb favorise la dissociation de DSIF et NELF de l'ARN Pol II.

Enfin, le Facteur 2 inhibe l'élongation mais n'agit pas sur la phosphorylation de la polymérase. Par contre, ce facteur pourrait provoquer une terminaison prématurée de la transcription. Le facteur 2 contient un domaine ATPasique et possède la capacité à lier l'ADN simple ou double brin. Ainsi, cette protéine est capable de provoquer le relâchement du transcrit par la polymérase d'une manière ATP dépendante. Ce facteur pourrait déstabiliser le complexe ADN-ARN Pol II.

-Modifications post transcriptionnelles des ARNm

La maturation des transcrits primaires à lieu dans le noyau de la cellule. Chez les procaryotes ce phénomène n'existe pas, le début de la traduction de l'ARNm se faisant avant la fin de la transcription.

Les transcrits primaires ne correspondent qu'aux produits de la transcription de l'ARN-polymérase II, ce qui ne veux pas pour autant dire que les produits de la transcription des autres ARN-polymérases ne sont pas soumis à des modifications post-transcriptionnelles. On parle de pré-ARNm, pré-ARNr et pré-ARNt.

a) Addition de la coiffe en 5' (ou capping)

La coiffe correspond à l'ajout d'un groupement, dit « m7G », par une liaison 5'-5' tri-phosphate. Ce groupement m7G correspond à l'addition de trois groupements phosphates et d'une molécule de GTP au niveau de l'extrémité 5' du transcrit primaire grâce à l'énergie libérée par l'hydrolyse de la molécule de GTP. Le nucléotide G va ensuite être méthylé sur le septième carbone (C7) pour donner la 7-méthyl-guanosine.

La coiffe est ajoutée grâce à un complexe protéique appelé « Cap-Binding-Complex » qui possédant une activité triphosphatase, une activité guanylyl-transférase et une activité méthyl-transférase.

b) Poly-adénylation en 3' par la poly-A polymérase

Correspond à l'ajout de jusqu'à 200 adénines à l'extrémité 3' du transcrit primaire et ceci sans matrice par la poly-A-polymérase. La poly-A-polymérase reconnaît le signal de poly-adénilation qui n'est autre que la séquence CPSF (AAUAAA).

c) Excision des introns et épissage des exons (ou splicing)

Après l'addition de la coiffe et la poly-adénilation, le transcrit primaire est encore soumis à l'excision des introns et l'épissage des exons ; les introns sont ainsi éliminés. Ceci est possible par la présence de site donneur d'épissage (dinucléotide GU) à l'extrémité 5' des introns et de site accepteur d'épissage (dinucléotide CAG) à l'extrémité 3' des introns.

Les jonctions d'épissage sont reconnues par les snRNPs (ou snurps pour Small-Nuclear-Ribonucleo-protein-Particules). Les snRNP correspondent à l'association de snRNA et de protéines et l'ensemble des snRNPs s'appelle le spliceosome.

L'excision des introns et l'épissage des exons se fait en plusieurs étapes :

- Le snRNP U1 permet la reconnaissance du site donneur d'épissage et entraîne la rupture de la liaison phosphodiester entre le premier exon et l'intron.

- Cette rupture de la liaison phosphodiester entraîne la formation d'un lasso, qui n'est autre que l'extrémité 5' de l'intron. Ce lasso forme une liaison avec le site de branchement, lui-même situé sur le même intron qui se repli ainsi sur lui-même. Le site de branchement est reconnu par le snRNP U2 et permet la liaison par l'intermédiaire d'une adénosine.

- Le snRNP U2 permet également la reconnaissance du site accepteur d'épissage. Suite à cette reconnaissance il y a rupture de la liaison phosphodiester au niveau de l'extrémité 3' de l'intron.

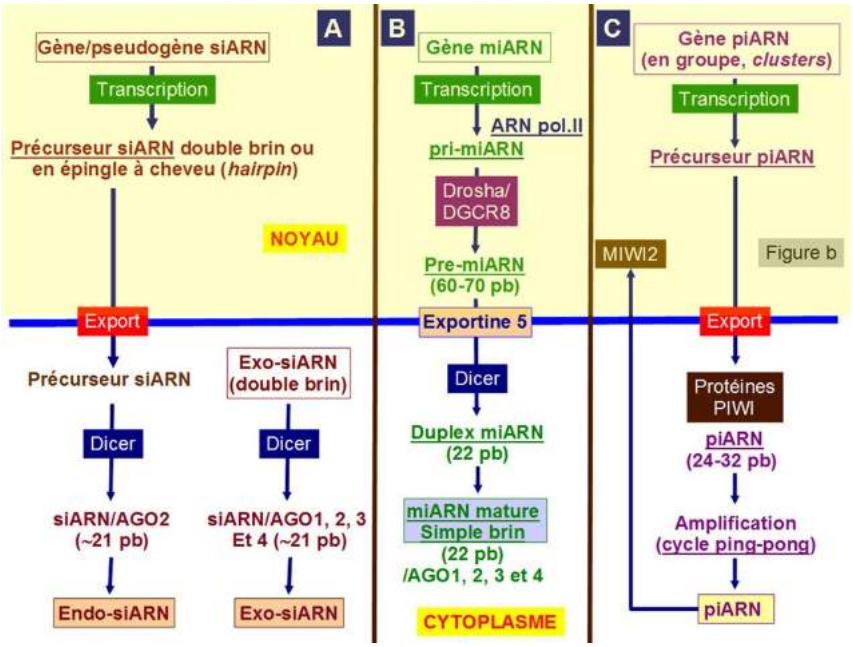
- Le groupement 3'OH du premier exon peut ainsi réagir avec l'extrémité 5'phosphate du deuxième exon pour former une liaison phosphodiester et permettre la libération de l'intron qui sera dégradé.

d) L'épissage alternatif

A partir d'un transcrit primaire on peut avoir deux ou plus ARNm matures qui seront à l'origine de la formation des protéines-isoformes. Ceci est possible grâce à l'épissage alternatif qui consiste en l'élimination de certains exons. En effet certains exons sont constants au niveau des différents ARNm matures et d'autres sont variables et spécifiques du tissu dans lequel se trouve la protéine isoforme.

Place des petits ARNs dans le génome eucaryote

Les trois classes de petits ARNs miARN, piARN et endo-siARNs se distinguent par leur biosynthèse et leur rôle au niveau cellulaire. Ainsi, de nombreuses études ont démontré la participation des piARNs et des siARNs à la défense contre les séquences génomiques parasites (par exemple, répression des transposons). De plus, tous deux peuvent être originaires de transposons, de gènes codant ou de pseudogènes, les piARNs s'exprimant dans les cellules germinales et les endo-siARNs dans les cellules somatiques. Par ailleurs, il est difficile de différencier les miARNs des siARNs, ceux-ci étant essentiellement distingués en fonction de leur origine et plus rarement selon leur taille ou leur fonction. Pour certains, les endo-siARNs seraient des intermédiaires dans l'évolution dont l'adaptation aurait abouti à la formation des miARNs.



Synthèse des petits ARNs (siARN, miARN et piARN).

Les piARNs, furent observés dans des régions du génome où se situaient des séquences répétées et des séquences de transposons. Ces piARNs semblent jouer un rôle important dans le développement de la lignée germinale et la maintenance de l'intégrité du génome. En effet, les protéines PIWI appartiennent à la famille des protéines Argonautes (constituée des protéines AGO et PIWI) et contrairement aux autres membres de la famille AGO, sont exprimées essentiellement dans les cellules de la lignée germinale. Ainsi, les mutants PIWI chez la souris présentent des défauts de gamétogenèse mais se développent normalement. Ainsi, bien que des miARNs et des siARNs soient aussi présents dans les cellules de la lignée germinale, les piARNs y prédominent. Pour contrôler les transposons, notamment dans les cellules de la lignée germinale, le rôle des piARNs est fondamental.

Bien que des inconnues sur le mécanisme précis demeurent, les piARNs en interaction avec les protéines PIWI reconnaissent les transposons dont ils sont complémentaires puis les clivent et les dégradent.

En sus de cette activité, les piARNs semblent au moins en partie contrôler la transcription et agir au niveau post-transcriptionnel. Par ailleurs, il n'est pas exclu que les piARNs en association avec les protéines PIWI agissent aussi au niveau de la chromatine.

Les endo-siARNs quant à eux jouent un rôle dans la répression des transposons. Cela a pu être démontré chez la drosophile dont on avait supprimé la voie piARN. Ce rôle a aussi été démontré dans la lignée germinale femelle de la souris. De plus, dans le génome des mammifères, on retrouve de nombreux pseudogènes normalement non fonctionnels. L'étude par clonage de petits ARNs à partir des ovocytes de souris a montré la présence inattendue de pseudogènes fonctionnels, source de formation d'endo-siARNs par un mécanisme complexe.

Méthodes d'analyse des génomes

Au début des années 1990, la communauté internationale a décidé d'organiser un projet public de séquençage dont le but était d'obtenir, en l'an 2000, la totalité des séquences du génome humain en caractères. En raison de l'ampleur de ce travail, 20 institutions ont convenu de partager la tâche. Chaque centre de séquençage s'était alors engagé à déposer les séquences dans des bases de données publiques dès l'acquisition, pour éviter un accès payant à ces informations. Le génome qui a été séquencé par le consortium international n'appartient pas à un seul et même individu mais à plusieurs donneurs anonymes. Les technologies utilisées en génomique ont été améliorées rapidement après ces grandes avancées.

1-Techniques de séquençage

Le séquençage de l'ADN, consiste à déterminer l'ordre d'enchaînement des nucléotides d'un fragment d'ADN donné.

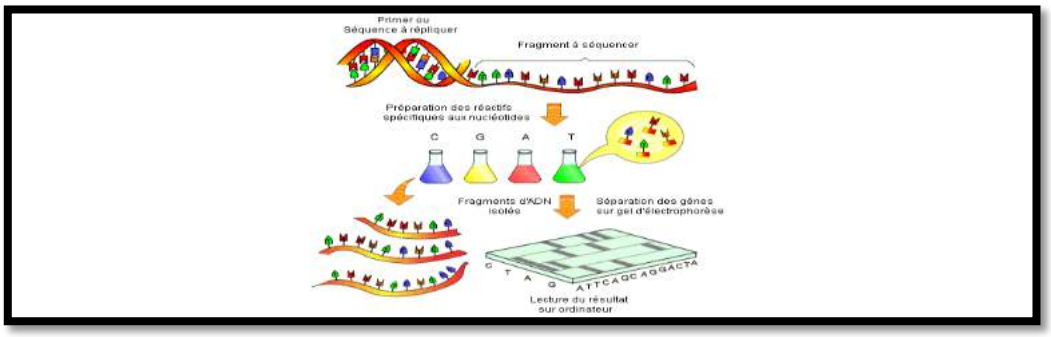
1.1. Méthode de Sanger

Dans quatre tubes à essais distincts, sont introduits en parallèle des brins d'ADN cibles dénaturés en grand nombre, des amorces (sur laquelle se fixe la polymérase), la polymérase elle-même et des nucléotides (dATP, dTTP, dCTP, dGTP).

Sont ensuite ajoutés un nucléotide modifié ou di-déshydro-nucléotide (ddATP, ddTTP, ddCTP, ddGTP) de type différent dans chaque tube (manque un groupement-HO nécessaire à l'assimilation du nucléotide suivant par la polymérase). Ainsi, lors de l'assimilation d'un de ces nucléotides par la polymérase, la réplication du brin est stoppée. Une fois les réactions terminées, on obtient dans chacun des quatre tubes des doubles brins d'ADN de tailles variables, en fonction de leur arrêt par les nucléotides modifiés.

On place alors le contenu des tubes dans quatre puits distincts (correspondant aux quatre nucléotides possibles) de gel d'électrophorèse. Après leurs migrations, dans ce gel on peut aisément retrouver l'ordre des nucléotides de la séquence concernée. Une simple lecture horizontale de ce gel permet de connaître l'ordre des bases de la séquence.

Afin de voir les fragments d'ADN sur le gel d'électrophorèse. On peut aussi marquer radioactivement les nucléotides puis les exposer à un film photographique: des bandes sombres apparaissent là où se trouvait de l'ADN sur le chromatogramme.



Principe de la technique de séquençage de Sanger

1.2. Automatisation des techniques de séquençage

Aujourd'hui, la plupart des techniques de séquençages sont réalisés par des séquenceurs industriels entièrement automatisés. Ceux-ci utilisent la technique de Sanger mais avec des méthodes de révélation différentes.

Les fragments d'ADN sont marqués par des marqueurs fluorescents; leur taille est ensuite déterminée par chromatographie ou électrophorèse assistée par ordinateur. Avec ces techniques, on peut séquencer jusqu'à 1000 bases avec les meilleurs séquenceurs contre 200 à 300 via une méthode manuelle. En effet, lors de l'électrophorèse manuelle, le nombre de bases est limité afin de ne pas rendre le chromatogramme illisible par la surcharge des bandes et ainsi ne plus permettre une lecture horizontale.

C'est notamment grâce à la rapidité de ces appareils que le séquençage du génome humain fut réalisé en un temps record par rapport aux prévisions effectuées lors du démarrage du projet.

- Pyroséquençage

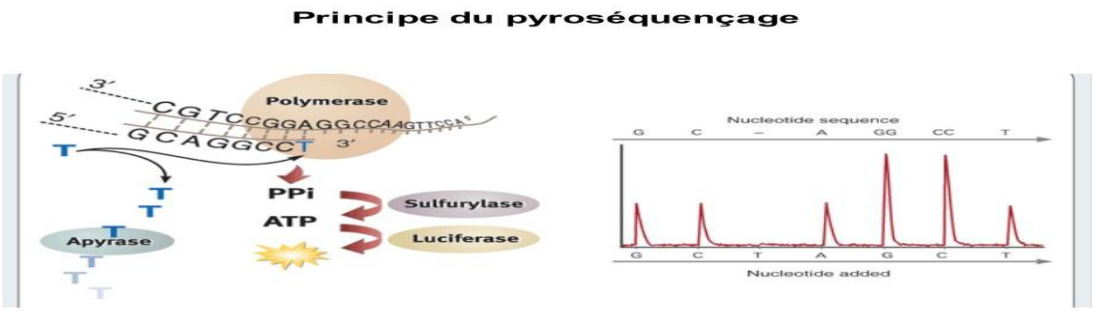
Il s'agit d'une technologie de séquençage direct utilisée principalement pour détecter des mutations sur des séquences ciblées par comparaisons de brins. Son principal avantage est sa rapidité qui autorise enfin l'analyse de grandes séries d'individus.

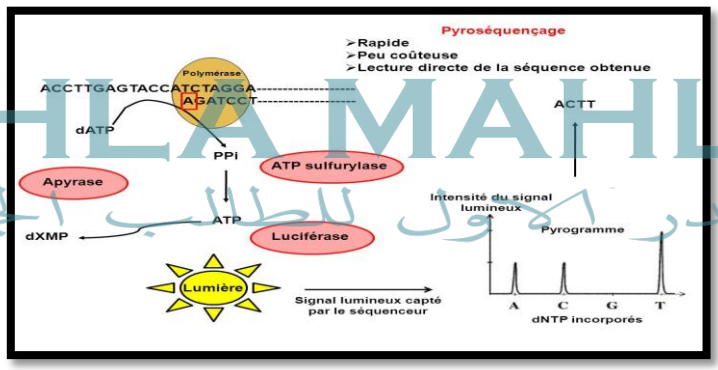
- Principe du pyroséquençage

Les enzymes et leurs substrats sont introduits dans une cartouche adaptée, et le brin d'ADN dans une micro cuvette. Le tout est mis dans un pyroséquenceur, appareil entièrement informatisé. Le pyroséquenceur propose alors tour à tour un type de base. Si la base correspond, elle est alors incorporée par la polymérase au brin.

Celui-ci rejette alors un phosphate inorganique (PPi) transformé en ATP par la sulfurylase puis en lumière par la luciférase. Cette émission lumineuse est alors captée par un détecteur photosensible qui transmet un signal à l'ordinateur.

La hauteur de ce pic est fonction de l'intensité du signal lumineux, elle-même proportionnelle au nombre de nucléotides incorporés en même temps. On peut donc déduire la séquence de la taille des pics obtenus. Par ailleurs, en cas de mélange de nucléotides à une même position (polymorphisme de séquence), la taille des pics permet d'avoir une quantification de la proportion de brins porteurs de l'un ou l'autre des nucléotides.





2. Analyse fonctionnelle de l'expression des gènes: Les puces à ADN

2.1. Définition

Un *microarray* est une surface solide et généralement plane sur laquelle sont fixées des molécules. Ces molécules sont le plus souvent des acides nucléiques : ARN, ADN, ADNc, ARNc (= ARNs de séquence complémentaire d'ARN messagers). . .

Cependant, il est utile de préciser qu'ont été également développés des arrays :

• de protéines, enzymes, peptides, anticorps; •de molécules lipophiles; •de molécules glycosylées; •de petites molécules (métaux).

L'appellation *microarray*, signifiant microalignement ou microarrangement, provient de la manière ordonnée de la répartition de sondes sur un support. Le mot ordonné est important : on sait ce que l'on fixe sur le support et où on le fixe. La nature du support est variable, mais il s'agit de verre dans la grande majorité des cas. Enfin, le fait que certain supports soient en silicium et que les systèmes d'hybridation soient miniaturisés ont amené à l'utilisation du terme « puces à ADN » (*DNA chips*) pour qualifier ces *microarrays*.

2.2. Principe d'analyse sur *microarray*

Il s'agit d'une application du principe d'hybridation moléculaire. Le procédé initial, décrit par E. Southern en 1975, consistait en un transfert d'ADN à analyser sur une surface solide (nitrocellulose). Afin de voir si une séquence d'intérêt était présente dans ce fragment, on ajoutait au milieu un ADN marqué comportant une séquence complémentaire, appelé sonde. Le nom de sonde (extrapolé de « sonder ») implique que l'acide nucléique utilisé est marqué (fluorophore, isotope, enzyme. . .). L'hybridation entre la cible (la séquence recherchée) et la sonde était mise en évidence grâce au marquage de la sonde. Cette technique fut fondatrice des premiers diagnostics en génétique (principe des polymorphismes de restrictions : *restriction fragment length polymorphism*, RFLP). Les *microarrays* sont nés de l'adaptation de l'hybridation moléculaire à une grande échelle sur un format réduit préconisée par certains chercheurs. Ici il s'agit cependant d'hybridation inverse. Les sondes, fragments d'acide nucléique ARN ou ADN de petite taille fixés sur un support, sont utilisées pour repérer de manière spécifique des séquences d'intérêt dans un milieu complexe contenant des milliers de séquences nucléiques différentes. Ces hybridations sont hautement spécifiques dans les conditions expérimentales utilisées. L'array sert ainsi à détecter des acides nucléiques extraits d'un échantillon biologique, et à les quantifier dans certaines conditions. Le marquage de ces extraits est généralement effectué à l'aide de fluorophores.

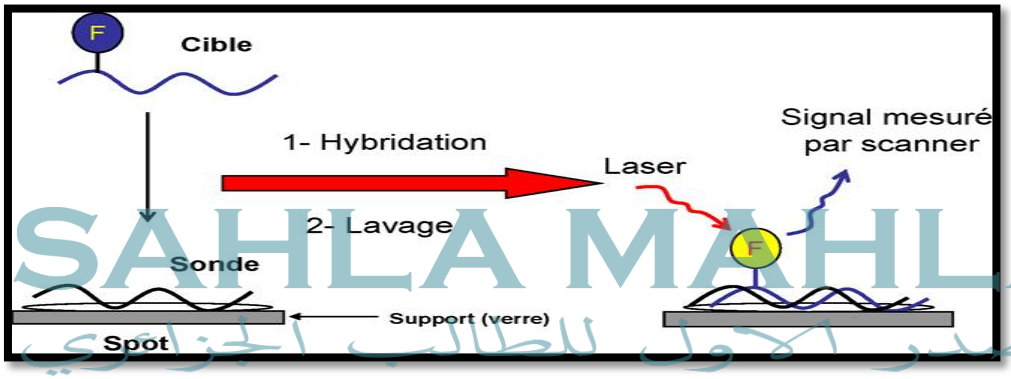
D'une manière générale, les études portent sur des acides nucléiques obtenus à partir d'ARN messagers ; l'array mesure alors l'expression des gènes dont sont issus les messagers, car le signal mesuré sera proportionnel au nombre de messagers transcrits. Cet outil s'avère avoir une grande puissance dans la mesure où l'on peut fixer une très importante quantité de différentes molécules d'ADN : il est possible de mesurer, de manière simultanée, l'expression de milliers de gènes. Il existe différentes terminologies pour nommer ce qui est fixé sur le

support solide et ce qui est mis à hybrider avec les ADN fixés dans l'array. Ce qui est fixé sur le support sera appelé sonde et que les acides nucléiques extraits à tester par hybridation moléculaire seront appelés extraits ou cibles.

2.3. Analyse de l'expression des gènes

Une matrice de quantification du signal en fonction des spots appelée matrice de quantification est réalisée suite au scannage du ou des *arrays*. Ensuite, les données quantifiées de cette matrice sont reportées sur une matrice d'expression de gènes dont les lignes horizontales représentent les gènes et les colonnes les tissus correspondants.

Une fois réalisée cette matrice des gènes différentiellement exprimés, on procède à des regroupements de gènes et/ou d'échantillons en fonction de leur profil d'expression. À cet effet, des algorithmes de classification sont utilisés afin de prédire des regroupements et des classifications.



2.4. Intérêt des puces d'expression

Une cellule fonctionne grâce à l'ensemble de ses gènes dont la grande majorité a pour finalité la synthèse de protéines.

Cependant, quoique chaque cellule d'un organisme contienne le même équipement en terme de gènes, on observe des différences significatives dans leur fonctionnement d'un type de cellule à un autre : certains gènes sont activés, d'autres réprimés. Ce mécanisme de contrôle de l'expression des gènes aboutit à la transcription ou non transcription sous forme d'ARN messager de l'information codée par le gène. De la formation de la première cellule embryonnaire à l'individu adulte, de nombreux tissus différents ont été générés, ce qui met en évidence l'importance de la régulation des gènes et de leur expression. Afin d'étudier ces processus, les chercheurs ne disposaient auparavant que de méthodes ne permettant d'étudier

un gène à la fois, en regardant dans différentes conditions s'il était actif ou non, ou bien en étudiant les effets de la suppression de l'activation du gène (*knockout* de gènes). L'avènement de la technique des *microarrays* a totalement changé cette approche. Il est maintenant possible d'analyser non seulement l'expression d'un gène isolé, mais d'observer dans sa globalité l'expression sous forme d'ARN messager ou d'ADNc d'un ensemble de gènes, voire de tous les gènes d'une culture cellulaire ou d'un tissu et de comparer cette expression dans différentes circonstances. On réalise alors de véritables « instantanés » de ce qui se passe à l'échelon cellulaire.

L'expression des gènes est analysée en préparant des sondes à partir d'ADNc, c'est-à-dire d'ADN complémentaire des ARN messagers, obtenus par transcription inverse.

Les sondes déposées sous forme d'array représentent la totalité des gènes exprimés dans une cellule. On estime qu'il y a entre 10 000 à 20 000 espèces d'ARN messagers dans une cellule de mammifères. La puce permet d'analyser par hybridation l'expression des gènes de cellules placées dans des conditions particulières. Dans ces conditions, les messagers (ou leur ADNc ou ARNc) des gènes actifs, marqués par un fluorophore, hybrident aux sondes de séquence complémentaire. Il est ainsi possible d'étudier, d'analyser et de comparer l'expression de milliers de gènes dans différentes circonstances telles que :

• comparaison de niveaux d'expression de gènes entre différents tissus, par exemple : sain versus cancéreux;

• analyse d'expression de gènes en présence d'agents toxiques, thérapeutiques;

• analyse d'expression de gènes de cellules placées dans diverses conditions différentes (croissance, différenciation).

Réaliser un profil de gènes, c'est étudier un ensemble de gènes dans deux conditions différentes.

Grossièrement, deux démarches peuvent être menées afin de produire des profils d'expression de gènes. Soit on réalise deux arrays indépendants, l'un hybridant les messagers (ou dérivés) dans la condition 1 et l'autre dans la condition 2, soit on hybride les deux extraits, marqués de manière différente dans chacune des conditions, sur le même array : il s'agit ici d'hybridation comparative. Dans ce cas, classiquement, les marqueurs fluorescents Cy3 et Cy5 sont utilisés.

2.5. Démarche expérimentale

2.5.1. Extraction de l'acide nucléique

Elle est réalisée selon les procédures habituelles. En utilisant des kits d'extraction par exemple. L'important, est de veiller à obtenir la meilleure reproductibilité possible, car si deux échantillons identiques ne donnaient pas un extrait qualitativement et quantitativement similaires, l'interprétation serait biaisée.

La qualité de l'acide nucléique extrait est primordiale. C'est pourquoi les processus de préparation de l'échantillon comportent une étape d'analyse de la qualité des extraits. Cela peut être effectué de différentes manières : électrophorèse sur gel d'agarose, mesure de l'absorbance au spectrophotomètre (rapport de DO260/280 nm), mesure quantitative de fluorescence…etc

2.5.2. Marquage

La méthode de marquage est fonction de la technologie de l'*array* utilisé. Il est intéressant de distinguer ici les deux approches : marquage unique (exemple : Affymetrix) versus hybridation compétitive.

2.5.2.1. Marquage unique

On n'hybride sur une puce qu'une seule sorte d'extrait. Donc si l'on désire comparer, par exemple, le taux d'expression de gènes entre un tissu sain et un tissu cancéreux, il faudra utiliser deux puces équipées de sondes identiques, et hybrider l'une avec l'extrait de tissu sain et l'autre avec l'extrait de cellules cancéreuses. Les deux bases de données de résultats pourront alors être comparées.

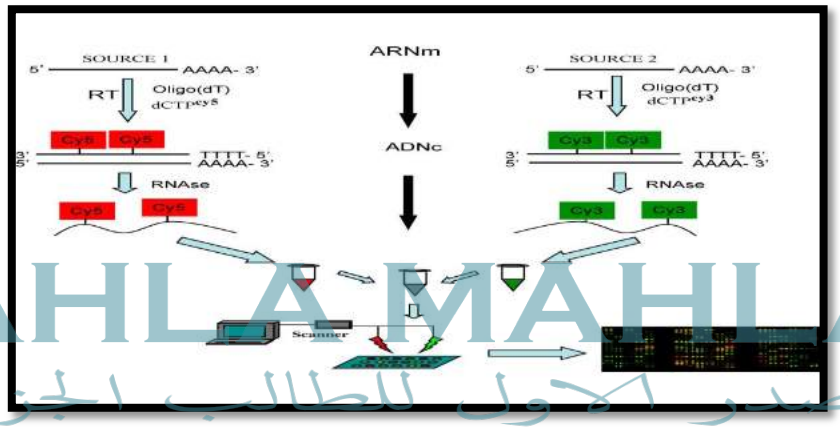
La société Affymetrix qui pratique cette méthode a défini des protocoles très précis, ayant pour but une comparaison facilitée des résultats obtenus par différents laboratoires utilisant leur technologie. Le produit à analyser est ici marqué par de la biotine. L'ajout ultérieur de streptavidine-phycoérhytrine permet de mesurer le signal fluorescent émis après excitation par laser.

2.5.2.2. Hybridation compétitive

Deux extraits différents sont hybridés sur la même puce, chacun étant marqué par un fluorophore différent. Rappelons ici que le couple Cy3/Cy5 est très utilisé. Ces formats facilitent les analyses comparatives : il est ainsi possible de comparer en une seule hybridation

l'expression de tissus pathologiques par rapport à des tissus sains, ou bien de deux types de tissus, ou de deux circonstances de cultures pour un type de cellule. Couramment pratiquée sur les *microarrays*, cette méthode limite la variabilité entre les expérimentations, car l'utilisation d'un seul array au lieu de deux élimine tous les biais potentiels issus de la comparaison de deux arrays.

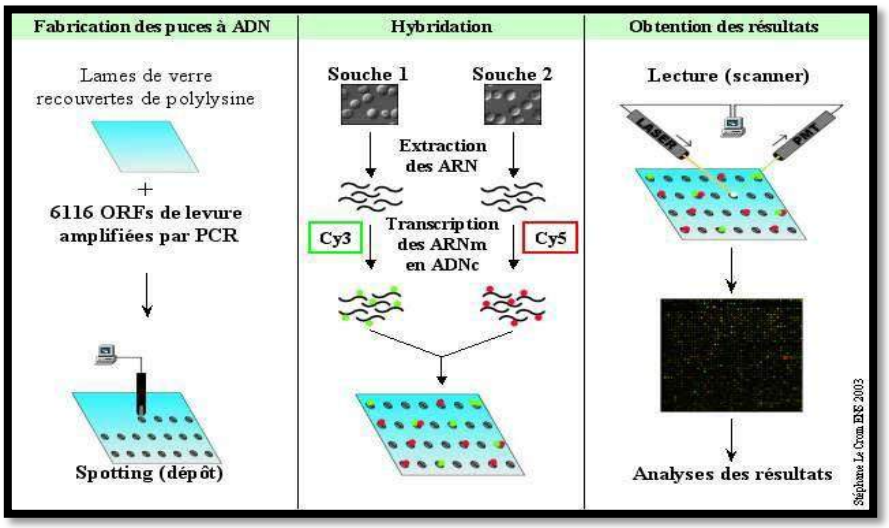
La reverse transcription des ARN messagers peut être effectuée, en utilisant une amorce poly-dT, permettant d'atteindre une taille de 3000 pb. Une méthode de marquage couramment utilisée consiste en l'incorporation de résidus dC comportant un *aminolinker* auquel un résidu Cy sera fixé en fin de synthèse. Cette seconde stratégie comporte un avantage : l'efficacité de marquage par Cy3 et Cy5 sera la même, alors que si l'on incorpore Cy3 et Cy5 directement en cours de synthèse, l'efficacité de la reverse transcriptase est moins bonne avec Cy5 qu'avec Cy3.



2.5.3. Hybridation

Les conditions d'hybridation des lames sont soumises aux paramètres usuels : concentration en sel, température optimale, humidité contrôlée (chambre d'hybridation scellée) et agitation. Ce processus est automatisable. La durée d'hybridation est de 12 à 24 heures. Pour éviter les hybridations avec les sites de polyadénylation de l'ADNc on ajoute du poly-A ou du poly-T au milieu d'hybridation. De plus on peut ajouter au milieu d'hybridation de l'ADN Cot1 isolé de placenta humain enrichi en séquences répétitives, telles que, celles des familles *Alu* et *Kpn* afin de limiter au maximum le bruit de fond causé par les hybridations non spécifiques.

Le lavage est d'une manière générale automatisé (solution stringente ou lavage à haute température). Une fois cette étape terminée, il faut interpréter l'*array*. Pour cela, il faut exciter les fluorophores, mesurer la fluorescence émise en retour et interpréter les images obtenues.

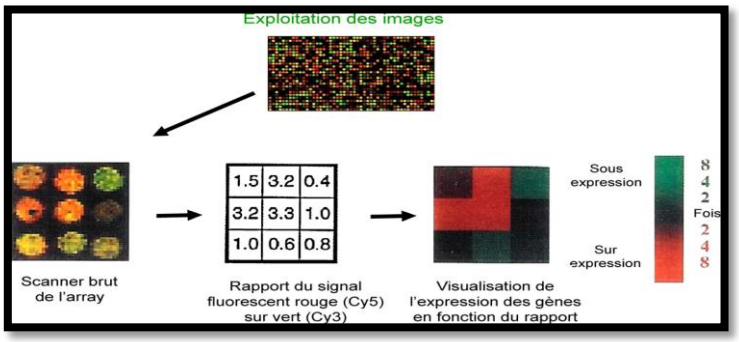


2.5.3. Analyse des images

Acquisition de la fluorescence des arrays. Pour exploiter l'émission de fluorescence des produits hybridés aux sondes, la lame de verre est introduite dans un scanner. Un rayon laser excitateur balaie la lame et un photomultiplicateur mesure la fluorescence émise en retour. Rappelons ici que sur les puces Affymetrix, une seule mesure est effectuée (phycoérhytrine) et que, en général, les autres arrays ont un double marquage. Le scanner peut donc être équipé de plusieurs lasers, permettant d'effectuer la détection de plusieurs molécules fluorescentes.

Mesure des signaux. Les images obtenues sont traitées par un logiciel, de manière à affecter, à partir de deux images monochromes, des couleurs aux signaux mesurés : vert pour Cy3 et rouge pour Cy5. D'autres couleurs peuvent être affectées.

Le photomultiplicateur mesure donc l'intensité du signal émis par le spot après excitation de Cy3 par un laser émettant à la longueur d'onde ad hoc, puis le signal émis par Cy5 excité par un second laser à une longueur d'onde différente. L'expression d'un des deux échantillons est analysée par rapport à l'autre servant de référence (example ; tissus cancéreux versus tissu sain). Pour analyser ces données, on calcule le rapport d'intensité de fluorescence, traité de manière appropriée, des deux échantillons hybridés de façon compétitive sur le même *microarray*. Si les deux extraits hybrident à une sonde en quantité équivalente, le spot sera jaune, ce qui indiquera que l'expression du gène concerné est similaire dans les deux conditions analysées. Si le signal d'un spot est vert, le gène est sous exprimé par rapport à la référence, s'il est rouge, il est surexprimé.



2.5.4. Analyse des données

Afin de pouvoir comparer de manière précise les deux échantillons d'ARN marqués de manière différente, il est nécessaire de procéder à une normalisation. En effet, l'extraction d'ARN et le marquage fluorescent sont effectués de manière indépendante : il faut donc normaliser l'intensité des signaux obtenus avant de comparer quoique ce soit. Ce processus comprend plusieurs étapes :

- ***Élimination des mauvaises images (décalées, noires, de mauvaise qualité).*** Le plus simple est d'éliminer ces mauvaises images au risque de perdre de l'information. Sinon il est possible de retourner sur l'analyse d'image et de tenter d'obtenir un meilleur signal. Cette démarche est consommatrice de temps.

- ***Soustraction du bruit de fond.*** Trois méthodes sont particulièrement utilisées.

Méthode globale. Certains arrays comportent des milliers de gènes dont le niveau d'expression de la plupart s'avère être similaire dans les deux conditions analysées. On ajuste alors le rapport de fluorescence de manière à ce qu'il soit égal à un pour les gènes s'exprimant de manière identique dans les deux conditions. Ce procédé n'est applicable qu'aux arrays comportant plus de 1000 gènes.

Utilisation de gènes de ménage. On introduit dans chaque array des spots correspondants à plusieurs gènes de ménage. Un gène de ménage (*house keeping gene*) est appelé ainsi, car il montre, du moins en principe, un niveau d'expression constant dans tous les tissus (exemple : β- actine). Le signal obtenu par l'expression de ces gènes permet d'ajuster la valeur du signal des autres spots.

Utilisation d'un contrôle interne. On introduit un volume égal de contrôle interne aux deux échantillons d'ARN avant le marquage. Après hybridation le signal sera adapté de manière a être égal pour les deux extraits.

Normalisation proprement dite : traitement du signal. La normalisation a pour objectif de corriger les biais systématiques des données et de supprimer dans la mesure du possible toute influence non biologique sur ces données. La normalisation doit donc permettre de tirer de vraies informations biologiques des données et de comparer les données d'un *array* à l'autre, voire d'une plate forme d'*array* à une autre.

3. Applications

- En pathologies cliniques

- Analyse de l'expression des gènes
- Analyse du fonctionnement des gènes dans les tissus sains ou pathologiques
- Identifier les gènes particulièrement actifs dans certaines cellules
- Identification de nouveaux gènes associés aux différentes pathologies
- Identification des mutations génétiques et leur application en génétique médicale

- En recherche pharmaceutique

- Mieux comprendre l'action des médicaments : Les puces à ADN permettent de savoir quels sont les gènes qui sont modulés dans une cellule sous l'effet d'un médicament. Ceci permet d'avoir une vision plus complète de son action.
- Des mutations de type SNP sont à l'origine des différences observées dans la réponse individuelle aux médicaments.
- Possibilité d'ajuster les doses de traitement en fonction du profil génétique de chaque patient.

3. Chromatin ImmunoPrecipitation

La régulation de la transcription est un processus fondamental des cellules vivantes qui contrôle la différenciation, l'expression spécifique des gènes, le développement, la prolifération ou encore pour l'adaptation à un environnement.

L'étude de la régulation de la transcription passe par l'identification des éléments du génome impliqués dans les différents processus biologiques. Les approches méthodologiques ChIP-Seq, permettent d'analyser les régions régulatrices de l'expression génique en identifiant les régions de l'ADN associées aux activités régulatrices. Ainsi lorsqu'on veut déterminer où la protéine est liée sur un génome à grande échelle, on peut utiliser une puce à ADN (Chip on chip) ou par séquençage à haut débit (Chip-Seq).

Définition

L'immunoprécipitation de la chromatine associée au séquençage –(Chromatin Immuno-Precipitation Sequencing) est une méthode utilisée pour analyser les interactions entre les protéines et l'ADN. Cette technologie combine l'immunoprécipitation de la chromatine (ChIP), en utilisant des anticorps spécifiques d'une protéine d'intérêt et le séquençage haut débit. Le but étant de cartographier tous les sites de liaison sur l'ADN de cette protéine à l'échelle du génome. Il est ainsi possible de cartographier toutes les régions génomiques associées avec une histone nucléosomale portant une modification biochimique (acétylation, méthylation, etc.) caractéristique de l'activité transcriptionnelle.

Principe du chip on chip

Le principe de cette technique consiste à fixer de façon covalente les protéines liées à l'ADN par un traitement chimique, de fragmenter la chromatine, d'immunoprécipiter les fragments en présence de l'anticorps d'intérêt et après purification et élimination des protéines, de séquencer les fragments d'ADN obtenus. Le principe de fonctionnement de la technique ChIP-chip se compose de deux parties, la partie ChIP pour Chromatine ImmunoPrecipitation et la partie chip qui correspond a une étude sur puce à ADN.

La partie ChIP consiste dans un premier temps à fixer les protéines à l'ADN *in vivo* à l'aide le formaldehyde. Vient en suite l'étape de "sonication" ou l'ADN est fragmenté aléatoirement, puis a l'aide d'anticorps spécifiques intervient le processus d'ImmunoPrecipitation qui permet de retenir les fragments porteurs de la protéine étudiée. L'ADN et la protéine sont alors séparés.

Dans la partie chip, on récupère des fragments d'ADN provenant de deux expériences, une ou les fragments d'ADN ont subi l'étape d'ImmunoPrecipitation et l'autre non, chacune des expériences étant marquée par un fluorochrome différent. Les fragments sont alors hybridés sur une puce à ADN et les niveaux de fluorescence des sondes vont permettre de déterminer si le fragment correspond à un site de fixation.

- **Interprétation des résultats**
- On utilise ensuite une puce à ADN en vue d'identifier les fragments recueillis dans la première partie. On a en effet une collection de fragments dont on sait qu'ils interagissent avec une protéine d'intérêt. Il est évidemment intéressant de mieux caractériser ces fragments, dans le but de les localiser sur le génome ou d'étudier leur séquence même.
- Des caractéristiques communes des fragments (polarité, séquences consensus...) peuvent être mises en évidence. D'autre part, on peut mettre en évidence de nouveaux gènes régulés par la protéine d'intérêt.
- Il est néanmoins à noter que l'utilisation de puces à ADN pour l'identification et la caractérisation des fragments n'est pas obligatoire. On peut aussi utiliser une

classique PCR pour identifier des fragments connus et vérifier leur présence ou leur absence.

- L'utilisation de puce à ADN en deuxième partie permet cependant d'accélérer le processus d'identification des séquences, puisqu'on teste potentiellement plusieurs milliers de séquences à la fois.
- L'avantage majeur du ChIP est qu'il s'effectue *in vivo*, c'est-à-dire que l'information tirée de cette expérience provient directement d'une analyse faite sur les cellules vivantes. Ainsi, après la croissance des cellules dans les conditions souhaitées, toutes les protéines qui touchent l'ADN y sont immobilisées par des liens covalents à l'endroit précis de leur interaction grâce à un traitement au formaldéhyde.
- Par la suite, les cellules sont lysées, puis les complexes ADN-protéines sont fragmentés en courts segments avant d'être immunoprécipités.
- L'immunoprécipitation, effectuée à l'aide d'un anticorps dirigé contre la protéine d'intérêt permet de récupérer les protéines cibles ainsi que toutes les régions d'ADN auxquelles elles étaient liées lors du pontage initial au formaldéhyde.
- L'analyse statistique des données consiste a rechercher des régions de pics significatifs synonymes de sites de fixation.

ChIP on ChiP Sequencing

- Liaison covalente *in vivo* des protéines à l'ADN, en général par l'utilisation de formaldéhyde
- Extraction de l'ADN de la cellule
- Fragmentation de l'ADN en courts brins(Par exemple par sonication)
- Sélection des fragments (associés à la protéine étudiée) grâce à un anticorps correspondant
- Précipitation des complexes ADN-protéine-anticorps, élimination du surnageant (i.e. ADN non associé à la protéine d'intérêt)
- Séparation du complexe ADN-protéine pour ne garder que l'ADN (protéinase K par exemple)

On obtient ainsi une collection de fragments d'ADN d'assez courte taille et dont on sait qu'ils interagissent avec une protéine d'intérêt.

L'objectif va être de détecter des pics significatifs, synonymes de sites de fixation

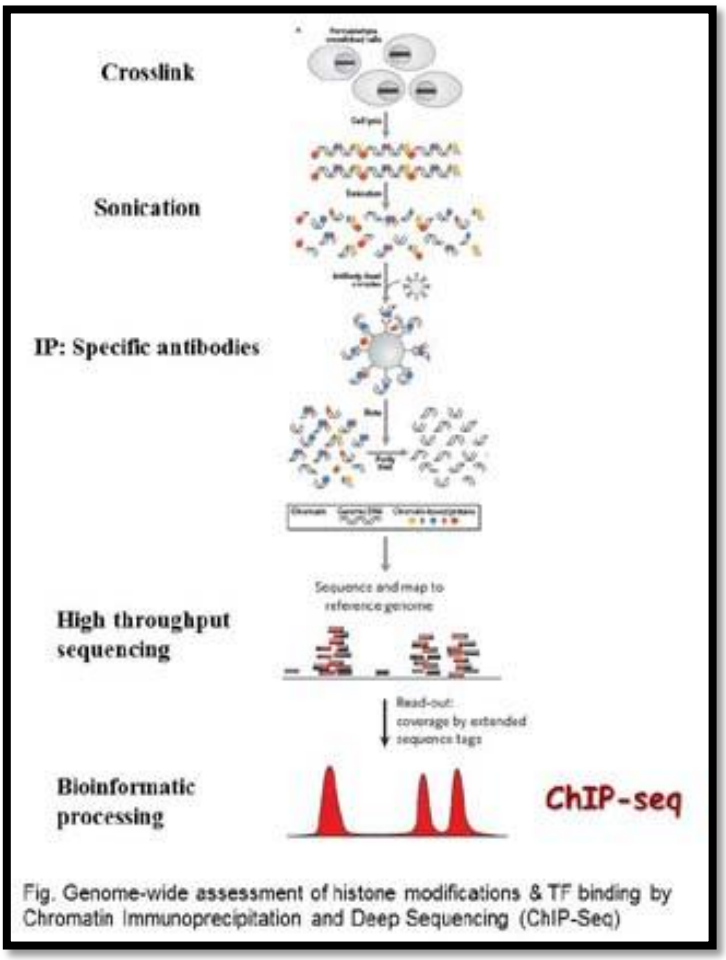


Fig. Genome-wide assessment of histone modifications & TF binding by Chromatin Immunoprecipitation and Deep Sequencing (ChIP-Seq)

Principe du Chip sequencing

Analyse des génomes, annotation et recherche de similarités par outil bioinformatique

L'alignement de plusieurs séquences et l'identification des portions de ces séquences qui sont conservées permet d'identifier des domaines ou des motifs fonctionnels (par exemple impliqués dans une réaction catalytique ou dans la reconnaissance de partenaires cellulaires).

1. Définition et Intérêt de la bioinformatique

Lors de sa création, la bioinformatique correspondait à l'utilisation de l'outil informatique pour stocker et analyser les données de la biologie moléculaire. Cette définition originale a maintenant été étendue et le terme bioinformatique est souvent associé à l'utilisation de l'informatique pour résoudre les problèmes scientifiques posés par la biologie dans son ensemble. Il s'agit dans tous les cas d'un champ de recherche multidisciplinaire qui associe informaticiens, mathématiciens, physiciens et biologistes.

Pour aboutir à la formulation de ces modèles et à ces prédictions, il est indispensable de tout d'abord collecter et organiser les données à travers la création de bases de données.

2. Les bases de données Une base de données est un ensemble structuré et organisé permettant le stockage de grandes quantités d'informations afin d'en faciliter leur utilisation (ajout, mise à jour, recherche et éventuellement analyse dans les systèmes les plus évolués). Ils permettent ainsi d'accéder à des informations dans les bases de données et d'effectuer des calculs, des analyses, des comparaisons, en ligne et sur des ordinateurs distants.

2.1. Les banques de données utiles dans le domaine de la génétique

Ils correspondent à différentes bases de données qui permettent d'accéder aux données du génome humain (et de celui d'autres espèces) à l'aide d'une interface. En plus des données de séquence, ces navigateurs permettent d'accéder à de nombreuses données d'annotation (gènes avec exons et introns, sites de fixation, régions d'homologie).

Les plus populaires sont : ● Ensembl (European Bioinformatics Institute / Wellcome Trust Sanger Institute) ● NCBI (National Cancer for Biology Information) ● UCSC (University of California Santa Cruz)

2.2. Banques de données cliniques

- **Orphanet** (Base de données sur les maladies rares et sur les médicaments orphelins)
- **GENDIAG** (syndromes génétiques et les anomalies du développement chez l'homme)
- Si on doit interroger une banque et utiliser un ou plusieurs logiciels de bioinformatique, on doit spécifier une **requête**, à chaque fichier est attribué un numéro d'identité **unique au sein de** la banque, appelé **numéro d'accession.** Ces **numéros d'accession permettent** d'accéder rapidement et sans ambiguïté à n'importe quel fichier dans les banques. Comme ces numéros ne sont pas les mêmes d'une banque à

une autre, alors, pour identifier une fiche dans une banque il faut fournir **Banque: Numero d'accession.**

Exemple:

- SWISSPROT: **P04156**, MEDLINE: **86300093**, OMIM: **123400**
- Le problème crucial de l'analyse de séquences génomiques est l'identification des séquences codantes.
- Chez les procaryotes, cette identification est grandement facilitée par la quasi absence de séquences non codantes, et par la possibilité de reconnaitre assez facilement les phases ouvertes de lecture, les promoteurs et les terminaisons des gènes.
- L'identification des unités transcriptionnelles reste possible grâce a des outils informatiques capables d'identifier un gène sur plusieurs critères : la présence d'une phase ouverte de lecture, de signaux d'épissage, la composition en bases. On utilise aussi pour déterminer si une séquence est codante la comparaison à l'ensemble des données acquises par les programmes de séquençage d'ADN.
- Une similarité entre une séquence génomique et une séquence ADNc permet de conclure que cette séquence est transcrite. On utilise l'information a priori concernant l'organisme pour identifier les gènes et l'organisation des chromosomes.

G	**20.4**	11.4	**5.4**	12.4
A	40.7	39.0	45.8	41.9
T	24.7	33.7	44.2	33.9
C	14.2	16.8	**4.6**	11.9

(les pourcentages en gras sont éloignés de la moyenne pour les symboles correspondants)

3. L'annotation : outils et bases de données

La connaissance de la séquence du génome humain n'aurait qu'une portée limitée si elle n'était annotée à différents niveaux. Ainsi l'annotation est un processus complexe qui peut être subdivisé en trois catégories : l'annotation syntaxique, l'annotation fonctionnelle et l'annotation relationnelle:

L'annotation syntaxique qui permet d'identifier les séquences présentant une pertinence biologique (gènes, signaux, répétitions, …)

L'annotation fonctionnelle qui permet de prédire les fonctions et produits potentiels des gènes préalablement identifiés (similitudes de séquences, motifs, structures, ...) et de collecter d'éventuelles informations expérimentales.

L'annotation relationnelle qui permet enfin de déterminer les interactions que les molécules biologiques préalablement identifiés sont susceptibles d'entretenir (familles de gènes, réseaux de régulation, réseaux métaboliques, …).

4. Annotation fonctionnelle en relation avec la structure des protéines

Ce type d'annotation en relation avec la structure des protéines sera d'un apport primordial pour l'interprétation des mutations responsables de maladies génétiques. Nous pouvons distinguer plusieurs niveaux dans la description de la structure des protéines :

● La structure primaire : elle correspond à la séquence des acides aminés constituant la protéine. Il s'agit d'un assemblage linéaire des acides aminés codés par l'ARN messager.

● La structure secondaire : elle décrit un niveau structural plus complexe : les structures secondaires qui sont représentées par les repliements locaux de la protéine. Elle comporte les structures en hélices α et les feuillets β et enfin les coudes (types I, II, III et γ).

● La structure tertiaire : décrit la structure tridimensionnelle de la protéine ou plus précisément d'une forme particulière que peut prendre dans l'espace la protéine d'intérêt dans des conditions expérimentales données et ceci à un temps t.

● La structure quaternaire : permet de décrire les interactions entre protéines.

Parallèlement à ces données classiques, des annotations complémentaires sont de plus en plus fréquemment disponibles (domaines protéiques en relation avec une structure ou une fonction particulières, structure de protéines mutantes …). Parmi les outils et bases de données qui sont indissociables dans le cas des structures protéiques nous pouvons citer ● Uniprot/Swiss Prot/Expasy (Uniprot Consortium) ● Protein Data Bank (Research Collaboratory for Structural Bioinformatics) ● Topspan (Open Protein Structure Annotation Network) ● NCBI (National Cancer for Biology Information)

5. Alignement des séquences

En bioinformatique, l'opération d'alignement vise à identifier des zones communes à un groupe de séquences. Trois Situations sont possibles pour une position donnée de l'alignement: .Les caractères sont les mêmes: Identité ; les caractères ne sont pas les mêmes: Substitution; L'une des positions est un espace: Insertion/ délétion

- **Intérêts de l'alignement**: d'une façon générale, un alignement permet l'identification

- de motifs fonctionnels ou structurels conservés
- étude phylogénétique
- étude comparative des génomes
- prédiction de gène
- prédiction de la structure 2D/3D des protéines

- caractérisation de la fonction des protéines
- prédiction de la structure et fonction des ARN

- Algorithme de programmation dynamique

Principe:

En informatique, la programmation dynamique est une méthode algorithmique pour résoudre des problèmes d'optimisation. La programmation dynamique consiste à résoudre un problème en le décomposant en sous-problèmes, puis à résoudre les sous-problèmes, des plus petits aux plus grands en stockant les résultats intermédiaires.

Trois types d'algorithmes d'alignement de deux séquences:

●**Alignement global:** l'Algorithme de référence est celui de Needleman and Wunsch. Les séquences vont être alignées sur toutes leurs longueurs. Utilise quand les séquences ont a peu prés la même longueur. Il est employé pour aligner les séquences dont on soupçonne l'homologie.

●**Alignement semi-global:** (pas de pénalités des gaps aux extrémités). Utilisé quand une séquence est plus courte que l'autre ou quand on recherche des chevauchements aux extrémités.

●**Alignement local:** (connu comme l'algorithme de Smith and Waterman). Utilisée lorsqu'on veut aligner deux séquences de tailles très différentes (par exemple pour une recherche de sous séquences). L'algorithme recherche les deux sous-régions les plus conservées entre les deux séquences. Seulement ces deux régions seront alignées.

L'alignement se fait par fonction de score: La méthode d'alignement la plus classique consiste à utiliser une fonction de score : on attribue un certain nombre de points a chaque alignement et on sélectionne l'alignement (ou les alignements) de score le plus élevé. Ceci sous-entend qu'on est capable de calculer le score de tous les alignements possibles. On considère par exemple, trois types d'appariements associés à trois scores différents : — $\sigma match = \sigma(a, a)$; — $\sigma mis = \sigma(a, b)$ pour $a \neq b$; et — $\sigma gap = \sigma(a, -) = \sigma(-, a)$. (Le coût σmatch est généralement nul ou positif alors que les coûts de substitution σmis et σgap sont négatifs).

Déterminer le meilleur alignement entre deux séquences S1 et S2, consiste donc à déterminer au moins deux séquences S1 et S2 qui ont le meilleur score d'alignement.

- Deux types de score en fonction des algorithmes:

Score d'homologie: La valeur du score diminue avec le nombre de différences observées entre les deux séquences.

Score de distance: La valeur du score augmente avec le nombre de différences observées entre les deux séquences

Alignement de 2 séquences par programmation dynamique

Pour aligner 2 séquences efficacement, il s'agit de remarquer que le meilleur alignement peut se calculer de manière récursive sur les deux séquences à aligner : le meilleur alignement entre deux séquences dépend du meilleur alignement pour les sous-séquences qui le composent, et du type d'appariement (match, insertion ou délétion) à une position donnée.

Alignement global de deux séquences

- L'alignement de deux séquences consiste à décider de la manière optimale de les mettre l'une en face de l'autre. Par exemple pour les deux séquences S1 =LALMEE et S2 =LAME on pourrait les aligner de la manière suivante : LALMEE et LA-M-E
- Chaque caractère - introduit dans S2 correspond au fait qu'il y a insertion de 1 caractère dans S1 en l'alignant avec S2 (ici un L et un E). On note - un gap. On substitue donc S1 et S2 à S1 et S2 en rajoutant des caractères de gap le long de la séquence (ici S1 =LALMEE et S2 =LA-M-E). S1 et S2 sont de même longueur (|S1 | = |S2 |).

Exemple : l'algorithme Fasta

- Le format classique utilisé pour manipuler les séquences nucléiques ou protéiques est le format FASTA. Une séquence de ce format comporte une **ligne de description** commençant par un **">"** puis les **lignes de la séquence**. On atteint la **fin d'une séquence** soit lorsqu'on atteint la fin du fichier, soit lorsqu'une autre séquence commence (ligne commençant par un ">").
- Exemple de séquence au format fasta : > sp | P62158 | CALM_HUMAN Calmodulin OS=HomosapiensGN=CALM1PE=1SV=2MADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTIDFPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREADIDGDGQVNYEEFVQMMTAK
- Il s'agit de comparer une séquence à l'ensemble des séquences répertoriées dans une banque. Le nombre de comparaisons à effectuer nécessite l'utilisation d'algorithmes de recherche rapides.
- Wilbur et Lipman proposent un algorithme reposant sur la notion de "mot". Un mot ou "k-tuple" est une suite ordonnée de symboles. La démarche consiste a fragmenter la séquence à analyser en mots chevauchants de longueur donnée, puis a dresser pour chaque mot la liste des positions ou il est rencontré ("lookup table" ou "hashing table").

Cette liste permet d'identifier pour chaque mot d'une séquence de la banque en quelles positions il se trouve sur la séquence à analyser, et donc quel est le déplacement à effectuer

pour les mettre en correspondance. La notion de mot implique la définition de similarités strictes

- Les séquences sélectionnées sont ensuite soumises à une analyse plus précise. FASTP par exemple retient par défaut les cinq meilleures diagonales au sens match/mismatch, puis recalcule (avec une matrice) le score des séquences réalignées et garde la meilleure score.

- FASTA est plus sensible que FASTP. FASTA détecte lui des alignements locaux. Il tient compte du fait qu'il peut y avoir un décalage léger entre les deux séquences, et le compte dans le score final des séquences de la banque.

L'évaluation statistique selon le format FASTA est effectuée grâce a un calcul des scores. On retient tous les scores de toutes les séquences de la banque, on calcule la moyenne (μs) et l'´ecart type (ss) de ces scores, et on calcule le z-score ((score−μs)/ss) pour les séquences de meilleur score (celle qu'on retient a la fin de la recherche). La valeur du z-score permet de juger de la signifiance des résultats obtenus (si le z-score est supérieur a 5 ou plus, la signifiance est assurée).

L'algorithme BLAST

Son avantage, par rapport à FASTA, qui recherche des coïncidences de doublets stricts, est de rechercher des coïncidences de mots de longueur W qui ne sont pas strictement identiques. Pour ce faire, BLAST calcule un dictionnaire de mots équivalents a ceux de la séquence requête, c'est a dire dont la distance d'´edition ne dépasse pas un certain seuil. Une fois ces mots rangés, BLAST parcours les séquences de la banque, et s'arrête a chaque mot d'une séquence de la banque ayant un correspondant dans la séquence. BLAST essaie d'étendre l'alignement local (sans insérer de gap) de part et d'autre des ces "amorces"

Alignements multiples: Alignement de n séquences simultanément

Alignements multiples de séquences

- Les méthodes précédentes requièrent un minimum de connaissances préalables sur les séquences. Lorsque d'une façon générale, on cherche à caractériser une famille de séquences partageant une même activité biologique, on ne sait rien de la nature, du nombre et de la position des motifs impliqués dans le(s) mécanisme(s) d'action, ou de reconnaissance, mis en jeu. L'approche la plus immédiate consiste à rechercher sur les séquences des régions conservées. L'alignement multiple fût l'une des premières réponses apportée à ce type de problème.

- L'algorithme de programmation dynamique (NWS) pour deux séquences est généralisable à l'alignement de N séquences.

- De même que dans le cas de 2 séquences, la programmation dynamique permet de calculer le meilleur chemin d'alignement dans une matrice carrée, pour N séquences, elle permet de trouver le meilleur chemin d'alignement dans un hyper-cube N-dimensionnel.

Un alignement multiple se présente sous cette forme :

Ici, cinq séquences nucléiques sont alignées. Chaque ligne représente une séquence alignée. Toutes les lignes comportent le même nombre de caractères ("A", "C", "G", "T" ou "."). Les numéros à droite de l'alignement indiquent les positions des derniers caractères dans leurs séquences respectives. La ligne *consensus* indique le degré de conservation de chaque colonne ("*" pour 3 caractères identiques, "!" pour tous les caractères identiques). La graphe *logo* représente le contenu en information (*"Information content"*) de chaque lettre à chaque position. Cette quantité est déduite de la proportion de chacune des lettres (plus une lette est caractéristique d'une position –sa proportion élevée, plus cette valeur est haute).

- L'accroissement du nombre de séquences dans les banques de données oblige le développement et l'utilisation de méthodes d'alignements multiples • Généralisation de la méthode d'alignement de 2 séquences à l'alignement de n séquences simultanément

Alignement multiple

L'alignement multiple permet de détecter les régions qui ont été conservés au travers de l'évolution, très souvent ces régions correspondent à des domaines associés à une fonction clé de la molécule. Contrairement aux algorithmes de programmation dynamique, les algorithmes heuristiques ne garantissent pas de trouver la solution exacte du problème (en l'occurrence, le meilleur alignement), mais devraient fournir une solution approchée raisonnable.

Principe : Une façon de résoudre le problème de l'alignement de N séquences consiste à utiliser une procédure séquentielle qui peut être décrite comme suit :

- 1. Alignement par programmation dynamique des deux premières séquences.
- 2. Alignement de la séquence n 3 avec l'alignement précédent.

- 3. Alignement de la séquence N avec l'alignement obtenu sur les N-1 premières séquences.
- La principale difficulté de ce type de méthode consiste toujours a définir l'ordre dans lequel les séquences doivent être prises en compte car l'alignement final en dépend.

Logiciel d'alignement multiple CLUSTAL

- La similarité de chaque séquence est évaluée par rapport à toutes les séquences. Un score de similitude est calculé pour chaque paire de séquences selon un alignement approximatif global rapide (seuls les fragments exactement appariés et les diagonales avec un grand nombre d'appariements sont pris en compte. On obtient ainsi une matrice de distances).
- Un arbre de guidage ("*guide tree*") est construit : il s'agit d'un arrangement traduisant les relations globales de parenté entre les séquences. Il indique l'ordre à partir duquel l'alignement multiple graduel sera établi.
- Le programme construit un premier alignement multiple (par programmation dynamique ou par une méthode semblable à celle de FASTA): les 2 séquences les plus similaires servent de base pour l'élaboration de cet alignement multiple primaire.
- On obtient une première séquence consensus qui est alignée avec la 3e séquence la plus similaire.
- Toutes les séquences (des plus proches aux plus distantes) sont ainsi progressivement ajoutées par construction de consensus successifs jusqu'à l'alignement multiple final.