République Algérienne démocratique et populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université A.MIRA-BEJAIA
Faculté des Sciences de la Nature et de la Vie
Département des Troncs Communs

Biologie moléculaire

(*Méthodes d'étude du génome*)

Dr. BOUREBABA Yasmina

Cours destiné aux étudiants en Licence et Master SNV

Préambule

Selon *Michel Morange*, la biologie moléculaire est "l'ensemble des techniques et découvertes permettant l'examen moléculaire des processus les plus singuliers du vivant, de ceux qui en assurent la pérennité et la reproduction". En effet, la biologie moléculaire ainsi que la génétique et la biochimie traitent l'un des domaines les plus captivants de la biologie : l'ADN, support de l'information génétique et du patrimoine de chaque organisme vivant.

Afin de mieux comprendre ce que la biologie moléculaire représente à l'heure actuelle, il est crucial de saisir l'origine, les fondements ainsi que le développement de celle-ci. Les techniques de manipulation des acides nucléiques en génétique moléculaire sont principalement basées sur des outils tels que les enzymes, l'électrophorèse, les cellules procaryotes ou eucaryotes, les vecteurs, l'hybridation via des puces à ADN ou autres,… offrant la possibilité d'étudier les processus biologiques complexes à l'échelle pangénomique.

Depuis la publication de la séquence complète du génome humain grâce au projet Génome Humain, en 2004, chaque année de grandes banques de séquences d'ADN sont générées. De puissants programmes bioinformatiques arrivent à comparer les génomes d'une grande variété d'organismes. Par conséquent, la génétique comparative favorise la compréhension de l'évolution du génome humain et celui de nombreux autres organismes modèles, ouvrant la voie à la recherche. Afin de favoriser le développement continu des nouvelles technologies, de nombreux projets sont constamment mis en œuvre, comme le génotypage à haut débit des SNP ou encore des micropuces à ADN fournissant des outils puissants et performant. Indéniablement, tout cela ouvre des perspectives prometteuses pour l'avenir de l'humanité.

Cet ouvrage offre aux lecteurs un soutien pédagogique et vise à leur faire découvrir les techniques les plus courantes appliquée à la biologie moléculaire et notamment la démarche scientifique élémentaires. La démarche pédagogique utilisée dans la conception de ce cours repose sur l'emploie d'un langage fondamental, accompagné d'exemples et d'un grand nombre d'illustrations rendant les concepts abordés plus accessible. Ce cours intitulé "*Méthodes d'étude du génome*" est divisé en trois chapitres principaux traitant successivement d'une description de l'ensemble des techniques d'extraction et de purification des acides nucléiques, des méthodes de quantification de l'ADN et ARN, étapes préalables à toute manipulation de ceux-ci, et des techniques de base pour la manipulation et l'étude du génome à savoir la digestion enzymatique, la réaction en chaine de polymérisation, l'hybridation moléculaire, le séquençage ainsi que la recherche et le traitement de données moléculaire.

« La richesse des acides nucléiques n'aura de cesse de nous surprendre »

Préambule

Table des Matières

Chapitre II. Quantification des acides nucléiques

Chapitre III. Manipulation et étude du génome : techniques de base

Liste des figures

Liste des tableaux

N° *Titre* *Pages*

Objectifs généraux du cours

A travers ce cours, il est voulu fournir aux apprenants des notions cohérentes sur l'historique, les techniques et méthodologie moléculaire ainsi que les démarches scientifiques d'analyse et de traitement de données moléculaire. La conception de ce cours a été faite de telle sorte à faciliter la tâche aux étudiants venant d'horizons différents, mais ayant eu, tout de même, au préalable des bases fondamentales en biologie moléculaire. De plus, afin de favoriser une meilleure compréhension de ce cours, il est indispensable que certaines notions scientifiques de bases aient été au préalable acquises.

Pré-requis

- Notions en biologie générale et en biologie cellulaire;
- Notions de Chimie/Biochimie élémentaire ;
- Notions en génétique et biologie moléculaire de base ;
- Notion en informatique.

A l'issue du cours, les apprenants devraient en théorie pouvoir répondre à la liste des objectifs suivants

- Connaître ce qu'est une extraction/purification d'acides nucléique ;
- Décrire les différentes techniques de base pour l'extraction et purification des acides nucléiques (ADN et ARN) ;
- Savoir quand et quel protocole d'extraction/purification appliquer en fonction de l'échantillon de base et de l'expérience en aval ;
- Comprendre quand et pourquoi réaliser une quantification des acides nucléiques ;
- Connaître les techniques d'étude du génome et comment les appliquer (digestion par enzymes de restrictions, amplification par PCR, hybridation moléculaire, séquençage,….) ;
- Etre capable de faire une recherche de données moléculaire et de les traiter ;
- Etre capable d'interpréter des résultats expérimentaux obtenus par les techniques les plus courantes de biologie moléculaire appliquées à la recherche ou à l'exploration de matériel génétique par les laboratoires de biologie.

Chapitre I

Techniques d'extraction, de purification et de conservation des acides nucléiques

Techniques d'extraction, de purification et de conservation des acides nucléiques

Objectifs spécifiques

Au terme de ce cours qui traite des différentes techniques d'extraction, de purification et de conservation des acides nucléiques, vous devez être capable de :

- Comprendre ce qu'est une extraction, une purification et la conservation des acides nucléiques ;
- Comprendre pourquoi faire une extraction, purification des acides nucléiques ;
- Connaitre quelles sont les différents procédés afin d'extraire et purifier le matériel génétique ;
- Distinguer les différences entre les procédés et quand les appliquer.

I. Introduction

Afin de pouvoir poursuivre l'étude et les analyses moléculaires via le génie génétique (Séquençage, PCR ou clonage,…), les acides nucléiques (ADN et ARN) doivent être isolés de tissus ou de cellules de quelque nature que ce soit ayant été récoltés sur terrain.

L'extraction et la purification des acides nucléiques sont les premières étapes dans la plupart des démarches entreprises dans les études de biologie moléculaire et dans toutes les techniques du génie génétique mais le rendement ainsi que la pureté des acides nucléiques sont deux éléments importants pour assurer l'efficacité et la fiabilité des analyses.

II. Extraction des acides nucléiques

L'**extraction des acides nucléiques** est une technique permettant d'isoler l'ADN ou l'ARN de cellules ou de tissus. Il existe différents protocoles pour extraire l'ADN/ARN, suivant sa nature :

- De l'ADN génomique (Issu du ou des chromosomes des cellules analysées) ;
- De l'ADN plasmidique (Provenant de plasmides portés le plus souvent par des cellules bactériennes comme *Escherichia coli*) ;
- De l'ARN (Issus de cellules procaryotes ou eucaryotes).

Tous ces protocoles suivent approximativement le même schéma ou principe :

- Lyse des cellules (Casser les membranes cellulaires et nucléaires) ;
- Elimination des protéines et des lipides ;
- Elimination des autres acides nucléiques (selon les cas ARN, ADN,…) ;
- Concentration de l'ADN/ARN par précipitation à l'alcool.

II.1. Echantillons d'extraction

L'extraction des acides nucléiques (ADN, ARN) s'effectue à partir de **matériels biologies variés** tels que :

- Les cellules (bactéries, levures, cellules sanguines,…) ;
- Les tissus (végétal, animal) ;
- La matière fraîche (feuilles de plantes,…) ;
- La matière sèche (phanères[1],…).

II.2. Principaux procédés d'extraction des acides nucléiques

Les méthodes d'extraction des acides nucléiques peuvent se classer en trois principales classes en fonction du principe auquel elles font appel :

- Les méthodes utilisant des solvants organiques ;
- Les méthodes utilisant des solvants non organiques ;
- Les méthodes basées sur l'utilisation de micro-colonnes de résines échangeuses d'ions (Kits commerciaux permettant des extractions rapides à l'aide de réactifs prêts à l'emploi).

II.3. Principe de chaque étape d'extraction/purification

L'extraction et la purification d'acides nucléiques d'un matériau biologique donnée requiert donc la lyse cellulaire, l'inactivation des nucléases cellulaires et la séparation de l'acide nucléique souhaité des débris cellulaires.

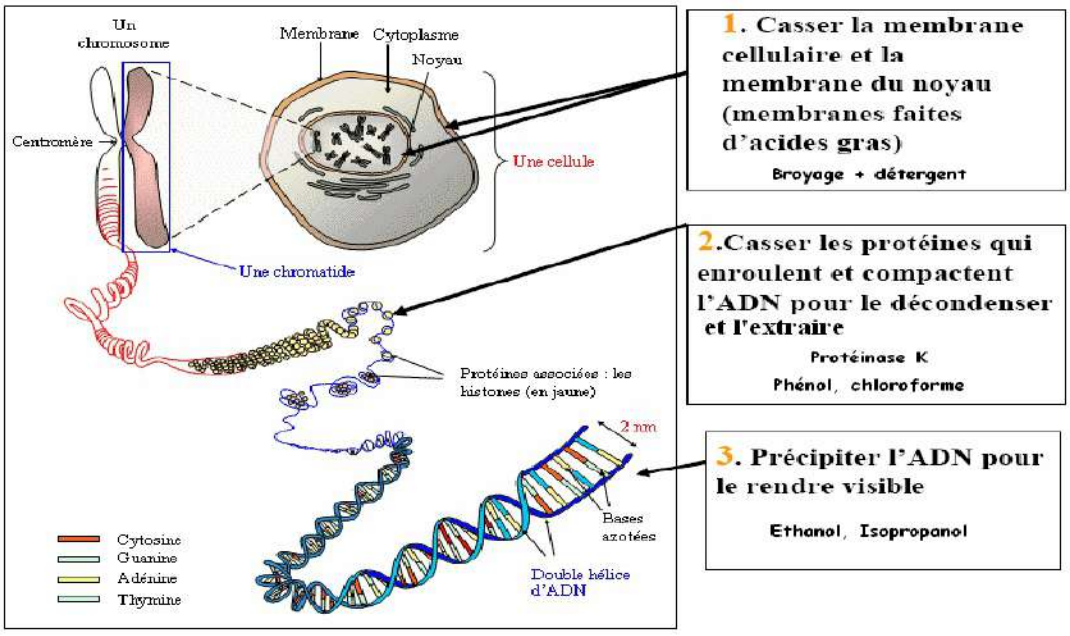


Figure 01 : Principales étapes d'extraction de l'ADN.

[1] Production épidermique apparente (cheveux, poils, plumes, écailles, griffes, ongles, dents).

II.3.1. Extraction de l'ADN génomique

a) Lyse cellulaire

La **lyse** cellulaire est un processus de destruction des membranes de cellules biologiques en utilisant des agents chimiques, physiques, ou biologiques, conduisant à la mort de la cellule. On obtient ce que l'on appelle **lysats**, produits résultant de cette désintégration.

Le procédé de lyse idéale est habituellement un compromis de plusieurs techniques et doit être suffisamment rigoureuse afin de briser la matière complexe de départ (telle que le tissu), mais assez douce pour protéger l'acide nucléique cible. Les procédés les plus courants de lyse sont les suivantes :

- **Rupture mécanique**

Pour la lyse mécanique, on utilise de préférence des **méthodes** qui ne dénaturent pas l'ADN, c'est-à-dire qui ne « **cassent** » pas les molécules d'ADN. Il est donc nécessaire que la méthode utilisée ne génère pas trop de forces de cisaillement[2] : généralement on utilise le choc osmotique ou le choc thermique (cycles de congélation/décongélation).

Les méthodes d'agitation ainsi que l'emploi des ultra-sons ont été délaissés du fait que ces méthodes ne sont pas adaptées à l'extraction d'ADN à partir de cellules procaryotes. En effet, les traitements mécaniques utilisés afin de détruire les cellules procaryotes doivent être considérablement plus drastiques que ceux employés pour les cellules eucaryotes (broyeur à bille, presse de French, …) cela est due au fait que les cellules de petite taille (bactéries,…) ont la capacité de s'insinuer entre les faisceaux de forces de cisaillement ce qui les rend donc parfois très difficiles à lyser mécaniquement. Les traitements mécaniques employés pour les procaryotes sont donc trop dénaturant pour l'ADN. La **lyse mécanique est de préférence réservée aux cellules eucaryotes**.

- **Choc thermique**

La fragilisation des parois cellulaires des échantillons peut être faite par chauffage. Cette méthode est fréquemment employée pour les spores et les champignons. Les cycles de congélation/décongélation (Par exemple, le passage par des cycles de $-20°C/-80°C$ à $37°C$) des échantillons peuvent également provoquer la rupture des membranes cellulaires.

- **Traitement chimique/enzymatique**

Cette méthode est préférée pour les cellules procaryotes que celle mécanique. La aussi on retrouve plusieurs procédés :

[2] C'est le rapport d'une force à une surface. Elle possède la dimension d'une pression.

↳ ***Fragilisation enzymatique des parois de cellules bactériennes, végétales et fongiques***

L'obstacle majeur à la lyse cellulaire est la présence de la paroi. Des hydrolases spécifiques peuvent être employées afin de rendre efficace la désorganisation de la membrane plasmique :

Organisme	Bactérie	Champignon	Végétaux
Type de paroi	Peptidoglycanne	Chitine	Cellulose Hémicellulose Pectines
Enzyme	Lysozyme	Chitinase	Cellulase, Hémicellulase, Pectinase

Ce traitement crée des brèches dans la paroi. Lorsque la protection assurée par la paroi contre la forte pression osmotique intracellulaire est perdue, la cellule gonfle (l'eau afflux vers l'intérieur de la cellule) jusqu'à ce que la membrane plasmique soit totalement rompue.

↳ ***Désorganisation des membranes par des détergents et solubilisation des lipides membranaires***

Le Sodium Dodécyl Sulfate (SDS) appelé aussi laurylsulfate de sodium, le triton X100, le sarcosyl sont des détergents qui vont solubiliser les lipides membranaires sous forme de micelles. Cela aide à créer au niveau membranaire des pores suffisamment larges afin de libérer le contenu du cytoplasme hors des cellules.

En fonction de leur force (qu'ils soient chargés ou non), les détergents dénatureront plus ou moins les protéines membranaires. Cette dénaturation protéique de la membrane plasmique contribue tout autant à la lyse de la cellule.

On commence en général par un broyage, suivi d'une extraction par des détergents, qui vont casser les membranes cellulaires et nucléaires et dénaturer les protéines. La solution obtenue est en général très visqueuse, car l'ADN ainsi libéré forme de très longs filaments.

Remarque

L'inactivation des nucléases intracellulaires ainsi que la rupture de la membrane peuvent être combinées. À titre d'exemple, une solution simple peut contenir des détergents pour solubiliser les membranes cellulaires et des sels chaotropiques[3] puissants pour inactiver les enzymes intracellulaires.

b) Elimination des protéines

Cette étape sera très importante pour l'**élimination des histones** mais aussi les **protéines membranaires**. On peut procéder de deux façons différentes.

[3] Agent chaotropique, est une molécule qui détruit la structure spatiale (tridimensionnelle) des macromolécules biologiques telles que les protéines et les acides nucléiques en les dénaturants. Les chaotropes interfèrent avec les interactions intramoléculaires faibles (non covalentes), comme les liaisons hydrogène.

- **Déprotéinisation par hydrolyse enzymatique**

Elle est réalisée en faisant agir une **endoprotéase non spécifique** comme la **protéinase K**, active jusqu'à 65°C, c'est donc une enzyme très stable. Cette digestion est généralement réalisée en présence de **détergents** dénaturants tels que le **SDS**. En plus de contribuer à la lyse de la cellule par la solubilisation des lipides, le SDS soutient également l'action de la protéinase K en déployant sa chaîne protéique.

La **protéinase K** est une enzyme de la famille des protéases très active. Elle est communément utilisée en biologie moléculaire pour digérer les protéines et enlever les contaminants de préparation d'acides nucléiques sans effet hydrolytique sur les acides nucléiques.

Elle permet d'inactiver les ADNases et les ARNases et tout en ayant la capacité de modifier les protéines cellulaires de surface. La protéinase K est une endopeptidase qui clive de préférence les liaisons peptidiques au niveau du groupement carboxyle d'un acide aminé à chaîne latérale hydrophobe ou aromatique. Son activité est stimulée par agents dénaturants comme le SDS ou l'EDTA.

- **Précipitation des protéines en utilisant un agent chaotropique**

La précipitation va se faire grâce à des agents chaotropiques qui sont des **sels** (donc des ions) **modifiant la solubilité** des molécules (protéines ou acides nucléiques) et qui peuvent provoquer leur **précipitation**. Certains sont dénaturants d'autres pas. L'agent chaotropique peut agir de plusieurs façons :

- Neutraliser une partie des charges ioniques requises en surface pour maintenir la solubilité.
- Interférer dans les interactions que les protéines établissent avec l'eau et modifier la solubilité des protéines. Pour établir des interactions avec les molécules d'eau disponibles dans la solution, l'agent chaotropique entre en compétition avec la protéine vis-à-vis des molécules d'eau qui l'hydrataient (la solvatait), conduisant à la précipitation de la protéine puisqu'elle sort alors de la solution.
- Pour certains, ils peuvent aussi **dénaturer les protéines**, on les qualifie alors de **dénaturants-précipitants**. Pour les protéines par exemple, ils induisent la rupture des liaisons hydrogènes qui maintiennent leur structure tertiaire ce qui conduit ainsi à l'exposition de leurs régions hydrophobes. La tendance à s'agréger des régions hydrophobes conduit à la précipitation des protéines.

On peut citer comme agent chaotropiques :

→ ***Des agents chaotropiques ne dénaturant pas les protéines :*** le chlorure de sodium (NaCl) à forte concentration et le sulfate d'ammonium ($(NH_4)_2SO_4$), amplement adopté dans la purification des protéines ;

→ ***Des agents chaotropiques dénaturant pour les protéines***, le perchlorate de sodium ($NaClO_4$), l'iodure de sodium (NaI), le thiocyanate de guanidine (TCG) et le chlorure de lithium.

Le thiocyanate de guanidine (GTC) ainsi que l'iodure de sodium (NaI) sont les agents chaotropiques essentiellement les plus utilisés. En effet, par leur action dénaturante sur les protéines, ces agents chaotropiques assurent deux fonctions à la fois disloquer les membranes des tissus en agissant sur les protéines membranaires en conduisant ainsi à la lyse des cellules et éliminer les protéines en provoquant leur précipitation.

Après centrifugation, peu de protéines subsistent au niveau du surnageant contenant, lui, la totalité des acides nucléiques.

c) Elimination des ARN lors de l'extraction de l'ADN

Les deux types d'acides nucléiques (ADN et ARN) sont contenus dans les extraits acellulaires bruts. Deux manières peuvent être utilisés afin d'hydrolyser les ARN contenus dans les extraits :

- **Hydrolyse chimique** avec du NaOH. A pH alcalin, l'ADN est protégé de la lyse alcaline contrairement à l'ARN, qui lui, est hydrolysé due à la présence de groupement hydroxyle en C2' sur le ribose mais pas sur le désoxyribose.
- Mais on préfère souvent opérer par **digestion enzymatique**. Afin de réduire la concentration en ARN par digestion à la RNase, une **RNase « *DNase free* »** ; à savoir dépourvue d'activité DNase ; est employée. A cet effet, les préparations de RNase du commerce sont traitées par chauffage (par exemple 5 min à 100°C), un chauffage auquel résiste la **RNase**, mais pas les éventuelles DNases contaminant. Cette enzyme (RNase) particulièrement thermostable, est habituellement ajoutée dès le début de l'extraction-purification due à son importante stabilité.

II.3.2. Extraction des ADN plasmidiques

Les plasmides sont de petites molécules d'ADN habituellement circulaire existant indépendamment des chromosomes de l'hôte. Ils sont présents chez de nombreuses bactéries, quelques levures et mycètes dont la réplication est autonome. Ils portent un nombre de gènes très réduit ($\leq$30), en nombre assez variable (Plasmide à copie unique = 1 seul/cellule hôte, et

plasmides à copies multiples = 40 ou plus/cellule hôte) et dont l'information génétique n'est pas indispensable à l'hôte.

1) Minipréparation de plasmides

La préparation d'ADN plasmidique à partir de bactéries est l'une des techniques les plus courantes de la biologie moléculaire, également connue sous le nom abrégé de ***miniprep, midiprep*** ou ***maxiprep*** en fonction du volume de la culture bactérienne utilisée lors de l'extraction. Le principe de l'extraction est connu sous la désignation de **lyse alcaline**. Cette méthode permet de préparer sélectivement l'ADN plasmidique contenu dans les bactéries, tout en éliminant l'ADN du chromosome bactérien.

Le principe de cette méthode consiste à effectuer la lyse des cellules au moyen d'un **détergent**, le SDS, en présence de **soude**, à pH 13. A ce pH très alcalin, l'ADN est dénaturé, c'est-à-dire que les deux brins de la double-hélice sont séparés. En effet, les forces de répulsion entre les deux brins sont à leur maximum due aux groupements phosphate intégralement chargés négativement à pH basique.

On neutralise ensuite rapidement la solution, ce qui provoque la renaturation brutale (réappariement des brins du duplex d'ADN). L'ADN chromosomique, très long (quelques Méga (106) paires de base), ne parvient pas à se rappareiller complètement et forme des enchevêtrements insolubles.

L'ADN plasmidique, court (~10^3 paires de base), parvient à se reformer et reste en solution. On sépare alors les espèces par centrifugation. Les protéines précipitées, sont également éliminées avec le détergent et l'ADN chromosomique. Un lysat clair est alors formé, celui-ci est une solution ne contenant que très peu d'ADN chromosomique et de protéines. L'ADN plasmidique est ensuite concentré par précipitation à l'alcool.

Remarque

A côté de cela, des kits commerciaux avec de petites colonnes de résine chromatographique échangeuse d'ion sont disponibles et améliore la pureté de l'ADN.

II.3.3. Extraction des ARNs

Les ARNs sont plus difficiles à étudier que les ADN parce qu'ils sont très vulnérables/sensible vis-à-vis de la ribonucléase (RNase A) donc son extraction est plus délicate que celle de l'ADN.

Celle-ci, RNase, est ubiquitaire (les doigts par exemple en sont couverts), extrêmement active et très résistantes/stable à toutes les agressions habituellement néfastes pour toutes les

enzymes ; par exemple un traitement à 90°C pendant une heure n'altère pas son activité et même si elles sont présentes en petite quantité, réaliseront la dégradation des ARN.

Par conséquence, toute recherche concernant les ARNs doit être exécuté dans des conditions strictes de stérilité en raison de la présence des RNases microbiennes. Tous les milieux (eau, tampons,…) ainsi que le matériel doivent être préalablement autoclavés ; les pièces ne pouvant subir un tel traitement (pièces en plastique par exemple) doivent être lavées à l'acide iodo-acétique 10 mM puis rincées à l'eau distillée autoclavée et le port de gants est indispensable.

1) Principe

Tous les différents protocoles utilisés pour extraire l'ARN ont le même schéma et principe que pour l'extraction de l'ADN : Lyse cellulaire mécanique (congélation, billes de céramique ou détergents) ou enzymatique (lysozyme ou la lyticase = Endoglucanase et Protease).

La méthode d'extraction la plus sûre est celle de Chirgwin *et al.* en 1979, qui homogénéisait les prélèvements dans une solution à 4 M de thiocyanate de guanidinium, agent puissant de dénaturation des protéines (agent chaotropique qui dénature les RNases endogènes), additionné de β-2-mercaptoéthanol ($HOCH_2CH_2SH$, agent réducteur) qui permet de rompre les ponts disulfures protéiques (inhibiteur de ribonuléases) en plus du détergent (SDS).

Cette méthode a, par la suite, été amélioré par Chomczynski et Sacchi afin d'accélérer le process d'extraction en combinant à la phase d'extraction au phénol-chloroforme, lyse au thiocyanate de guanidinium. Lors de la réalisation de grandes séries d'échantillons ou lorsque l'on ne dispose que de petites quantités cellulaires ou tissulaires, cette modification s'est avérée significativement intéressante.

Une perte ou une fragmentation de l'ARN peut être obtenus lors de l'utilisation de solvant organique. De même, des résidus salins peuvent interférer sur les résultats en modifiant la vitesse et l'efficacité des réactions enzymatiques ultérieures.

Une fois les cellules convenablement lysées, les ARNs, peuvent être extraits. L'extraction des ARNs peut être réalisée selon deux grands principes :

- Les différences de propriétés physico-chimiques entre les différents acides nucléiques et les protéines ;
- L'adsorption sélective des acides nucléiques sur un support solide.

III. Purification des acides nucléiques extrait

Afin de pouvoir procéder aux études fonctionnelles en biologie, la purification des acides nucléiques s'avère être une étape clé du processus. En raison de la grande diversité des organismes vivants dans la nature, la purification devient un réel défi pour les chercheurs.

La séparation de l'ADN ou de l'ARN des autres composants du lysat représente l'**étape de purification**. La séparation doit être spécifique pour éliminer tous les composants du lysat et ne conserver que l'acide nucléique souhaité.

III.1. Purification par extraction au phénol-chloroforme

La méthode utilise la solubilité différentielle des molécules (acides nucléiques / contaminants comme les protéines et les lipides) entre deux phases non miscibles.

En pratique, on mélange vigoureusement la solution d'acides nucléiques en phase aqueuse à une phase non miscible hydrophobe. Après centrifugation, la phase aqueuse contenant les acides nucléiques (phase aqueuse = phase supérieure) est préalablement récupérée à la pipette.

III.1.1. Deux extractions successives

Les deux extractions successives qui se distinguent selon la phase non miscible sont :

- ***L'extraction phénolique***

Elle est utilisée pour débarrasser les acides nucléiques des protéines car le **phénol** est un **déprotéinisant** puissant. Les protéines **précipitent**, et sédimentent au fond de la phase aqueuse mais elles **restent à l'interface** c'est-à-dire qu'elles restent à la surface de la phase phénolique qui est une phase hydrophobe. Les différents débris lipidiques membranaire entreront dans la phase phénol hydrophobe.

Afin d'extraire l'ADN, le phénol doit être très pur et saturé en tampon (pH 8).

Le phénol est un produit à **manipuler avec précaution** car son inconvénient réside dans son **caractère très corrosif**.

- ***L'extraction au chloroforme***

Elle complète toujours l'extraction précédente pour éliminer toutes traces de phénol aqueux. Pour permettre l'action ultérieure d'enzymes (de restriction par exemple) sur l'acide nucléique extrait, toute trace de phénol doit être préalablement éliminée. L'**alcool isoamylique** (AIA = 3-méthyl-1-butanol = $(CH_3)_2CHCH_2CH_2OH$), **agent antimousse**, est habituellement additionné au chloroforme afin de **stabiliser** la séparation des phases (agent déstabilisant de l'émulsion).

Remarque

Il est à noter que l'élimination de l'ARN peut être effectué à la fin ou avant l'étape phénolique lors de la préparation d'un extrait d'ADN. S'il est exécuté après l'étape phénolique, un deuxième traitement de déprotéinisation doit suivre le précédent traitement afin d'éliminer la RNase résiduelle.

L'extraction phénol-chloroforme appliquée à la purification de l'ARN = extraction phénol acide-chloroforme.

III.1.2. Purification au phénol acide

Le phénol utilisé est un **phénol acide** c'est-à-dire qu'il a été mis en solution et équilibré avec un tampon de pH acide (pH 5). Dans ces conditions, les protéines histones riches en acides aminés basiques (portant dans leur radical un groupement NH_2), vont avoir à pH acide une forte charge positive (NH_3^+), elles vont alors étroitement s'associer à l'ADN génomique, qui lui, est chargé négativement ce qui va entraîner sa précipitation en même temps que les protéines.

Ainsi, à la surface de la phase phénolique subsiste l'ADN et les protéines ayant précipité et sont par conséquent rassemblés à l'interface. La phase aqueuse (supérieure) comprendra les ARN en solution débarrassés de l'ADN. La phase phénolique (inférieure) contiendra les lipides.

Remarque

D'autres techniques rapides ont été élaborées sans l'utilisation du phénol acide-chloroforme : purification sur mini-colonne contenant une membrane de gel de silice présentant une affinité pour les ARN avec élimination des ADN par l'emploi d'une DNAse.

III.2. Purification par centrifugation sur chlorure de césium

La centrifugation est une technique utilisant la force centrifuge, mouvement de rotation très rapide, pour séparer par exemple un solide en suspension dans un liquide ou deux phases de densité différentes. Il est à noter que l'une des méthodes de purification les plus robuste est la centrifugation sélective.

III.2.1. Séparation ADN/ARN par centrifugation isopycnique sur gradient de chlorure de césium

Une solution de chlorure de césium (ClCs) de densité donnée soumise à une **force de gravité intense** forme spontanément un **gradient de densité continu**. La résolution de tels gradients atteint le centième d'unité de densité.

Au cours de la migration, les molécules additionnées à la solution de ClCs vont se positionner au niveau de la zone du gradient, dont la densité est semblable à la leur. D'où la désignation « isopycnique » du grec *iso* (égal) et *puknos* (dense).

Cette technique permet de distinguer l'ADN natif c'est-à-dire double brin de l'ADN dénaturé (simple brin) et de l'ARN (lui aussi monocaténaire). En effet, il y a des liaisons possibles entre des groupements ionisés de l'intérieur de la molécule d'ADN et le chlorure de césium qui est lui aussi chargé (Cs+). Il y a d'avantage de liaisons qui s'établissent quand l'ADN est dénaturé ou avec l'ARN car il y a d'avantage de groupements ionisés libres pour fixer les ions Cs+. Par conséquent, l'**ADN dénaturé** ou l'**ARN** seront **plus denses** que l'ADN natif en présence de chlorure de césium.

La présence du ClCs provoque ainsi la fixation différentielle d'ions Cs+ sur ces macromolécules et leur densité apparente est alors modifiée : protéine = 1,3 g/ml, ARN = 1,75 −1,89 g/ml et ADN = 1,6−1,79 g/ml. Ces macromolécules seront ainsi facilement séparées durant la centrifugation.

Ce type d'ultracentrifugation est très résolutif mais d'un maniement délicat.

III.2.2. Purification de l'ARN par centrifugation sur coussin de chlorure de césium

L'extraction d'ARN est un procédé beaucoup plus difficile du fait que les molécules de **ribonucléotides**, soient aisément **dégradées** due à leur **très grande vulnérabilités** et **très grande sensible** aux **ribonucléases** (RNAses), **très stables**. Cela impose de prendre des précautions particulières aussi bien dans le recueil et la conservation des échantillons que dans la mise en œuvre des procédures d'extraction.

Les tissus ou les cellules, frais ou congelés dans l'azote liquide, sont homogénéisés le plus vite possible dans un tampon qui doit **très fortement inhiber** les RNases endogènes tout en **dissociant** les protéines des acides nucléiques. Le tampon contient en règle générale :

- Un **détergent** (SDS) ;
- Un agent dissociant dit encore **agent chaotropique**, le thiocyanate de guanidine, qui précipite les protéines et inhibe en même temps les RNases ;
- Un **agent réducteur**, le dithiothréitol ou le 2-mercaptoéthanol, des inhibiteurs de RNases ou d'enzymes en général puisqu'ils hydrolysent les ponts disulfures présents au niveau des protéines provoquant ainsi leur dénaturation (ponts disulfures = liaison covalente établie entre deux acides aminés soufrés cystéine ou méthionine : S-S devient SH + SH sous l'action de l'agent réducteur).

L'**ARN monocaténaire** lie d'avantage de molécules de chlorure de césium que l'**ADN génomique bicaténaire,** l'ARN présente alors une densité supérieure. Il n'y a que l'ARN qui

peut traverser les coussins de CsCl et être récupéré dans le culot, cela est notamment dû à sa densité. L'ARN est ensuite lavé dans de l'acétate de sodium et précipité à l'alcool.

III.3. Purification par chromatographie

Les deux techniques, **chromatographie** sur **colonne d'échange d'anions** et **chromatographie** sur **colonne de silice**, sont les plus employées pour purifier les acides nucléiques.

De nombreux **kits commerciaux** offrant différentes variantes et/ou améliorations des méthodes précédemment décrites contiennent couramment de **petites colonnes** de résine **chromatographique** échangeuse d'anions ou de silice, favorisant l'**amélioration** de la **pureté**. La chromatographie sur colonne met en œuvre 05 étapes :

- Equilibration de la colonne → travail en tampon et en force ionique adéquats ;
- Fixation de la molécule à purifier ;
- Lavage de la colonne → on élimine ce qui n'est pas fixé ;
- Elution → on décroche ce qui a été fixé ;
- Régénération → retour à l'état de départ de la colonne ou tampon de conservation.

III.3.1. Chromatographie d'adsorption sur colonne de silice

Cette technique est actuellement amplement employée et de nombreux fournisseurs proposent ces kits d'extraction/purification fondés sur ce principe, adaptés à de nombreux échantillons.

- **Le support**

Composé chimique et un minéral, la silice (dioxyde de silicium) est de formule SiO_2. Préparé à partir de silicate de sodium, le **gel de silice** est un **polymère** d'acide silicique $Si(OH)_4$. Elle est soluble dans l'eau sous la forme de $Si(OH)_4$, l'acide silicique (acide faible), a une limite de solubilité à 25°C de 0,140 g/l. Surface du gel de silice sec et hydraté.

La silice peut être proposée sous forme de billes ou bien de gel sur colonne :

- Dans le cas d'une **colonne**, la séparation est réalisée par **centrifugation**. Ce procédé est basé sur l'emploi de petites colonnes adaptées à des tubes de 1,5 à 2 ml (tube eppendorff).
- S'il s'agit de **billes**, la séparation sera effectuée à l'aide d'un **dispositif aimanté** via l'utilisation de billes **paramagnétiques** recouvertes de silice.

- **Les interactions**

Il s'agit d'une **chromatographie par adsorption**. La liaison de l'ADN à la silice implique des interactions dipôle-dipôle (interaction hydrogène).

- **Les étapes de la chromatographie**

Les acides nucléiques peuvent être **adsorbés sélectivement** sur membrane de gel de silice uniquement lorsque l'on travaille dans les conditions suivantes : à savoir en présence de **concentrations élevées** d'agents chaotropiques ou par exemple à force ionique élevée (NaCl élevée) et en présence d'**alcool**.

L'agent chaotropique en tant que concurrent interfère avec les relations que l'ADN établit avec l'eau. L'agent chaotropique se lie à l'eau et l'ADN devient moins hydraté, conduisant l'ADN à établir des liaisons avec la silice ce qui fait qu'il se fixe au support de la colonne.

Un lavage de la colonne à l'éthanol est suivie, ce dernier est un solvant moins polaire que l'eau ce qui évite le décrochement l'ADN.

Un **tampon** à **basse force ionique** va être utilisé afin d'éluer l'ADN. On peut utiliser un tampon avec une faible concentration en sels ou de l'eau pure, l'ADN va à nouveau établir des liaisons avec l'eau, s'hydrater et repasser en solution.

Remarque
1 et 2 → sont réalisées à **force ionique élevée** : on neutralise les groupements phosphate PO_4^- de l'ADN avec Na^+ (il devient moins soluble).
3 → la colonne est lavée en présence d'**éthanol** qui est un solvant moins polaire que l'eau (l'ADN ne repasse pas en solution).
4 et 5 → on utilise de l'eau ou un tampon de **force ionique très faible** (exemple avec de l'EDTA).

- **Avantages**

- Pas d'utilisation de substances toxiques si le NaCl est employé.
- Aucune précipitation alcoolique n'est requise, puisque l'ADN obtenu est de bonne qualité et de ce fait prêt à l'emploi.
- La méthode est rapide et peut être appliquée à divers spécimens.
- La taille de l'ADN obtenu est de l'ordre de 20-30 kb, pouvant également aller jusqu'à 50 kb.

- **Inconvénient**

Il est impossible d'obtenir des fragments de plus de 50 kb de long nécessaire aux expériences de clonage ou de *blotting* avec cette technique.

III.3.2. Chromatographie sur colonne échangeuse d'anions

- **Le support**

→ Un échangeur d'anions de type **DEAE cellulose** ou équivalents. La diéthylaminoéthyl-cellulose (DEAE-cellulose) est un support échangeur d'anions, préparé en substituant la cellulose (polymère de molécules de glucose liées en β1− 4) par des groupements diéthylaminoéthyls.

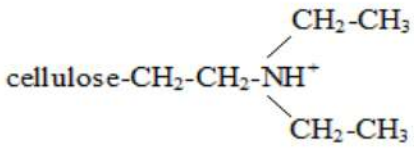


Figure 02 : Molécule de diéthylaminoéthyl-cellulose (DEAE-cellulose).

Le DEAE est un échangeur faible, pourvu de groupements issus d'un acide faible ou d'une base, de telle sorte que le support soit sous forme ionisée ou neutre.

→ Un échangeur d'anions de **type Q-sepharose** qui comporte l'ion **ammonium quaternaire R_4N^+** (désigné par la lettre Q) = un atome d'azote portant une charge positive et dont les 4 valences sont occupées chacune par un radical R. L'ammonium quaternaire est un échangeur fort issu d'une base ou d'un acide fort qui est donc tout le temps ionisé.

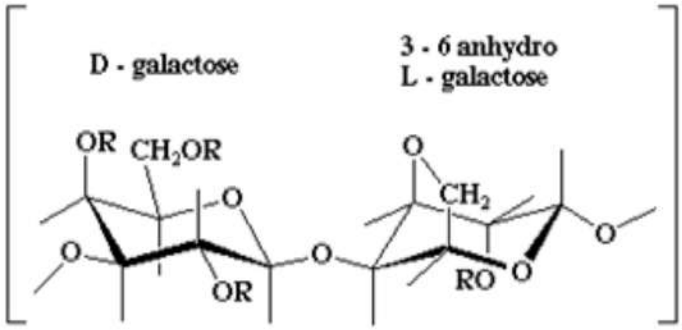


Figure 03 : Motif de base de l'agarose.

Le gel "Sepharose™" est constitué de chaînes d'agarose. Chaque chaîne, de masse molaire d'environ 120.000 Da, est un polymère d'un diholoside dont le motif est présenté ci-contre.

- **Les interactions**

La chromatographie échangeuse d'anions repose sur l'**interaction** entre les **charges positives** du **support de chromatographie** et les **charges négatives** des **groupements phosphate** de l'ADN. La rétention de l'ADN fait intervenir des **interactions ioniques** et cela après échange de l'ion (avec Cl^- par exemple).

- **Etapes de la chromatographie**

Le lysat doit être déposé sur la colonne de chromatographie dans des conditions de **concentration en sels basse**, afin que l'ADN puisse exposer ces charges négatives de telle sorte que les ions positifs contenus dans le tampon ne puissent pas neutraliser les groupements phosphate.

La **fixation** de l'ADN à la colonne se fait grâce aux **liaisons ioniques**. Ensuite, les protéines, les débris cellulaires et les autres molécules en solution sont éliminés par lavage à l'aide d'un tampon moyennement chargé en sels. Finalement, à l'aide d'un **tampon** à **haute force ionique** (solution concentrée de NaCl), l'ADN est **élué**. L'ADN est précipité par de l'alcool afin d'éliminer les sels.

Remarque
1, 2 et 3 → sont réalisées à force ionique **basse** ou **moyenne**.
4 et 5 → sont réalisées à force ionique **élevée** (élution par compétition avec le NaCl).

- **Avantages**

- Aucune substance toxique n'est utilisée : intérêt pour une préparation d'ADN en vue d'une thérapie génique par exemple.
- L'ADN isolé est d'une taille de 50-100 kb en moyenne et peut atteindre 150 kb.

III.3.3. Purification des ARN messagers eucaryotes par chromatographie d'affinité

Un **ligand spécifique** est mis en œuvre lors d'une chromatographie d'affinité. Il est connu qu'une **queue poly(A)** en **3'** d'environ une centaine de nucléotides est présente chez la plupart des ARNm eucaryotes. Cette structure particulière est mise à profit pour les purifier par une **chromatographie d'affinité** où il se fixe sélectivement par complémentarité de séquence sur des colonnes d'oligo (dT).

L'ARN poly(A) est purifiée via des passage sur une **colonne oligo(dT)** : oligo(dT) cellulose ou oligo(dT) fixé sur des billes magnétiques.

Cette approche permet ainsi d'augmenter la sensibilité de la méthode en éliminant les ARN ribosomaux et les ARN de transfert les plus abondants, notamment pour les transcrits faiblement exprimés. Elle est notamment utilisée pour :

- La réalisation de *northern blot* lorsque les ARNm sont présents en faible quantité ;
- La réalisation de la technique de *RNA mapping* à l'aide de la méthode d'extension d'amorce dans le but de déterminer la position du site d'initiation de la transcription ;
- La construction de banque d'ADNc, pour l'étude de l'expression des gènes à l'aide de puces d'ADNc ;

- La réalisation de traduction *in vitro*.

Deux approches peuvent être utilisées : soit les ARNm sont purifiés à partir des ARN totaux préalablement extraits, soit directement à partir du prélèvement source sans passage intermédiaire sous forme d'ARN total.

IV. Conservation des acides nucléiques

Une fois l'ADN ou l'ARN extrait et purifié, il est donc prêt à être utilisé pour des analyses moléculaires. Cependant, ces analyses prennent du temps et ne peuvent se faire en une seule fois (jour), nécessitant souvent des répétitions. De ce fait, cet acide nucléique doit être conservé dans des conditions optimales afin de pouvoir garder son intégrité et pour une utilisation ultérieure.

Le temps de conservation des acides nucléiques (ADN, ARN) dépend du choix du milieu (tampon) et de la température de conservation.

IV.1. Conservation de l'ADN

L'ADN peut être conservé dans un **tampon (10mM Tris, pH=8)** additionné d'**EDTA (1mM)** à **4°C**.

- A pH 8, la dégradation de l'ADN est notablement plus faible qu'à pH 7.
- L'EDTA permet de chélater les ions divalents (nécessaires à piéger pour inactiver les nucléases) et ils évitent aussi la croissance de microorganismes.

L'ADN peut être également conservé à **−20°C** dans le **même tampon** mais des cycles successifs de congélation/décongélation entraînent des cassures des acides nucléiques de grande taille (>10kb). Par conséquent, afin de procéder à une bonne conservation il est nécessaire de réaliser des **fractions aliquotes**.

IV.2. Conservation de l'ARN

Du fait de la très faible stabilité des ARNs, les échantillons sont stockés à**−80°C** après addition de **trois volumes d'éthanol**. En l'absence de sels, ils restent en solution. Après agitation, une aliquote représentative peut être prélevée puis ajustée à 0,3M d'acétate de sodium pH 5.2, ce qui provoque la précipitation de l'ARN brièvement à−20°C et recueillie par centrifugation.

Chapitre II

Quantification des acides nucléiques

Quantification des acides nucléiques

Objectifs spécifiques

Au terme de ce cours qui traite des différentes techniques de quantification des acides nucléiques, vous devez être capable de :

- Comprendre quand et pourquoi réaliser une quantification des acides nucléiques ;
- Connaitre les différentes techniques pouvant être utilisées ;
- Connaitre les avantages et les inconvenants des différentes méthodes de quantification ;
- Distinguer les différences entre les procédés et quand les appliquer.

I. Définition

En biologie moléculaire, il est indispensable de quantifier et d'analyser la pureté des acides nucléiques après leur extraction/purification. Le dosage des acides nucléiques est une procédure systématique permettant d'estimer approximativement la quantité ainsi que la qualité de l'ADN ou de l'ARN après son extraction et purification à partir d'un matériel biologique.

Afin d'éviter l'incohérence et la variabilité dans les projets de recherche, un haut niveau d'intégrité moléculaire est essentiel. La qualité des acides nucléiques est essentielle pour plusieurs techniques utilisées à des fins d'analyses génomiques. Des procédures de tests doivent être misent en place pour contrôler et évaluer la qualité des échantillons avant toute manipulation ultérieure.

II. Rappels

Les ondes électromagnétiques visibles ont une longueur d'onde qui s'étant de 400 à 800 nm alors que les ondes UV s'étendent de 200 à 400 nm. Ces rayons électromagnétiques transportent une énergie 'E' exprimée par la relation suivante : $\mathbf{E = h.\upsilon = h.c / \lambda}$

Ces énergies concordent avec les énergies de transition électronique de molécules ⇨ à T° ambiante, la majorité des particules sont dans leur état de vibration fondamental et leur état électronique.

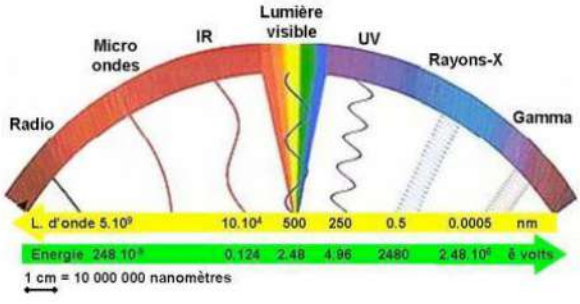


Figure 04 : Représentation des ondes électromagnétiques.

III. Techniques de quantification des acides nucléiques

Le contrôle de la qualité ainsi que de la quantité des acides nucléiques (ADN, ARN) est fondamental pour le succès des opérations d'étude de ceux-ci et donc à des fins de recherche. Pour ce faire, plusieurs techniques ont été misent en place et peuvent être utilisé afin de procéder au control. Ces outils ont leurs propres spécificités et sont parfois complémentaires.

III.1. Absorptiométrie UV

La spectrophotométrie est la méthode la plus répandue pour le dosage d'acides nucléiques en mesurant l'absorbance (ou densité optique) de ces acides nucléiques à 260 nm (absorbent dans l'ultraviolet). Conjointement, l'absorbance à 280 nm, 230 nm et 320 nm sont mesurées afin de déterminer leur pureté.

Cette méthode découle des caractéristiques physico-chimiques des acides nucléiques puisque les bases puriques et pyrimidiques absorbent fortement dans l'ultraviolet à 260 nm.

III.1.1. Correspondance $A_{260\ nm}$ et concentration en acides nucléiques

La mesure de l'absorbance à 260 nm est une méthode précise si l'ADN est pur et natif mais peu sensible, et ne peut pas être utilisée pour des concentrations inférieures à 250 ng/ml ($A_{260\ nm}$ = 0,005 UA).

La DO lue sur le spectrophotomètre à savoir une unité d'absorbance à 260 nm va correspondre à :

- Une solution d'ADN double brin à 50 µg/ml ;
- Une solution d'ADN simple brin à 37 µg/ml ;
- Une solution d'ARN à 40 µg/ml.

Il est à noter que ces valeurs s'appliquent à des acides nucléiques parfaitement pures et en solution homogène. Mais il est rarement nécessaire d'effectuer un dosage très précis et dans la pratique une simple estimation de la concentration suffit.

III.1.2. Rapport $A_{260\ nm}$ / $A_{280\ nm}$ et contrôle de la pureté d'une solution d'ADN

Les valeurs de ratio 260/230 et 260/280 définissent la pureté d'un échantillon d'ADN ou d'ARN.

Les protéines absorbent également à 260 nm mais leur maximum d'absorption se situe, lui, à 280 nm. Ainsi, le rapport $A_{260\ nm}$ / $A_{280\ nm}$ constitue un moyen d'apprécier une éventuelle contamination de la solution d'ADN/ARN, de ce fait :

- Un rapport compris entre 1,8 et 2 correspond à une solution pure d'ADN, ce ratio se situe autour de 2,0 (+/- 0,1) pour une extraction d'ARN ;
- Un rapport inférieur à 1,7 est le signe d'une contamination par des protéines ;
- Un rapport supérieur à 2 est le signe d'une contamination par l'ARN.
- Une éventuelle contamination par du phénol peut être recherchée en mesurant l'absorption à 270 nm (*rarement pratiqué*).

Remarque

A mettre en relation avec le fait que l'ARN et l'ADNsb absorbent plus à 260 nm que l'ADNdb (effet hyperchrome) et avec le fait que les acides aromatiques des protéines absorbent à 280 nm.

III.1.3. Ratio A_{260}/A_{230}

Le ratio A_{260}/A_{230}, généralement plus élevé que le ratio A_{260}/A_{280}, c'est un deuxième indicateur de pureté. La valeur attendue de ce ratio se situe entre 2,0 et 2,20. Si ce rapport est plus faible, celui-ci indique forcément qu'un composé absorbe à 230 nm. Ces composés peuvent le plus souvent être de l'EDTA, des sucres, ou encore une fois, le phénol et particulièrement le TRIzol.

Ces valeurs de ratios sont parfois négligeables, néanmoins elles indiquent la présence d'éventuels inhibiteurs qui pourraient perturber le devenir de vos acides nucléiques (PCR et autres réactions enzymatiques).

Il est primordial d'effectuer une valeur de blanc avec la même solution qui aura servi à éluer les acides nucléiques (comme précédemment mentionné, l'EDTA absorbe à 230 nm,… et bien souvent le TE 1X (Tris EDTA) sert à reprendre et conserver des acides nucléiques).

Tableau I : Valeurs des rations de pureté des échantillons d'ADN et d'ARN.

Pureté d'un échantillon d'ADN		
Ratio	**Valeur**	**Indication de pureté**
260/280	1.8	ADN pur
	< 1.8	Présence de protéines, phénol et autres contaminants
	> 1.8	Contamination à ARN
260/230	1.8-2.2	ADN pur
	< 1.8	Contaminants co-purifiés (solvants, sel, contaminants organiques)

Pureté d'un échantillon d'ARN		
Ratio	**Valeur**	**Indication de pureté**
260/280	2	ARN pur
	> 1.8	Présence de protéines, phénol et autres contaminants
260/230	1.8-2.2	ARN pur
	< 1.8	Contaminants co-purifiés (solvants, sel, contaminants organiques)

Les valeurs minimales acceptables sont :
$260/280 \geq 1,8$
$260/230 \geq 1,6$

III.1.4. Absorbance à 320 nm

La présence d'impuretés présentes sur la paroi de la cuvette ou à la surface de mesure (sur le piédestal) et/ou de particules dans la solution peut être indiqué via l'absorbance à 320 nm.

III.1.5. Facteurs influençant les ratios de pureté

Les facteurs comme le pH, la force ionique de la solution d'acide nucléique, la présence de bulles d'air, les fluorochromes (pour les oligonucléotides marqués) mais également la précision du spectrophotomètre impactent les ratios de pureté.

Selon les résultats publiés par Wilfinger en 1997, le ratio A_{260}/A_{280} d'un ARN pur dans une solution plus acide est significativement plus faible, pouvant induire en erreur l'interprétation de la pureté. Lors du dosage par le spectrophotomètre, il est nécessaire d'utiliser le même tampon pour le blanc et pour la solution d'acides nucléiques.

Un léger décalage de la précision au niveau de la longueur d'onde d'un spectrophotomètre peut modifier significativement les ratios A_{260}/A_{280}. Par exemple, un décalage de +/- 1 nm de précision de longueur d'onde se traduira par une variation de +/- 0,2 pour le ratio A_{260}/A_{280}. Il est possible d'avoir une dissimilitude jusqu'à 0,4 pour les ratios A_{260}/A_{280} d'un même échantillon d'acides nucléiques obtenus avec deux spectrophotomètres différents car beaucoup de spectrophotomètres réclament une précision de 1 nm. Il est ainsi indispensable d'inspecter de façon régulière la précision et la performance du spectrophotomètre utilisé.

III.2. Fluorimétrie

La fluorimétrie est une méthode de dosage utilisant la propriété de certaines molécules à être fluorescente. La fluorescence proportionnelle à la concentration est généralement mesurée.

La quantification par fluorescence des acides nucléiques est plus sensible que la spectrophotométrie et permet d'obtenir une sensibilité de l'ordre du pg/µl (picogramme/microlitre). Elle est plus spécifique en utilisant des fluorochromes spécifiques.

De plus cette méthode n'est pas impactée par la présence de contaminants dans la solution d'acide nucléiques tels que les ARNs ou les nucléotides libres (par rapport à une

mesure d'absorbance UV, qui ne discrimine pas les acides nucléiques simple brin, double brin, nucléotides libres,…).

Elle est particulièrement intéressante lorsque la solution contient à la fois de l'ADN et de l'ARN, car elle permet de quantifier uniquement l'ADN et non l'ARN. Mais contrairement à la spectrophotométrie, la fluorimétrie ne permet pas d'analyser la pureté et ni d'identifier les contaminants dans une solution donnée.

En plus de cela, cette méthode présente un autre inconvénient car elle est sensible à la composition en bases (le fluorochrome peut se fixer préférentiellement sur les ADN riches en A-T ou G-C). Le standard utilisé devra donc avoir une composition en GC proche de celle de l'ADN mesuré (les cellules eucaryotes ont une composition en GC de 39 à 46 %, en règle générale, et pour les cellules procaryotes de 26 à 77 %).

La fluorimétrie est très largement utilisée en **cytométrie de flux** pour quantifier l'ADN à l'intérieur de cellules individualisées.

III.2.1. Principe de la méthode

La fluorescence moléculaire est un processus d'émission de lumière par des molécules (généralement des polyaromatiques ou hétérocycles), en solution ou à l'état solide sans dégagement de chaleur, après excitation de celles-ci par des photons appartenant au domaine du visible ou proche de l'ultraviolet.

Lorsqu'une molécule fluorescente absorbe un photon d'énergie, les électrons vont passer de l'état fondamental E_0 à l'état excité E_1' (instable). Ils vont ensuite revenir à l'état vibrationnel le plus bas de E_1 par perte de microchaleur ⇨ C'est la relaxation. Ils vont finalement retourner à l'état fondamental E_0 avec émission d'un photon de fluorescence = énergie. L'énergie ainsi émise, est inférieure à l'énergie d'excitation. Le spectre d'émission est l'image du spectre d'absorption dans un miroir.

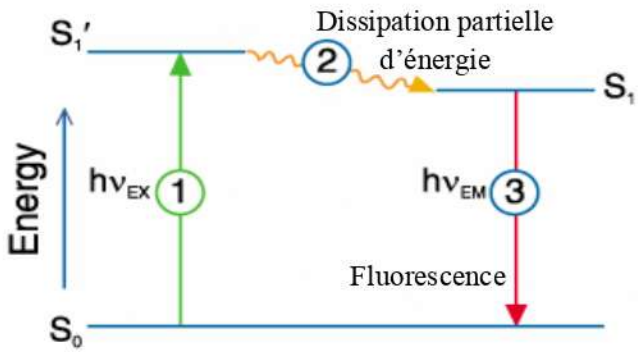


Figure 05 : Diagramme de Jablonski des états électroniques pour la fluorescence.

S_1' et S_1 sont des états singuliers excités ;
S_0 état fondamental ;
hv_{EX} c'est l'absorption d'un rayonnement électromagnétique ;
hv_{EM} c'est la libération de l'excès d'énergie sous forme de photons.

III.2.2. Espèces fluorescentes

La plupart des molécules ne sont pas fluorescentes, car leur structure est telle que la relaxation non rayonnante se produit plus rapidement que l'émission fluorescente.

Les composés produisant une intense émission moléculaire fluorescente et de ce fait la plus utilisée sont celles qui contiennent des noyaux aromatiques. Le rapport entre le nombre de molécules fluorescentes et le nombre total de molécules excitées (où le rapport entre les photons émis et les photons excités) correspond au rendement quantique de fluorescence moléculaire Φ. Pour des molécules très fluorescentes, $\Phi \approx 1$.

Le dosage d'ADN par fluorimétrie est basé sur l'interaction non-covalente de colorants dans l'ADN contenus dans des solutions aqueuses à pH neutre. Le principe étant que le fluorophore se fixe sur l'ADN et le taux de fixation est directement lié à la quantité d'ADN (fluorophore émettra une intensité de fluorescence qui est proportionnelle à la quantité d'ADN présente).

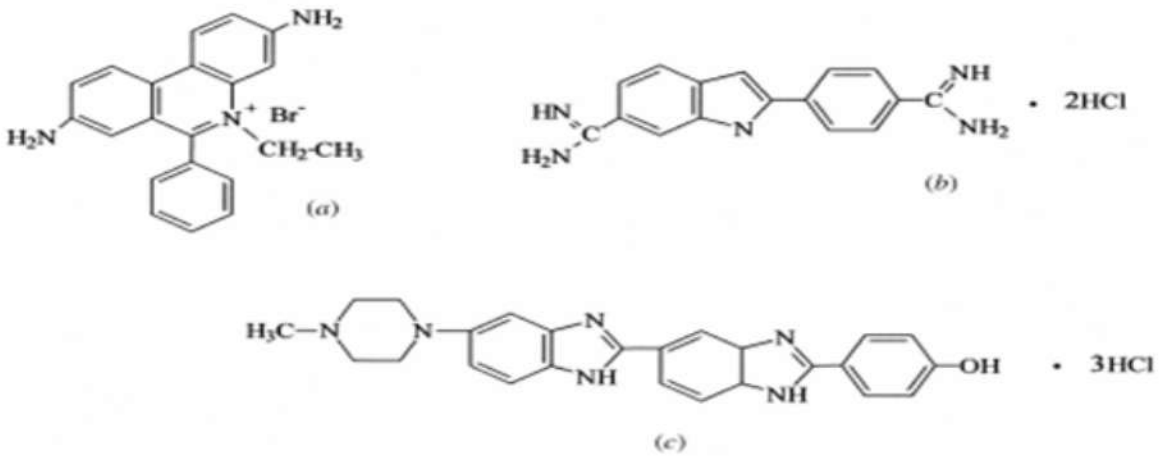


Figure 06 : Intercalants d'ADN : (a) bromure d'éthidium, (b) DAPI et (c) Hoechst 33258.

Différents fluorochromes peuvent être utilisés pour le dosage de l'ADN tels que :

- Ceux se fixant sur les paires de bases de façon spécifique :
  - Bases A-T: Hoechst 332 et DAPI
  - Bases G-C: mithramycine

- Intercalants :
  - Acridine orange, bromure d'ethidium et iodure de propidium. Ceux-ci sont connus pour être moins chers.

Ces intercalants exhibent une fluorescence limitée lorsqu'ils sont libres en solution. En se liant à l'ADN, leur fluorescence augmente considérablement.

Le bromure d'éthidium (λ_{ex} = 254 nm) se place entre les bases, se liant mieux à l'ADN qu'à l'ARN. Le Hoescht 33258 ainsi que le DAPI interagissent de façon sélective avec l'ADN.

III.3. Electrophorèse sur gel et coloration au bromure d'éthidium

L'emploi de l'électrophorèse classique est quotidien aussi bien à des fins analytiques que préparatoires. L'**électrophorèse** est une méthode utilisée en biochimie et en biologie moléculaire pour séparer l'ADN, l'ARN ou des protéines en fonction de leur poids moléculaire.

La technique est basée sur la séparation des acides nucléiques chargés négativement sous l'effet d'un champ électrique. Cette séparation s'effectue à travers la matrice d'un gel : les molécules de plus petites tailles se déplacent plus rapidement et migreront plus loin que les molécules de tailles supérieures. Donc c'est une séparation selon la taille des molécules.

De ce fait la **vitesse de migration** d'une molécule d'acide nucléique dépendra et sera en fonction de :

- Sa **masse moléculaire** donc le **nombre de bases** (ou de paires de bases). Plus les molécules sont de masse moléculaire faible, plus leurs vitesses de migration sont élevées.
- La **concentration** d'acrylamide ou d'agarose **du gel**.

La séparation d'un mélange d'acides nucléiques donné est en fonction du :

- Choix de la nature du support de l'électrophorèse (gel d'agarose ou de polyacrylamide) ;
- Diamètre des pores du gel utilisé dépendant de la concentration du gel.

III.3.1. Choix du support

Plusieurs supports peuvent être utilisés pour séparer les acides nucléiques car chacun a ces propres caractéristiques qui sont :

- **Gel d'agarose**

C'est le support le plus souvent utilisé. Les tailles des fragments qu'il est possible de séparer sont comprises entre 0,5 et 20 kb. Les gels sont coulés à l'horizontale dans des appareils transparents aux UV de manière à pouvoir suivre périodiquement la migration. **La migration est de ce fait aussi horizontale**.

- **Gel de polyacrylamide**

Il est utilisé pour la séparation de petits fragments c'est-à-dire de moins de 1000 pb. Le gel est coulé entre deux plaques de verre à l'abri de l'oxygène. **La migration se fait de façon verticale**. Ses utilisations majeures sont :

- Purification des oligonucléotides de synthèse et élimination des nucléotides libres après leur marquage radioactif ;
- Détermination des séquences d'ADN ;
- Séparation des petits fragments d'ADN.

III.3.2. Visualisation des acides nucléiques

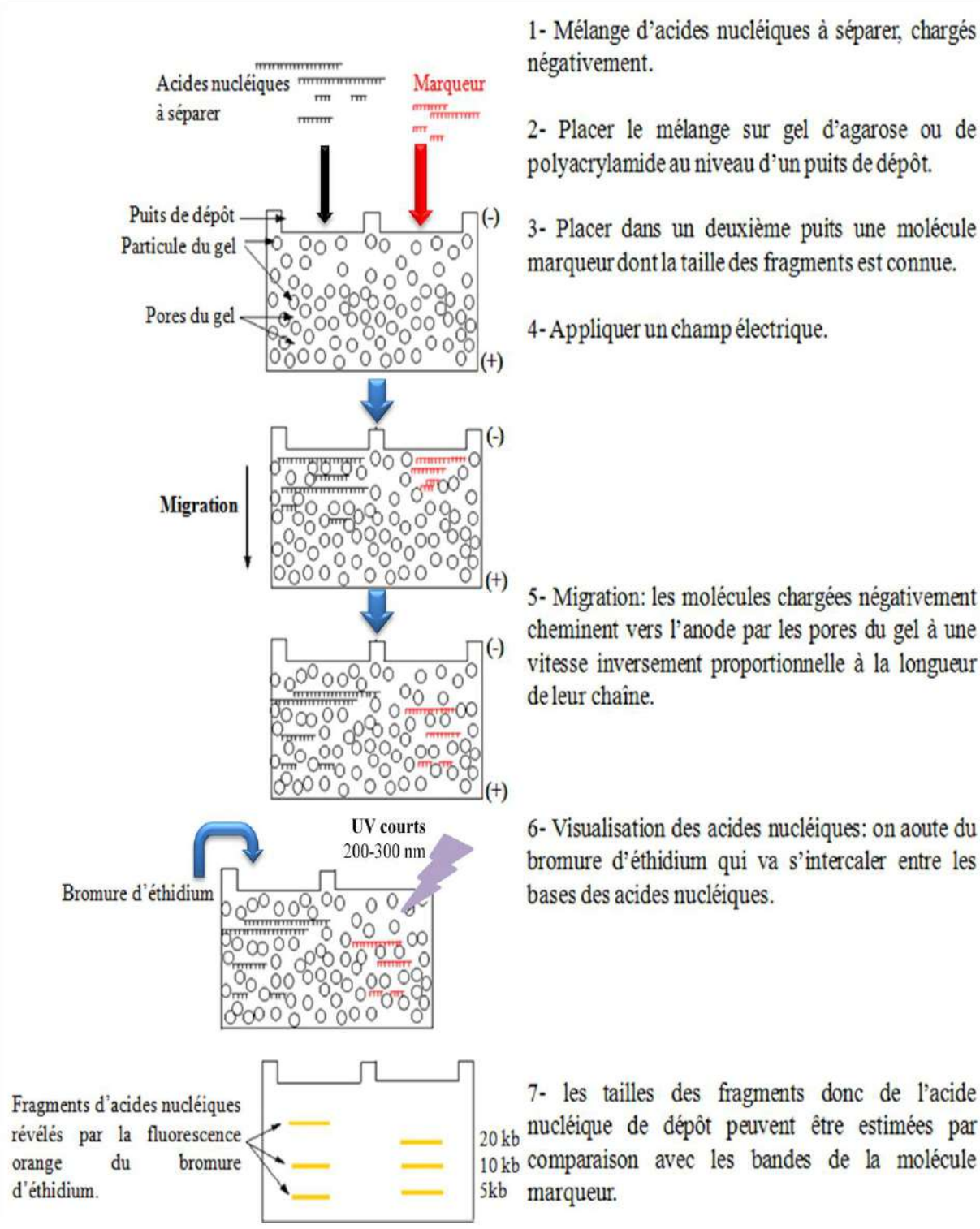
Elle se fait par coloration du gel au bromure d'éthidium (BrEt) qui est un agent s'intercalant entre les plateaux de paires de bases et émettant une fluorescence orange lorsqu'il est éclairé par des UV courts (200-300 nm). Le seuil de détection est de quelques ng. La comparaison visuelle de la fluorescence d'un échantillon avec celle d'une quantité d'ADN connue (le marqueur de taille) permet **d'estimer la quantité d'acides nucléiques déposée**.

III.3.3. Détermination de la taille d'un fragment

Elle se fait par rapport à la migration d'un marqueur approprié contenant des fragments de tailles connues.

III.3.4. Electrophorèse des ARN

Les ARN sont le plus souvent des molécules de taille importante (> 0,8kb) ; les gels utilisés seront donc d'agarose, autoclavés. Ils forment souvent des structures secondaires assez stables qui perturbent la migration électrophorétique et faussent l'estimation du poids moléculaire, c'est pour cette raison que le gel employé pour la migration des ARN doit contenir des agents dénaturants tel que le formaldéhyde provoquant la déstabilisation des appariements entre les bases complémentaires.



1- Mélange d'acides nucléiques à séparer, chargés négativement.

2- Placer le mélange sur gel d'agarose ou de polyacrylamide au niveau d'un puits de dépôt.

3- Placer dans un deuxième puits une molécule marqueur dont la taille des fragments est connue.

4- Appliquer un champ électrique.

5- Migration: les molécules chargées négativement cheminent vers l'anode par les pores du gel à une vitesse inversement proportionnelle à la longueur de leur chaîne.

6- Visualisation des acides nucléiques: on aoute du bromure d'éthidium qui va s'intercaler entre les bases des acides nucléiques.

7- les tailles des fragments donc de l'acide nucléique de dépôt peuvent être estimées par comparaison avec les bandes de la molécule marqueur.

Figure 07 : Représentation schématique d'une électrophorèse.

Chapitre III

Manipulation et étude du génome : techniques de base

Manipulation et étude du génome : techniques de base

« Enzymes de restriction »

Objectifs spécifiques

Au terme de ce cours qui traite des enzymes de restriction, vous devez être capable de :

- Comprendre ce qu'est une enzyme de restriction ;
- Connaitre ce qu'est un site de restriction ;
- Connaitre quelles sont les différentes classes d'enzymes de restriction ainsi que leur intérêts en application ;
- Comprendre le mode de fonctionnement des enzymes de restrictions ainsi que les précautions à prendre lors de leur utilisation ;
- Distinguer entre les enzymes compatibles, isoschisomères et famille d'enzymes.

I.1. Définition

Les enzymes de restriction ou endonucléases sont des protéines spécifiques, principalement d'origine bactérienne, capables de reconnaître de courtes séquences spécifiques appelés « ***sites de restriction*** » allant de 4 à 10 paires de bases et de couper/cliver l'ADN double brin, et ce quelque soit son origine, au site reconnu.

Ces sites spécifiques sont pour la plupart palindromiques, c'est-à-dire qu'ils sont composés de séquences nucléotidiques identiques sur les deux brins mais en orientations antiparallèles. L'utilisation de ces endonucléases permet donc de fragmenter l'ADN en segments de taille réduite aux extrémités bien caractérisées, dont certains peuvent contenir des gènes.

Certaines enzymes coupent le site en son milieu et produisent deux fragments dont les extrémités sont franches appelés "***bouts francs***" (blunt ends). Cependant, la plupart réalisent une coupure dissymétrique : on parle dans ce cas d'extrémités cohésives ou "***bouts collants***" (sticky ends), chaque fragment possède une chaîne qui dépasse l'autre de quelques bases.

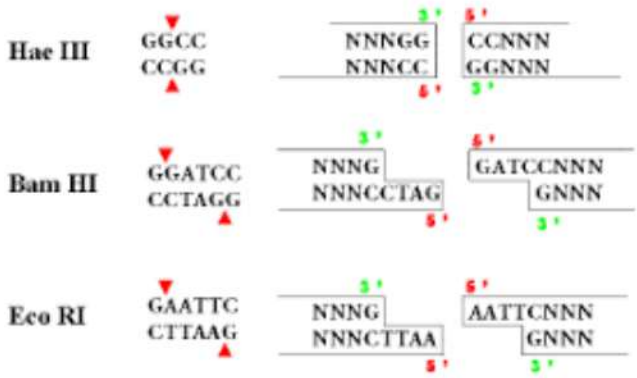


Figure 08 : Différents types d'extrémités engendrées après clivage via une enzyme de restriction.

I.2. Historique

La structure de l'ADN, son mode de réplication, les modalités d'expression des gènes et le code génétique ont été découverts entre les années 1950 à 1970. Toutefois, on ne dispose d'aucun outil permettant de manipuler aisément l'ADN ainsi que les gènes qu'il renferme. En ce sens, la découverte des enzymes de restriction constitue une avancée capitale.

En 1965, ***Werner ARBER*** découvre que les bactéries infectées par des virus (bactériophages) ont un moyen de défense contre ces parasites : elles découpent l'ADN viral en petits fragments grâce à des enzymes, véritable ciseaux moléculaires. Parallèlement, la bactérie possédait des méthylases capables de modifier (méthyler) son propre ADN afin qu'il ne soit pas reconnu par les enzymes de restriction.

Cela restreint considérablement l'aptitude des virus à répliquer leur ADN et donc à se multiplier à l'intérieur des bactéries. De ce fait, la destruction des bactéries qui suit normalement la multiplication virale n'a plus lieu.

Hamilton SMITH isola les enzymes responsables de ce phénomène (enzymes de restriction) et ***Daniel NATHANS*** transforme les enzymes de restriction en outils. Ces trois scientifiques obtiendront le prix Nobel en 1978. Depuis, beaucoup d'autres différentes enzymes de restriction ont été progressivement extraites de diverses bactéries.

I.3. Intérêts des enzymes de restriction

Des centaines de ces enzymes ont été identifiées et caractérisées. Elles reconnaissent une grande variété de sites de coupure. Pour la plupart commercialisés, elles font parties des outils indispensables aux biologistes moléculaires.

Ces outils permettent de couper l'ADN afin d'isoler certains fragments et faciliter de ce fait son étude.

Les endonucléases de restriction peuvent être employées afin de préparer un fragment d'ADN d'un gène donné qu'on appelle « insert » à être intégré/inséré dans un vecteur comme un plasmide par exemple.

Les enzymes de restriction sont utilisées pour établir une carte génétique ou carte de restriction de toute molécule d'ADN que l'on souhaite caractériser. Il s'agit de déterminer, l'ordre des sites de restriction, le long de cette molécule, qui produira, des fragments de différentes tailles, après « digestion enzymatique » de cette molécule d'ADN, celles-ci pouvant être définie par électrophorèse.

Créer de nouvelles recombinaisons d'ADN. C'est-à-dire, un fragment découpé par une enzyme de restriction se termine par « un bout collant » qui peut établir des liaisons hydrogène avec la séquence de bases complémentaire d'un autre fragment. L'ADN ligase, une autre enzyme, va réaliser la soudure définitive des deux fragments. On fabrique de cette façon une nouvelle molécule qui est un ADN recombinant.

I.4. Nomenclature des enzymes de restriction

La nomenclature des enzymes de restriction est assez bien précise. Leur nom indique leur origine et comporte plusieurs lettres (3 ou 4).

- La première lettre de dénomination de l'enzyme est écrite en majuscule, elle correspond au genre de la bactérie d'où a été extraite l'enzyme.
- La seconde lettre et la troisième lettre (en minuscules) correspondent à l'espèce de la bactérie d'où l'enzyme est extraite.
- On peut avoir une quatrième lettre écrite en majuscule correspondant à la souche bactérienne.
- Enfin pour terminer, un chiffre romain indiquant l'ordre de caractérisation de ces enzymes.

Exemples

EcoRI Extraite d'***Escherichia coli*** **R**YB site reconnu: G / AATTC.
SmaI Extraite de ***Serratia marcescens*** site reconnu: CGC / GGG.
Hind*III** signifie que c'est la troisième (**III**) enzyme de restriction isolée et caractérisée de la souche bactérienne ***Haemophilus influenza R**d**.

Tableau II : Exemples d'enzymes de restriction de type II.

Enzymes	Organisme d'origine	Séquences reconnus
*Bam*HI	*Bacillus amyloloquefaciens*	G / GATCC CCTAG / G
*Bgl*I	*Bacillus globigii*	GCCNNNN / NGGC CGGN / NNNNCCG
*Dra*II	*Deinococcus radiophilus*	RG / GNCCY YCCNG / GR
*Eco*RI	*Escherichia coli* RY13	G / AATTC CTTAA / G
*Eco*RII	*Escherichia coli* RY13	/ CC (A ou T) GG GG (T ou A) CC /
*Eco*RV	*Escherichia coli* J62/pGL74	GAT / ATC CTA / TAG
*Hpa*II	*Haemophilus parainfluenzae*	C / CGG GGC / C

*Mbo*I	*Moraxella bovis*	/ GATC CTAG /
*Nde*II	*Neisseria denitrificans*	/ GATC CTAG /
*Not*I	*Nocardia otitidis-caviarum*	GC / GGCCGC CGCCGG / CG
*Sau*I	*Staphylococcus aureus*	CC / TNAGG GGANT / CC

N: *n'importe quelle base ;* **R**: *n'importe quelle purine ;* **Y**: *n'importe quelle pyrimidine et /: Position de la coupure.*

I.5. Classification

Les enzymes de restriction font partie de la classe des endonucléases, à savoir des enzymes ayant la capacité de couper/cliver les liaisons phosphodiester entre deux nucléotides à l'intérieur d'un acide nucléique.

A l'inverse des endonucléases, les exonucléases sont caractérisé par le fait qu'elles dégradent la molécule d'ADN à partir de l'une de ses extrémités (3' ou 5'). Il existe trois types d'enzymes de restriction isolées des bactéries.

1) Type I

Dont l'action nécessite la présence de Mg^{++}, d'ATP comme cofacteur, et de S-adénosyle-méthionine. Leur site de coupure est éloigné de leur site de reconnaissance (jusqu'à plusieurs milliers de nucléotides, plus loin dans certain cas).

2) Type II

Pratiquement les seules utilisées en génie génétique, reconnaissent l'ADN à des sites particuliers et coupes dans ces sites ou à proximité immédiate d'eux. Ce sont des nucléases qui coupent à l'intérieur, et donc des endonucléases (Type II restriction endonucleases).

3) Type III

Ressemblent à celles du type I pour la séparation des sites de reconnaissance et de coupure, mais qui s'apparentent à celles du type II par leur mode d'action en coupant une vingtaine de nucléotide plus loin.

I.6. Notion d'enzymes compatibles

Lorsque deux enzymes de restriction génèrent après digestion des fractions aux extrémités cohésives complémentaires, elles sont alors dites compatibles. Ces fragments

peuvent être facilement ligaturés. Cependant, cela ne veut pas dire que les deux enzymes ont reconnu les mêmes sites.

Prenons deux exemples pour clarifier tout cela.

- L'enzyme 1 reconnaît les séquences 5' TGATCA 3' et coupe entre T et G. donc on aura :

5' T ↑**GATC**A 3'
3' A**CTAG**↓T 5'

- L'enzyme 2 reconnaît les séquences 5' GGATCC 3' et coupe entre G et G, donc on aura :

5' G↑**GATC**C 3'
3' C**CTAG**↓G 5'

Ces deux enzymes créent des extrémités cohésives 5'- **GATC** -3', on peut donc dire que les enzymes de restriction sont compatibles.

I.7. Isoschizomères et famille d'enzymes

En règle générale, les différentes enzymes de restriction reconnaissent des sites de restrictions différents. Cependant, plusieurs enzymes isolées d'organismes variés reconnaissent des séquences identiques mais elles ne clivent pas la séquence reconnue toujours au même endroit. On les appelle isoschizomères.

Exemples

***Asp* 718** ⇒ 5'- G↑GATCC -3' / 3'- CCTAG↓G -5' ***Kpn* I** ⇒ 5'- GGATC↑C -3' / 3'- C↓CTAGG -5' ***Bam* HI** ⇒ 5'- G↑GATCC -3' / 3'- CCTAG↓G -5'

Il est aussi à noter que certaines enzymes en reconnaissant des séquences différentes, vont produire les mêmes extrémités, c'est ce qui caractérise une famille d'enzymes.

Exemples

- **GATC :** *Sau* 3Al, *Bgl* I, *Bam* HI, *Bcl* II, *Xho* II.
- **CTAG :** *Mae* I, *Spe* I, *Nhe* I, *Avr* I, *Xba* I.

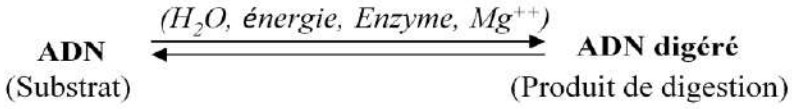
I.8. Enzymes de restriction utilisées en génie génétique

Les enzymes de restriction de type II sont pratiquement les plus utilisées en génie génétique, à cause de leurs propriétés de reconnaissance et de coupure. Le E.C. de ces enzymes est le suivant: **E.C.3.1.21.4.**

Leur **nom commun** est les endonucléases de restriction de type II. Elles réalisent des réactions de coupure à l'intérieur de la séquence d'ADN doubles brins au niveau des sites de reconnaissance et de coupure spécifiques, générant des fragments d'ADN double brin se terminant par 5'-phosphate. Elles font partie de la **classe 3** à savoir celle des **Hydrolase** et de

la sous classe 1, agissant sur les liaisons esters (**Estérases**), de la **sous-sous classe 21**en donnant des produits qui gardent leurs phosphates **5'** initial. Elles nécessitent la présence de $\mathbf{Mg^{++}}$ dans le milieu.

- **Réaction générale**



La réaction générale catalysée par les enzymes de restriction (E.C.3.1.21.n) implique la présence dans le milieu des facteurs suivants :

- **Enzyme** (E.C.3.1.21.n)
- **Substrat** ADN double brin non digéré
- **Cofacteur** : En général, seul le $\mathbf{Mg^{++}}$ est indispensable. Quelques enzymes de restriction font appel à d'autres cofacteurs. Les enzymes de méthylation ont pour coenzymes le S-adénosyl-Méthionine.
- **Facteurs physico-chimiques**: Ces facteurs (**pH**, potentiel d'**oxydoréd**uction, **forces ioniques**) contrôlent aussi ces réactions. L'énergie générée par l'hydrolyse est assez considérable afin que la réaction ne soit quasiment non réversible dans les conditions habituelles.
- Pour les enzymes de restriction, différents tampons sont utilisés. Ils permettent l'optimisation des réactions de multiples enzymes avec le même tampon (pH:7-7.9, PotOxyRed : dithiothreitol 1mM, Force ionique : NaCl ou CH_3COOK de 0 à 100 mM ; varient d'un tampon à l'autre).

- **Exemples d'enzymes de restriction de type II**

EcoRI : E.C. 3.1.21.4: C'est la première enzyme caractérisée et identifiée chez *E. coli* souche R, le site de reconnaissance et de coupure est formé de six paires de bases (nucléotides).

***HaeIII* :** C'est la troisième enzyme de restriction caractérisée et identifiée chez la bactérie *Haemophilus aegypticus*.

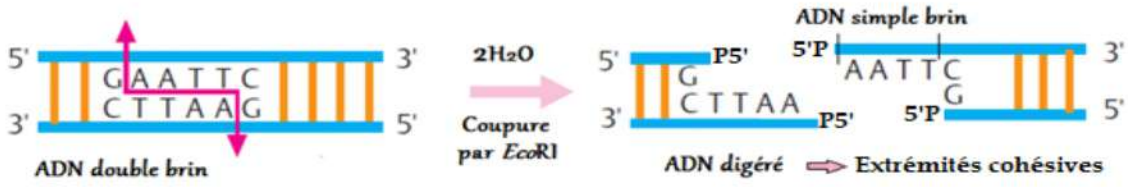


Figure 09 : Coupure de l'ADN double brin par *Eco*RI, générant des extrémités cohésives et *Hae*III, générant des extrémités franches.

- **Utilisation des endonucléases de restriction pour établir le map de restriction**

La digestion d'un fragment d'ADN de 5 kb par deux enzymes de restriction (*Bam*HI et *Eco*RI) est représentée en figure ci-dessous. L'analyse des résultats de la digestion (fig. 10.a) nous permet de déterminer les différentes possibilités de la localisation des sites de restriction pour chaque enzyme (fig. 10.b, 10.c et 10.d).

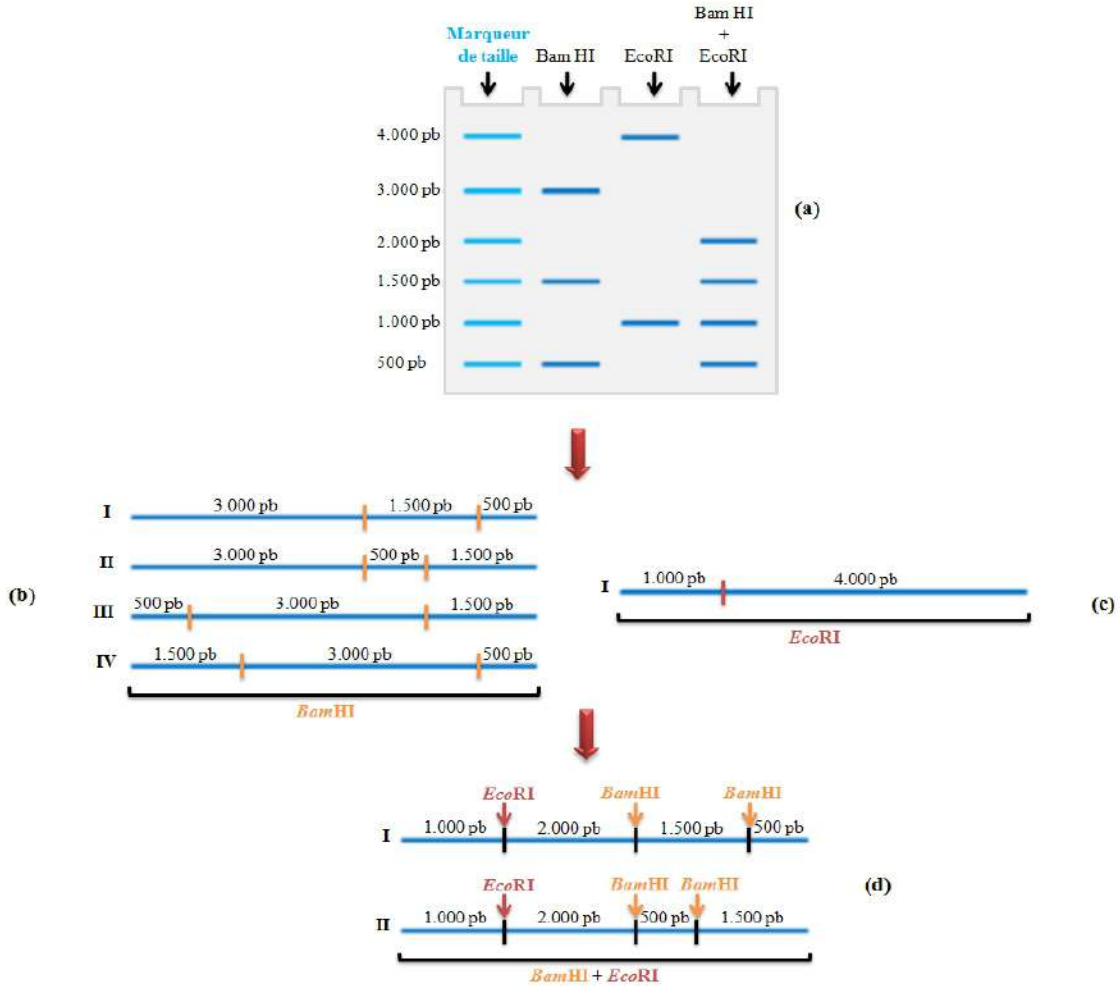


Figure 10 : Représentation schématique d'une digestion par des endonucléases de type II.

(a) Gel d'agarose des produits de digestion d'un fragment de 5kb par deux enzymes de restriction. **(b et c)** Les différentes possibilités de la localisation des sites de coupure par *Bam*HI et *Eco*RI. **(d)** Les différentes possibilités de localisation des sites de restriction des deux enzymes (*Eco*RI, et *Bam*HI).

I.9. Utilisation des endonucléases de restriction : précautions pratiques

Avant d'utiliser des enzymes en pratique, il est nécessaire de prendre certaines précautions afin d'assurer les meilleures conditions de travail et garantir des résultats fiables.

I.9.1. Précaution à prendre lors de l'utilisation d'une endonucléase de restriction

- Se laver les mains avant l'utilisation d'une ER et mettre des gants à usage unique avant la manipulation du tube contenant l'ER.
- Ne jamais toucher la surface interne du bouchon et du tube contenant l'ER (la peau possède des protéases et des nucléases).
- Ne retirer le tube contenant l'enzyme du congélateur à– 20°C qu'au dernier moment, juste avant son addition au mélange réactionnel. Le replacer au congélateur aussitôt après emploi.
- Lors du retrait d'un tube d'ER du congélateur, le mettre aussitôt dans de la glace pilée ou alors dans un bloc métallique réfrigéré.
- Utiliser toujours un cône neuf et stérile pour prélever l'enzyme.

I.9.2. Précautions lors de la digestion par une endonucléase de restriction

L'utilisation des endonucléases de restriction (ER) se fait dans un microtube de 1,5ml maintenu dans de la glace. II faut introduire successivement :

- L'eau distillée stérile.
- Le tampon de réaction 10X (de manière à obtenir une concentration finale 1X).
- L'ADN en solution, de préférence dans l'eau distillée stérile (Quantité entre 0,2 à 10 µg)
- L'endonucléase de restriction (2/3 unités/µg d'ADN). L'enzyme est toujours rajoutée en dernier dans le mélange réactionnel. Le volume réactionnel est compris entre 10 et 100 µl.

Ensuite agiter doucement le tube et centrifuger dans une minicentrifugeuse (10 000 tr/min) quelques secondes pour collecter la totalité du liquide au fond du tube.

Incuber à 37°C pendant le temps requis. Stopper la réaction enzymatique soit par addition d'EDTA à 0,5 mol.l $^{-1}$ (pH 7,5) à la concentration finale de 10 nmol.l $^{-1}$, soit par addition de solution de charge si l'ADN est aussitôt analysé dans un gel d'électrophorèse. Mélanger rapidement en vortéxant le tout et centrifuger à 12000 tr/min afin d'assembler la totalité du liquide et éliminer les bulles.

Manipulation et étude du génome : techniques de base

« Hybridation moléculaire : concepts de base »

Objectifs spécifiques

Au terme de ce cours qui traite de l'hybridation moléculaire : concepts de base, vous devez être capable de :

- Comprendre ce qu'est une hybridation moléculaire ;
- Connaître ce qu'est une sonde ;
- Saisir quels sont les éventuelles applications de l'hybridation moléculaire ;
- Connaitre quelles sont les différentes techniques relative à l'hybridation moléculaire ;
- Distinguer les différences entre les procédés et quand les appliquer.

II.1. Définition

L'hybridation moléculaire est une technique qui utilise la propriété que présente une molécule d'acide nucléique monobrin de s'associer spontanément et de façon spécifique et réversible à une autre molécule monobrin qui lui est complémentaire conduisant à la formation d'un double brin ou duplex.

Cette association s'effectue par l'établissement de liaisons hydrogènes entre les bases azotées selon leurs spécificités.

La composante en bases, la complexité de la séquence ainsi que sa longueur sont autant de facteurs impliqués dans la formation et la stabilité du duplex. L'hybridation est à la base de nombreuses techniques de biologie moléculaire impliquant la mise en présence d'au moins deux simples brins d'acides nucléiques dans des conditions physico-chimiques précises. Le brin avec au moins une partie de la séquence connue est ce qu'on appelle une sonde, et l'autre brin, le brin voulue caractériser constitue la cible. Il est important de procéder au marquage d'un des deux brins par couplage chimique avec une molécule capable de produire un signal.

II.2. Applications

Les applications de l'hybridation moléculaire sont nombreuses. Cette technique permet notamment la :

- Comparaison des tailles de génomes sans séquences répétitives (virus, bactéries, mitochondries..). Le $Cot_{1/2}$ sera directement fonction de la taille du génome.
- Analyse globale des pourcentages d'homologie de séquence entre deux espèces proches : A réaction complète, le % d'hybrides peut être assimilé grossièrement au % de séquences homologues.

- Analyse des séquences répétitives : qui auront tendance à provoquer une déformation du début de la courbe représentant la cinétique d'hybridation du fait que les séquences répétitives possèdent une cinétique de réassociation plus rapide.

II.3. Facteurs influençant l'hybridation moléculaire

Plusieurs facteurs peuvent influencer l'hybridation des acides nucléiques on citera notamment :

II.3.1. Température

Il est largement admis que les molécules d'acides nucléiques doubles brins peuvent être dénaturées par la chaleur. En fonction de la température, les molécules vont se présenter sous forme double brin, simple brin ou encore un mélange des deux. Sachant que la température de fusion (T_m) représente la température à laquelle la moitié de l'ADN est sous forme double brin et l'autre moitié sous forme monobrin, la dénaturation, processus très rapide (quelques secondes), d'une molécule d'ADN initialement double brin nécessite une température plus élevée que la T_m.

La renaturation de deux monobrins dissociés nécessite une température plus faible par rapport à la Tm. Ce processus de renaturation est considéré comme étant assez lent, voir de quelques secondes à plusieurs dizaines d'heures relative à la complexité des deux brins d'ADN et leur concentration dans une solution. Il est à noter qu'une amorce s'hybridera en quelques secondes avec une matrice d'ADN du fait que l'amorce a une séquence très courte et qu'elle est présente à une concentration extrêmement élevée. Cependant, un génome entier mettra plusieurs jours à se renaturer parce que sa séquence est très longue et complexe et que la concentration de chacune des régions de l'ADN est relativement faible.

II.3.2. Composition du milieu réactionnel

La composition du milieu réactionnel peut influencer le processus d'hybridation moléculaire à savoir :

a) Teneur en sel et en particulier en cations monovalents type Na^+

Les ions Na^+ permettent de stabiliser les ions phosphate des brins d'ADN et limiter la répulsion entre les chaînes monobrins. De ce fait, la présence de Na^+ favorise la réassociation entre eux.

b) La teneur en produits chimiques : formamide

La formamide, un amide issu de l'acide formique, crée des liaisons Hydrogène avec les bases azotées de la molécule d'ADN, ce qui rend la dénaturation plus rapide. Cette molécule a tendance à déstabiliser les hybrides imparfaits.

II.4. Sondes nucléotidiques

Une sonde nucléotidique est un segment de nucléotides obligatoirement monobrin qui permet de rechercher de manière spécifique un fragment d'acide nucléique (ADN ou ARN) que l'on désire étudier.

Elles sont de taille assez variable (oligonucléotide de 20-30 nucléotides ou a l'opposé de plusieurs centaines de nucléotides) et doivent être complémentaire et antiparallèle du fragment recherché. Dans un mélange complexe où s'effectue l'hybridation moléculaire, la sonde doit être facilement repérable grâce à un marquage avec un radio-isotope (marquage chaud), mais il existe également des sondes appelées sondes froides sans marquage par un radio-isotope.

II existe plusieurs possibilités pour obtenir une sonde nucléotidique :

- Dans le cas où la séquence de l'ADN à repérer est connue, la sonde peut être obtenue par synthèse chimique. Si la séquence cible est inconnue, on peut étudier la protéine correspondante et remonter grâce au code génétique à la séquence d'ADN. Dans ce dernier cas, le travail est particulièrement laborieux (nombre de codons élevé pour un même acide amine).
- Une sonde peut être un ADN complémentaire (ADNc). Une partie seulement de cet ADNc est utilisée (après action d'enzymes de restriction et clonage des fragments obtenus).
- Une sonde peut être théoriquement de l'ARNm.

II.5. Techniques d'hybridations

II.5.1. Southern blot

Cette méthode a été initialement mise au point en 1975 par Edward M. SOUTHERN afin de rechercher des fragments d'ADN en les hybridant avec une sonde complémentaire par électrophorèse. Elle consiste à détecter spécifiquement des fragments d'ADN transférés sur filtre par leur hybridation à des séquences complémentaires marquées par un radio-isotope. Cette technique permet de repérer un fragment d'ADN particulier dans un génome. Les étapes de cette méthode sont les suivantes:

- Extraction de l'ADN génomique : Cette extraction s'effectue à partir d'un échantillon donnée. Le choix de la technique d'extraction dépendra de l'échantillon.

- Digestion par des enzymes de restriction de l'ADN génomique. L'ADN génomique est digéré par des enzymes de restriction différentes: dans le tube 1, on réalisera une digestion par l'enzyme 1; dans le tube 2, une digestion par l'enzyme 2, ...). On peut également réaliser des digestions par deux enzymes dans un même tube. Dans ces conditions, un très grand nombre de fragments sont obtenus, toutefois seuls quelques fragments correspondent à la totalité ou à une partie du gène étudié.

- Séparation électrophorétique des fragments de restriction de notre molécule d'ADN via une électrophorèse sur gel d'agarose. Suivant la taille du génome étudié, on peut obtenir jusqu'à 106 fragments voir plus, on verra sur le gel non pas des fragments distinctement séparés mais une trainée d'ADN due aux fragments de taille très proche les uns des autres.

- Après électrophorèse des fragments d'ADN bicaténaires obtenus par digestion enzymatique, une dénaturation des fragments par un traitement alcalin par la soude du gel d'électrophorèse (immersion dans une solution alcaline) est réalisée afin d'obtenir des fragments d'ADN monobrin.

- Transfert des fragments monocaténaires du gel d'agarose à un support souple de nitrocellulose (ou nylon) préalablement imprégnée d'une solution saline concentrée. Le transfert des fragments monocaténaires du gel d'agarose à une membrane de nylon s'effectue par simple capillarité en appliquant un poids pour permettre le maintient de la membrane en sandwich comprise entre le gel et une couche épaisse de substance absorbante.

- Fixation permanente des fragments monocaténaires d'ADN sur le support souple se fait soit par chauffage (cas de la nitrocellulose) soit par exposition aux rayonnements ultra-violets (cas du nylon).

- Une étape de préhybridation permettant de traiter les sites laissés libres est faite, ils seront saturés par une solution contenant une protéine telle que la sérum albumine et/ou traités à l'aide d'un détergent comme le laurylsulfate de sodium (SDS).

- Hybridation avec une sonde spécifique complémentaire marquée par un radio-isotope. L'hybridation est réalisée en milieu liquide sous agitation, cette opération est suivie de lavages pour éliminer les sondes non appariées.

- Visualisation de l'hybridation se fera par autoradiographie dans le cas de l'emploi d'une sonde marquée à l'aide d'un radio-isotope (phosphore 32 par exemple) ou par développement d'un film photographique impressionné après réaction de chimioluminescence. C'est la trace de la sonde sur un film d'autoradiographie qui va être mesurée. Les fragments hybridés avec la sonde apparaissent sur l'autoradiographie.

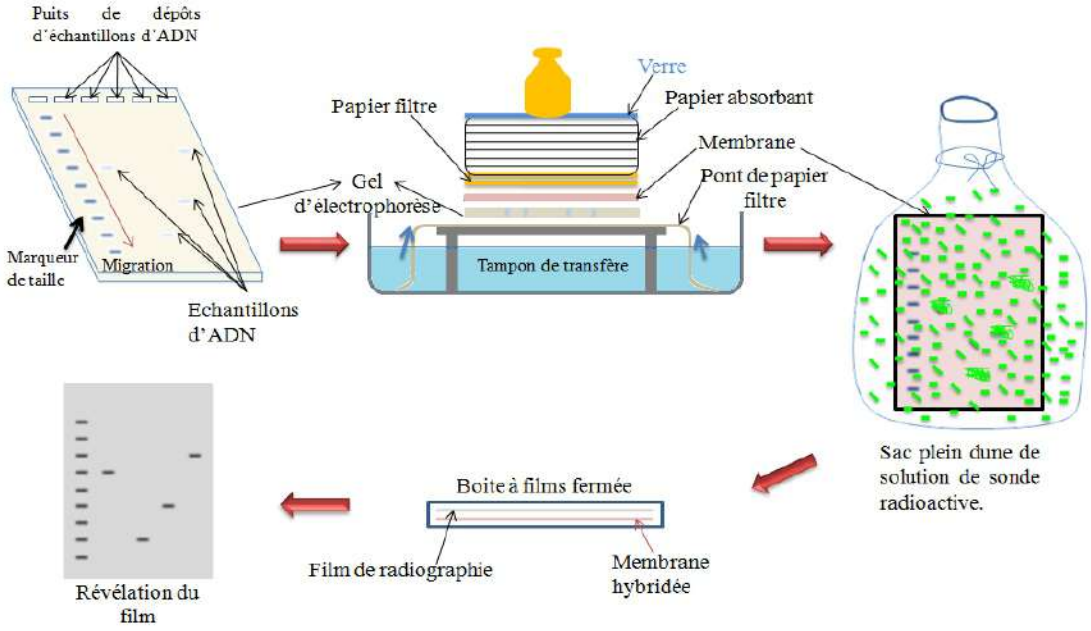


Figure 11 : Représentation des différentes étapes du Southern blot.

II.5.2. Northern blot

C'est une méthode permettant l'analyse des ARN cellulaires de manière qualitative et semi-quantitative. Le principe de cette technique est le même que pour le Southern blot mais ici ce sont les **ARN** qui sont étudiés au sein d'une population hétérogène en contacte avec une sonde reconnaissant au moins 20 bases de la partie codante du gène correspondant.

Elle permet notamment d'apprécier la distribution de l'ARN dans les tissus, d'affirmer sa présence ou son absence dans une cellule ayant servi à la préparation de l'échantillon d'ARN, étudier son abondance relative, déterminer sa taille et détecter les intermédiaires de maturation et les différentes formes d'épissage de l'ARN.

Dans cette technique, il n'y aura pas besoin de digérer l'ARN par des enzymes de restriction car les ARNs totaux sont des entités isolables, à l'inverse des gènes.

Il va être analysé directement par électrophorèse ce qui permettra de séparer les ARNs en fonction de leur taille en référence avec un marqueur de poids moléculaire radioactif. Durant l'électrophorèse, on va ajouter du formaldéhyde au gel d'agarose comme agent

dénaturant en replacement de l'hydroxyde de sodium utilisé pour le Southern blot car ce dernier a tendance à dégrader l'ARN. L'addition du formaldéhyde permet de casser les structures secondaires pouvant se former au niveau de la molécule d'ARN (bien que celle-ci soit généralement monobrin, il peut y avoir appariement entre des segments au niveau de la molécule d'ARN par complémentarité de bases). Les différents fragments d'ARN préalablement séparé seront transféré sur membrane par un phénomène de capillarité, d'où le nom de « blotting » donné à cette étape, par la suite hybridation et détection des ARNs grâce à une sonde oligonucléotidique marquée au phosphore 32 et révélée par autoradiographie.

Il est à noter que l'extrême sensibilité des ARN vis-à-vis des ribonucléases nécessite de travailler dans des conditions de stérilité (port de gants indispensable, matériel et solution préalablement autoclavés).

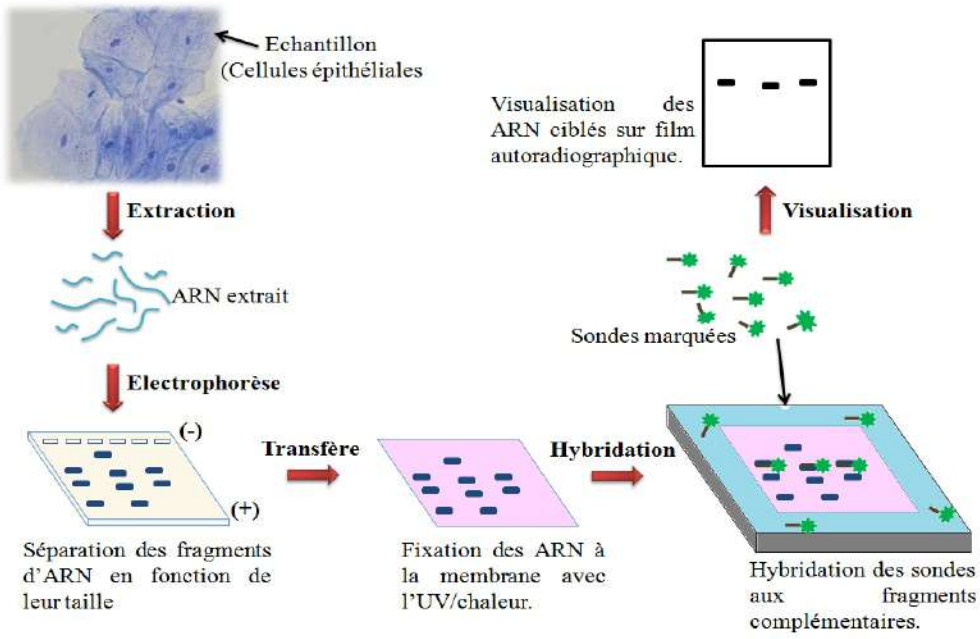


Figure 12 : Etapes du Northern blot.

II.5.3. Puces à ADN

Les puces à ADN est une autre méthode permettant l'hybridation moléculaire. Le principe général d'une puce à ADN se base sur une hybridation entre l'ADN d'un échantillon biologique et un ensemble de sondes immobilisées et organisées sur un support solide généralement une lame de verre.

Ce processus suit la succession de quatre étapes distinctes à savoir : la fabrication de la puce, la préparation de la cible, l'hybridation et enfin la lecture. L'avantage de celle-ci étant la possibilité d'analyser une grande quantité d'information génétique cible simultanément (une

centaine d'espèces différentes sur une même puce), dont les positions sur le support sont parfaitement connues.

- ***Fabrication et types de puce***

Une puce est constituée tout d'abord d'une surface solide, habituellement du verre mais peut être également d'une autre nature tel que le nylon ou autre, recouverte de polylysines, molécules capables de fixer des sondes. Chaque sonde correspond à une séquence spécifique d'un gène recherché dont la taille varie entre 200 et 2000 pb mais peuvent aussi être des oligonucléotides de 50 à 70 pb. Via des interactions électrostatiques, il y'aura dépôt des copies de la même sonde sous forme de « spot », à des emplacements précis sur la puce. Les sondes sont ensuite dénaturées afin que leurs fragments d'ADN se retrouvent sous forme simple brin et soient en mesure de capter leur brin complémentaire (cible) présent dans l'échantillon à analyser.

Les molécules de polylysines, en périphérie des spots, n'ayant pas reçu d'ADN lors de cette étape de fabrication, sont bloquées afin d'empêcher qu'elles ne captent des cibles ou qu'elles ne faussent la lecture des résultats, lors de l'analyse de l'échantillon. Différentes techniques de dépôt des sondes sur la puce existent et ce dépôt peut se faire après synthèse ou bien on réalisant une synthèse in situ par photolithographie.

Différentes puces à ADN peuvent être utilisées en fonction de la densité des sondes, du support utilisé, de l'objectif de l'expérience et des résultats attendue. On citera : les macroarrays, les microarrays et les puces à oligonucléotides.

Au niveau des **macroarrays**, on a une densité de fragments d'ADN d'environ 25 spots/cm^2 et les dépôts se font sur une membrane de nylon, pour les **microarrays** la densité du dépôt est plus élevée (1000 ADN/cm^2) et il se fait sur une lame de verre. Les **puces à oligonucléotides** représentent la miniaturisation la plus poussée (300 000 oligonucléotides/cm^2) dans ce cas la taille des oligonucléotides est d'environ 25 pb.

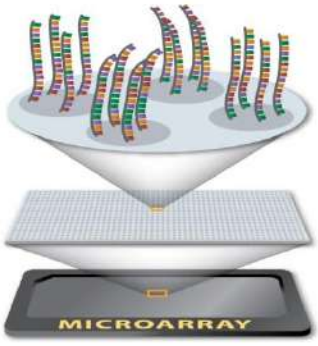


Figure 13 : Représentation de la formation d'un microarray.

- ***Préparation de la cible***

L'échantillon à analyser va subir une extraction d'ARN puis ce dernier subira une transcription inverse afin d'obtenir un ADNc. L'ADNc ainsi formé sera préalablement marqué par un fluorochrome en vue de sa révélation une fois fixé à une sonde de la puce.

- ***Hybridation***

Les ADNc auparavant marqués sont mis en contacte avec la puce puis placé à incuber à 60°C pendant toute une nuit. Dans de telles conditions, les brins d'ADN simple brin marqués au fluorochrome formant les cibles peuvent s'apparier avec leurs sondes complémentaires présentes sur la puce et former de l'ADN double brin. La phase d'hybridation est suivie d'un lavage destiné à débarrasser la puce des cibles nucléiques non hybridées.

- ***Lecture***

Les spots sont alors excités par un laser et la fluorescence émise est visualisée par un photo-multiplicateur couplé à un microscope confocal. Une image va être obtenue, dont le niveau de gris représente l'intensité de la fluorescence lue. En se basant sur le/les spots émettant une fluorescence, il est possible de déterminer la présence de la cible recherchée.

- ***Applications des puces à ADN***

Les puces à ADN sont très utilisées afin de procéder à l'**évaluation de l'expression des gènes** tels que celui du métabolisme respiratoire et de fermentation chez la levure ainsi que d'autres organismes, l'**analyse des voies biochimiques** telle que la cascade de signalisation des récepteurs à activité tyrosine kinase et bien d'autres, **validation de mécanismes d'action de médicaments** en étudiant leurs mécanismes d'action, en **classification phénotypique et prédiction**, en **analyses d'ADN génomique** et de ces variations, d'une grande utilité en recherche bio-médicale et en médecine car elles permettent de diagnostiquer certaines maladies ou d'en déduire les prédispositions.

Exemple

Le processus étant assez simple puisqu'il s'agit d'extraire l'ARN messager des cellules du patient (ADN cible) puis de les mettre en contact, sur une puce, avec l'ADN de la maladie (ADN sonde). Si les ADN sont complémentaires, c'est-à-dire si leur bases (A, T, G et C) s'apparient, cela signifie la présence d'une maladie qui peut être détectée.

Manipulation et étude du génome : techniques de base

« Techniques de réaction de polymérisation en chaine (PCRs) »

Objectifs spécifiques

Au terme de ce cours qui traite de la réaction de polymérisation en chaine et de ces différentes variantes, vous devez être capable de :

- Comprendre ce qu'est une réaction en chaine de polymérisation (PCR) ;
- Comprendre le principe et les étapes de la PCR ;
- Connaitre quelles sont les différents types de PCR ;
- Distinguer les différences entre chaque variante de PCR et quand les appliquer.

III.1. Définition

La réaction de polymérisation en chaine ou *Polymerase Chain Reaction* (PCR) est une technique rapide d'amplification *in-vitro* d'une séquence donnée d'ADN. Elle permet d'obtenir plusieurs millions de copies identiques de cette séquence (de 1 à 10^9 copies).

C'est une méthode qui dérive des connaissances du mécanisme moléculaire de réplication de l'ADN et de ces propriétés de dénaturation-renaturation.

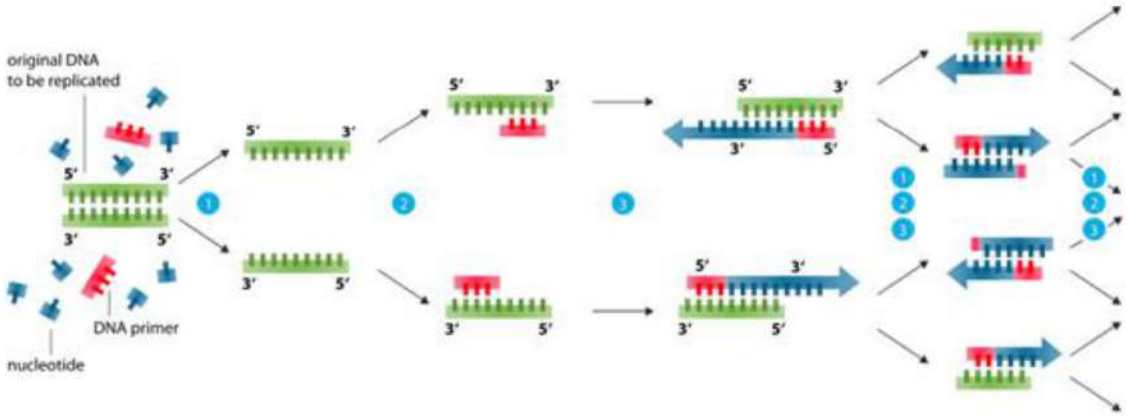


Figure 14 : Amplification d'ADN (PCR).

III.2. Historique

La PCR fut inventée par **Kary Brank MULLIS** en 1983 et brevetée en 1985, et il a eu pour cela le Prix Nobel de Chimie en 1993. La première publication parut à ce sujet date du 20/12/1985 dans *Science*.

III.3. Composantes du mélange réactionnel

La réaction de polymérisation en chaîne est réalisée dans un mélange réactionnel qui comprend plusieurs éléments indispensable afin qu'il y'ai amplification du fragment d'ADN cible. Ces éléments sont les suivants :

- **Matrice d'ADN**

Il faut obligatoirement un ADN double-brin ou ADNc. Cette matrice doit-être pure, c'est-à-dire qu'elle ne doit pas contenir de substances chimiques susceptibles d'inhiber la polymérase telles que le phénol, l'acétate, l'éthanol, l'EDTA, autrement dit les réactifs habituels des extractions/purifications d'ADN. La matrice doit y être en quantité suffisante, c'est-à-dire entre 10 et 100 brins d'ADN.

- **Amorces (Primers)**

Pour parvenir à amplifier sélectivement des séquences nucléotidiques à partir d'un extrait d'ADN par PCR, il est indispensable de disposer d'au moins une paire d'oligonucléotides, qui servira d'amorces à la réplication.

Ce sont de coutes séquences d'ADN monocaténaires contenant entre 10 et 30 nucléotides, afin de garantir une hybridation suffisamment spécifique avec les séquences d'intérêt de l'ADN matriciel. Ces amorces doivent contenir une plus grande proportion (30 à 40%) de GC car la triple liaison hydrogène qu'il peut y avoir entre ces bases lors de l'hybridation des amorces favorisent une meilleure stabilité de l'association amorce-ADN matrice.

Elles sont synthétisées par voie chimique et il doit s'agir de séquences complémentaires de l'ADN à amplifier de part et d'autre de la séquence d'intérêt ciblée par l'amplification. Mais il ne faut pas non plus que les amorces présentent de fortes complémentarités de séquence entre elles sans quoi elles s'associeraient l'une à l'autre et n'effectueraient pas correctement l'amorçage de la réplication. Elles doivent avoir leur extrémité 3'-OH libre afin de permettre le démarrage de la polymérase.

L'une des amorces est conçue pour reconnaître par complémentarité une séquence située en amont du brin 5'-3'; l'autre pour reconnaître, toujours par complémentarité, une séquence située en amont du brin complémentaire (3'-5') du même fragment d'ADN.

Plus les amorces sont riches en C et G, plus leur température d'hybridation (Tm) peut être élevée (elle est en général comprise entre 40 et 65°C, mais peut atteindre 70°C) et donc plus la spécificité de l'hybridation est grande. À des températures d'hybridation basses, les amorces ont tendance à s'hybrider de façon beaucoup moins sélective ce qui peut se traduire par une amplification de séquences d'ADN non souhaitées (amplification non spécifique).

La détermination de la température d'hybridation des amorces avec la matrice est donc cruciale. Une formule, simplifiée à outrance, du calcul de la température optimale d'hybridation d'une amorce, en degrés Celsius, est la suivante :

$$T_m = 4\,(G+C) + 2\,(A+T) - 4$$

Où G, C, A et T sont respectivement les nombres de bases G, C, A et T composant l'amorce.

Exemple

Pour une amorce qui contiendrait 5 fois chacune des bases, la température optimale d'hybridation serait : $T_m = 4 \times 10 + 2 \times 10 - 4 = 56°C$

- **Désoxyrhibonucléotides**

Les quatre dNTP (desoxyribonucléotides) fournissent à la fois l'énergie et les nucléotides nécessaires à la synthèse de l'ADN lors de la polymérisation en chaîne. Ils doivent être fournis en excès dans le mélange réactionnel.

- **Taq polymérase**

Lors d'une PCR, on utilise des polymérases purifiées thermorésistantes qui peuvent travailler à hautes températures. Elles sont extraites d'archaebactéries extremophiles (*Thermus aquaticus*) qui vivent dans les sources chaudes et résistent à des températures supérieures à 100°C. Cette température de 100°C est généralement suffisante pour dénaturer la plupart des protéines sauf la polymérase (Taq polymérase) qui trouve sa température de confort à 72°C, température optimum pour l'activité polymérase.

- **Tampon ou Buffer**

Les enzymes ont besoin de tampons particuliers, le tampon utilisé pour la réaction PCR sert à maintenir stable le pH du milieu réactionnel au niveau optimal pour la Taq polymérase. En général on utilise du Tris-HCl à pH basique entre 8 à 9.

Il contient des cations bivalents Mg^{2+}, cofacteurs indispensables à la réaction de polymérisation avec la Taq polymérase. La présence de cations bivalents Mg^{2+} et de cations monovalents (K^+ ou NH_4^+) dans le milieu réactionnel vont permettre de neutraliser, au niveau de l'ADN, les charges négatives des groupements phosphates et ainsi stabiliser les hybrides ADN/amorces.

- **Thermocycleur**

Il s'agit d'un appareil contenant une enceinte fermé où l'on dépose des microtubes contenant le mélange réactionnel avec l'échantillon d'ADN, qui permet d'exposer les

échantillons à une variation de température de manière très rapide (de 0 à 100°C, le chauffage et refroidissement) lors des différentes étapes que comptent les cycles d'une PCR donnée. Cet appareil permet la programmation de la durée et de la succession ainsi que le nombre des cycles de paliers de température. Chaque cycle comprend trois périodes de quelques dizaines de secondes ce qui fait que la réaction PCR ne dure que quelques heures (2 à 3 heures pour une PCR de 30 cycles).

Remarque

Le mélange réactionnel se prépare toujours dans de la glace pilé et les différents éléments qui le compose seront placés selon l'ordre suivant dans les microtubes: H_2O, le tampon, $MgCl_2$, les nucléotides, les amorces, l'enzyme (Taq polymérase) et enfin l'ADN à amplifier.

On réalise toujours un témoin négatif dans lequel l'ADN est remplacé par de l'eau afin de déceler une éventuelle contamination des réactifs (La contamination par de l'ADN précédemment amplifié représente le principal risque). Aucune bande ne doit être visible lors de la vérification des produits PCR finaux et permet donc d'éviter les FAUX POSITIFS.

Un témoin positif contenant un extrait d'ADN connus et déjà identifié par PCR permettra de valider la bonne conservation des réactifs, de l'enzyme responsable de la polymérisation et des performances du thermocycleur. Donc, il va obligatoirement produire le signal attendu, si dans le cas contraire, il n'y a pas us de signal dans le contrôle positif, il n'est pas possible de savoir si les échantillons sont négatifs ou si l'amplification n'a pas fonctionné (permet donc d'éviter les FAUX NEGATIFS).

Un témoin interne d'extraction présent dans le milieu réactionnel renseignera sur la présence ou l'absence d'inhibiteurs de l'amplification et donc sur l'efficacité de l'extraction de notre ADN.

III.3. Etapes de la PCR

La réaction PCR passe par plusieurs cycles, chacun de ces cycles comprends trois étapes qui sont :

1) Dénaturation thermique de l'ADN (94°C)

La première étape est réalisée à une température de 94°C, désigné température de dénaturation. À cette température, l'ADN, utilisé comme matrice durant le processus de réplication, est dénaturé : les liaisons hydrogène ne peuvent être maintenues à des températures supérieure à 80°C, de ce fait, les ADN double-brin sont dénaturés en ADN monocaténaires (ADN simple-brin).

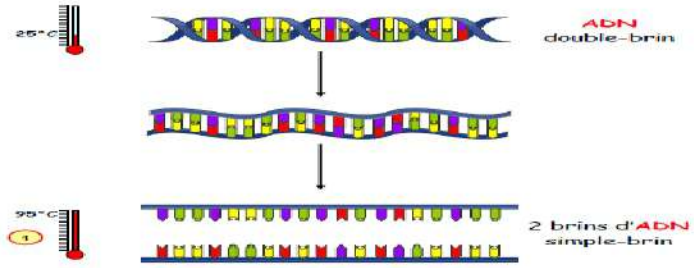


Figure 15 : Dénaturation de la matrice d'ADN.

2) Hybridation des amorces (40 – 70°C)

La deuxième phase est généralement réalisée à une température comprise entre 40 et 70°C, appelée température d'hybridation des amorces. La baisse de la température favorise la formation de liaisons hydrogène et ainsi aux brins complémentaires de s'hybrider. Les courtes séquences monocaténaires, les amorces, complémentaires des régions qui flanquent l'ADN à amplifier, s'hybrident plus aisément que les longs brins d'ADN matriciel. Afin que l'hybridation soit plus sélective, donc plus spécifique, la température d'hybridation doit être plus élevée.

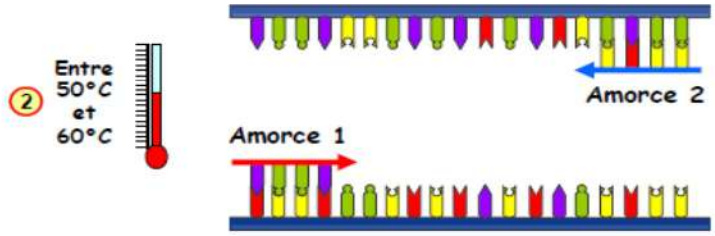


Figure 16 : Hybridation des amorces au fragment d'ADN cible.

3) Elongation ou extension des amorces (72°C)

La troisième étape se fait à une température de 72°C, nommée température d'élongation. À 72°C, la Taq polymérase s'associe aux ADN simple brin préalablement amorcés et utilise les désoxyribonucléosides triphosphates présents dans le mélange réactionnel pour catalyse la réplication. Par conséquent, les régions de l'ADN matriciel en aval des amorces sont synthétisées de manière sélective.

Figure 17 : Elongation du fragment d'ADN cible.

Remarque

Au cycle qui suie, les fragments synthétisés au cycle antécédent servent de matrice à leur tour et après quelques cycles, l'espèce prépondérante coïncide avec la séquence de l'ADN comprise entre les régions hybridées de part et d'autre avec les amorces.

Il est nécessaire de compter entre 20 à 40 cycles afin de synthétiser une quantité analysable d'ADN (environ 0,1 microgramme). En théorie, chaque cycle double la quantité d'ADN présente lors du cycle précédent (2^n).

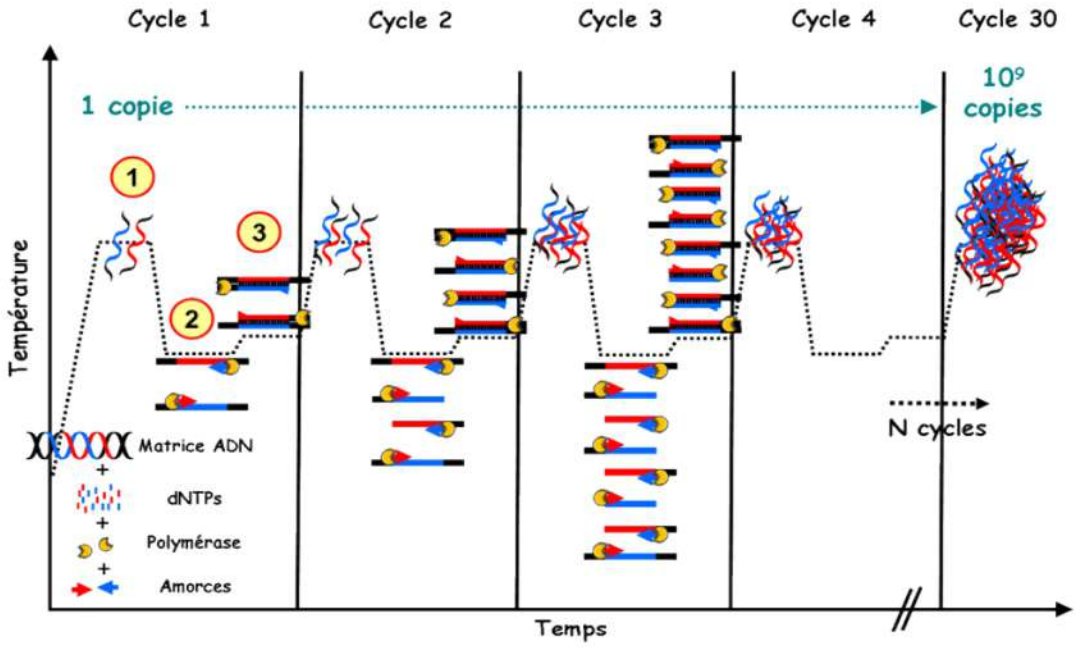


Figure 18 : Représentation des différents cycles d'une PCR.

4) Révélation des produits d'amplification

Lorsque la PCR est achevé, on se retrouve avec des produits d'amplification sensé contenir des fragments de taille déterminée et en quantité suffisante. La détection, la visualisation et l'analyse des produits PCR peuvent être très rapidement réalisées par électrophorèse sur gel d'agarose. Cette électrophorèse permet de faire migrer les acides nucléiques au travers du gel additionné de ***bromure d'éthidium*** (BET), produit intercalant qui se glisse entre les bases des acides nucléiques faisant apparaitre à la molécule d'ADN une fluorescence orange sous illumination par des UV courts (environ 280 - 320 nm) grâce au trans-illuminateur. La vitesse de migration étant dépendante de la masse de la molécule, donc de la taille (nombre de bases) de l'ADN amplifié, la présence et la taille des amplicons peuvent être facilement vérifiable sur le gel grâce à un marqueur de taille connus.

Des produits de très petite taille, correspondant généralement à des dimères d'amorces et parfois aux amorces elles-mêmes, sont habituellement visibles très près du front de migration sous forme de bandes plus ou moins diffuses. Selon les conditions réactionnelles, il

arrive que des fragments non spécifiques d'ADN soient amplifiés en quantité plus ou moins abondante, formant des bandes nettes ou des « traînées » (smear).

III.5. Différents types de PCR

Depuis 1985, beaucoup de méthodes ou de variantes de la PCR ont vu le jour : real-time PCR, PCR compétitive, PCR in situ, RT-PCR,…

III.5.1. Reverse transcriptase PCR (RT-PCR)

Ce type de PCR permet d'amplifier indirectement des ARN. Dans ce cas là, la PCR est précédée d'une transcriptase inverse qui permet le passage de l'ARN vers l'ADNc grâce à une enzyme qui est la « Reverse Transcriptase » (enzyme virale).

À l'issue de la transcription inverse, les ARN sont hydrolysés (traitement alcalin, RNase ou température). Les étapes suivantes sont réalisées dans l'enceinte du thermocycleur. Les ADNc monocaténaires sont alors répliqués par l'ADN polymérase au cours d'un premier cycle de température. D'autres cycles sont réitérés afin d'amplifier les ADNc bicaténaires en grande quantité.

Dans un phénotype cellulaire donné, on estime que 10 à 15 000 gènes sont exprimés chez l'homme et la plupart des mammifères. Certains transcrits cellulaires sont exprimés à quelques centaines voire quelques milliers de copies par cellule, mais la majorité des transcrits représente un faible nombre de copies. Des variations qualitatives ou quantitatives que connaissent les profils d'expression des transcrits reflètent la dynamique biologique de la cellule. Dans un contexte physiologique ou pathologique donné, l'identification des variations d'expression de gènes peut par conséquent fournir des informations inestimables relative à la fonction des gènes ainsi que l'influence de facteurs de modulation de leur expression, qu'ils soient physiologiques ou d'origine environnementale. L'analyse des changements d'expression génique intervenant dans une pathologie peut conduire à de nouvelles cibles thérapeutiques ou diagnostiques. Enfin, d'un point de vue fondamental, étudier le profil d'expression des gènes permet d'avancer dans la compréhension des mécanismes de physiologie cellulaire.

III.5.2. Nested-PCR

C'est une méthode mise au point afin de pouvoir amplifier d'une manière plus spécifique un fragment d'ADN. Il y a deux PCR successives. La première PCR avec un couple d'amorces définies. Les amplifiats obtenus servent de matrice pour la seconde PCR qui possède alors un couple d'amorces plus internes dans la séquence.

III.5.3. PCR quantitative en temps réel

La PCR quantitative en temps réel (*Quantitative real-time PCR*) est une application de la PCR qui a été mise au point au milieu des années 1990. En détectant la fluorescence émise par les produits de PCR néo formés, le processus d'amplification PCR peut être suivi en continu à savoir « en temps réel ».

L'augmentation du signal de fluorescence est proportionnelle au nombre d'amplicons produits pendant la réaction PCR. Il devient possible de suivre la réaction PCR durant sa phase exponentielle, en observant la quantité de fluorescence émise à chaque cycle, où la première augmentation significative dans la quantité d'amplicons est en directe corrélation avec la quantité initiale de la matrice originale cible (template).

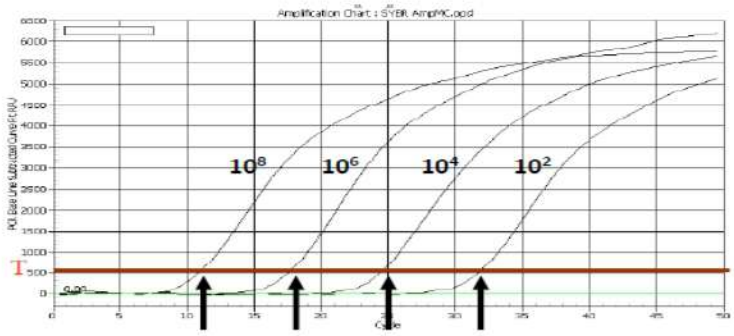


Figure 19 : Graphe montrant la corrélation entre le nombre de copie d'ADN.

Les appareils servant à la PCR en temps réel utilisent généralement un système en tubes fermés (plaques de 96 puits) et la quantification ne requiert aucune manipulation post-amplificatrice, ce qui minimise ou élimine les problèmes de contamination des amplicons suite à la réaction de PCR et réduit le temps d'analyse. Cette technologie est assez intéressante pour des applications d'analyses à grande échelle (high-throughput) due au fait que ce processus en sa totalité est du début à la fin rendant automatisé.

La fluorescence pendant cette PCR est rendue possible grâce à des agents se liant à l'ADN double brin et aux sondes nucléotidiques fluorescentes.

- ***Agents se liant à l'ADN double brin (Double-stranded DNA binding dyes: Light cycler assay)***

Les molécules qui se lient à l'ADN double brin peuvent être divisées en deux classes: les agents intercalants comme le bromure d'éthidium, le YO-PRO-1, le SYBR Green I et les agents se fixant au sillon mineur (minor groove binders) comme le Hoeschst 33258. Leur émission fluorescente augmente lorsqu'ils sont liés à l'ADN double brin. Pour être utilisés

dans une réaction de PCR en temps réel, ces agents doivent rencontrer deux exigences : augmenter en fluorescence lorsqu'ils sont liés à l'ADN double brin et ne pas inhiber la réaction de PCR.

Le colorant libre en solution exhibe peu de fluorescence lors de la réaction d'amplification par PCR. Une augmentation de la fluorescence est directement liée à la quantité de colorant se fixant à l'ADN double brin naissant durant l'étape d'élongation. Quand cela est suivi en temps réel, l'élévation du signal de fluorescence est observée durant l'étape de polymérisation et lorsque l'ADN est dénaturé à l'étape suivante, l'émission fluorescente décroît complètement. Par conséquent, l'émission de fluorescence est mesurée à la fin de chaque étape d'élongation pour chacun des cycles par un système de lecture intégré à l'appareil de PCR en temps réel qui permet de suivre l'augmentation de la quantité d'ADN amplifié durant la réaction.

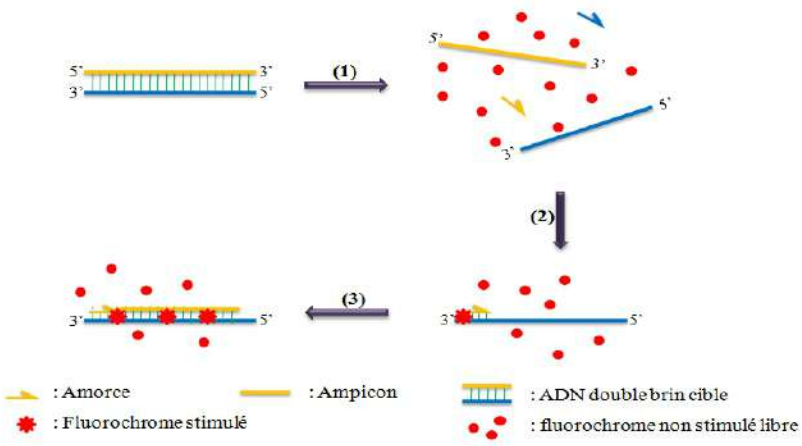


Figure 20 : Agents se liant à l'ADN double brin.

- ***Hydrolyse de sondes (Hydrolysis probes: Taqman assay)***

La technologie *Taqman* repose sur l'activité 5'-exonucléasique de la Taq polymérase, qui durant l'étape d'hybridation/extension de la PCR, peut hydrolyser les sondes s'hybridant à sa séquence cible sur l'amplicon. Un fluorochrome émetteur (reporter) tel que le FAM : 6-carboxyfluorocein se fixe à l'extrémité 5' de la sonde d'hybridation et son émission est de ce fait inhibée par un second fluorochrome suppresseur (quencher), qui lui, est présent à l'extrémité 3' tel que le TAMRA : 6-carboxy-tetramethyl-rhodamine.

Lorsque stimulé, le fluorochrome émetteur transfert son énergie au fluorochrome suppresseur voisin par le principe FRET (fluorescence resonance energy transfer) qui dissipe cette énergie sous forme de chaleur plutôt que d'émettre de la fluorescence. Les sondes libres en solution demeurent intactes et aucune fluorescence n'est émise vue que l'activité 5'-

exonucléasique de la Taq polymérase est spécifique à l'ADN double brin. Les amorces ainsi que la sonde se fixent à leurs séquences complémentaires respectives, lors de la phase d'hybridation. A l'étape suivante, la Taq polymérase débute l'élongation du nouveau brin d'ADN à partir de l'amorce jusqu'à ce qu'elle rencontre sur son passage la sonde hybridée qu'elle déplace et hydrolyse avec son activité 5'-exonucléasique. Ensuite, la molécule reportée est libérée de l'environnement du suppresseur, de sorte que l'émission de fluorescence augmente à chaque cycle proportionnellement au taux d'hydrolyse de la sonde.

La Taq polymérase n'hydrolysant la sonde que lorsqu'elle s'hybride avec sa séquence complémentaire, durant l'étape de polymérisation, les conditions de température doivent être ajustées afin de maintenir la sonde dans l'état hybridé pendant cette étape. La majorité des sondes ont une température de dissociation (Tm) autour de 70ºC ou de 5 à 10°C plus élevée que les amorces. En conséquence, la technologie *Taqman* utilise une combinaison d'étapes d'hybridation et de polymérisation à 60-62ºC pour assurer durant l'extension, l'hybridation et la stabilité de la sonde.

Ceci permet aussi une activité 5-exonucléasique maximale de la Taq polymérase mais, l'efficacité de l'activité de polymérisation de l'enzyme sera légèrement réduite à cette température sub-optimale. Pour de longs amplicons, une étape d'hybridation/polymérisation plus longue ou encore une augmentation de la concentration en Mn^{2+} ou Mg^{2+} pourrait s'avérer nécessaire pour stabiliser l'hybridation de la sonde à sa séquence cible.

Les principes à suivre dans la conception des sondes *Taqman* sont également applicables à d'autres sondes linéaires et sont inclus en tant que règles générales :

1) Une longueur de 20-40 nucléotides,

2) Un contenu en G-C variant de 40- 60%,

3) Aucun patron de séquence répétée,

4) Aucune séquence permettant une hybridation ou un chevauchement avec les amorces,

5) Un A, un C ou un T à l'extremité 5' parce qu'un G supprime la fluorescence de l'émetteur même après clivage et,

6) Une Tm de 5 à 10°C plus élevé que les amorces afin de s'assurer qu'elles s'hybrideront avant les amorces et qu'elles demeureront hybridées pendant l'étape combinée d'hybridation et de polymérisation.

Les sondes fluorescentes possèdent comme avantage par rapport aux agents se liant à l'ADN une spécificité accrue et une meilleure capacité de multiplexage. La spécificité

d'hybridation entre la sonde fluorescente et sa séquence d'ADN cible réduit significativement l'émission de fluorescence non spécifique due à de mauvais appariements ou des dimères d'amorces (primer-dimers).

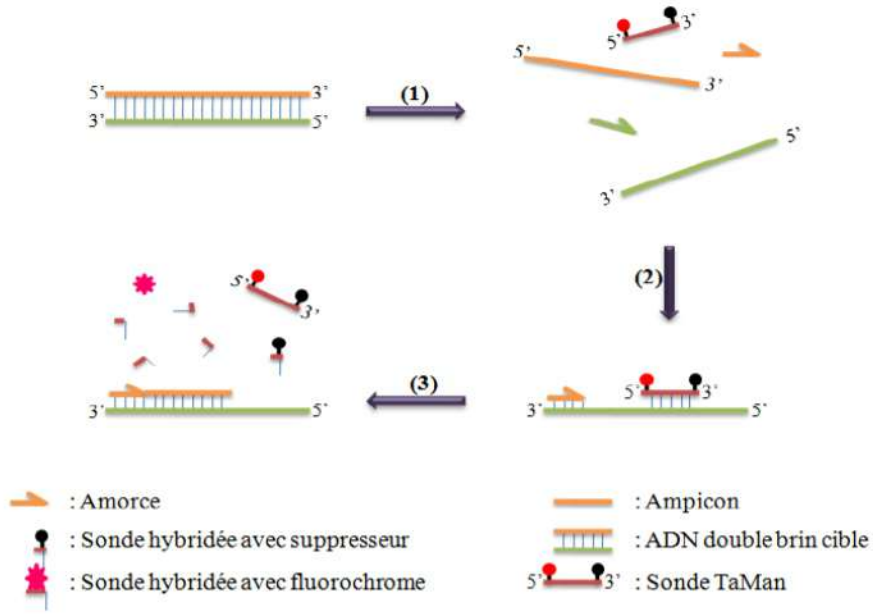


Figure 21 : Hydrolyse de sondes.

III.5.4. PCR semi-quantitative ou compétitive

Le taux d'ARN ou d'ADN d'intérêt est mesuré en tant que quantité absolue dan le cas de la PCR quantitative. Pour la PCR semi-quantitative ou de PCR compétitive, cela concerne la mesure des quantités relatives à l'aide de standards correspondant généralement à des ARN ou plus rarement à des ADN.

Ces standards peuvent être internes ou externes. Le standard est un ARN (plus rarement un ADN) qui est présent dans l'extrait d'ARN (standard interne) ou qui est rajouté en quantité connue dans le mélange réactionnel (standard externe). Le standard est amplifié en même temps que l'ARN d'intérêt. Il y a donc compétition entre l'amplification du standard et celle de l'ADN d'intérêt.

Plus la quantité de standard est importante (il s'agit néanmoins que l'amplification soit toujours en phase exponentielle) moins l'ARN d'intérêt sera amplifié et donc plus sa quantité sera faible à la fin. Bien sûr, la méthode d'analyse de l'échantillon PCR doit permettre de discriminer le standard par rapport à l'ARN d'intérêt d'une part et d'autre part d'évaluer la quantité relative d'ADN d'intérêt par comparaison avec la quantité de standard qui est connue.

Les standard internes sont des ARN endogènes, correspondant aux ARN de gènes dont l'expression est présumée constante (Actine, bêta2- microglobuline,…) et qui sont présents au sein de la population d'ARN matrices lors de la transcription inverse.

Ces standards présentent un désavantage majeur : ils nécessitent l'emploi d'amorces différentes de celles qui sont utilisées pour l'ARN d'intérêt. Par conséquent, les cinétiques d'amplification sont significativement plus différentes et il est très laborieux voir impossible de garantir une constante expression entre divers échantillons.

Les standards d'ARN externes homologues sont des ARN synthétiques qui partagent les mêmes sites d'hybridation des amorces que l'ARN d'intérêt et qui possèdent la même séquence globale, à une légère mutation, délétion ou insertion près qui vont permettre l'identification et la quantification de celui-ci par rapport au signal rendu par l'ARN d'intérêt. Ces standards permettent d'une part d'apprécier la variabilité introduite au niveau de la RT et, d'autre part, présentent globalement la même efficacité d'amplification que l'ARN d'intérêt que ce soit au niveau de la RT ou de la PCR.

Les standards d'ARN externes hétérologues sont des ARN exogènes et leur taux peut donc être contrôlé. Ils présentent toutefois, à la différence des standards externes homologues une efficacité d'amplification différente comparée à celle de l'ARN d'intérêt. Dans le cas de la RT-PCR quantitative (PCR semi-quantitative), le standard consiste en une solution titrée d'ADN de séquence identique à celle de l'ADN d'intérêt à quantifier.

Le standard entre en compétition avec l'ARN d'intérêt vis-à-vis de la polymérase et des amorces. Plus la concentration en standard augmente, plus le signal du gène d'intérêt diminue. Ici, la PCR n'a pas besoin d'être réalisée en phase exponentielle et les résultats présentent une correcte reproductibilité. Cependant, la méthode est lourde et ne permet pas de gérer beaucoup d'échantillons simultanément.

III.6. Intérêts et applications de la PCR

Les applications de la PCR sont multiples. C'est une technique actuellement incontournable en biologie cellulaire et moléculaire.

Elle permet notamment en quelques heures le « clonage acellulaire » d'un fragment d'ADN grâce à un système automatisé, alors qu'il faut plusieurs jours avec les techniques standard de clonage moléculaire.

La **PCR** est très couramment utilisée dans de nombreux domaines comme en :

- ***Médecine***

Dans les laboratoires de diagnostic moléculaire, l'emploie de la **PCR** se fait de manière routinière. Pour diagnostiquer des maladies génétiques (**myopathie**, **mucoviscidose**,…), des infections virales (**Covid-19**, **SIDA**, **Hépatite C**, **SRAS**), bactériennes (**tuberculose**) ou parasitaires (**toxoplasmose**), mais aussi des cancers.

La PCR peut être un outil très avantageux en cancérologie, le but est de détecter des **mutations** connues dans certains **gènes** spécifiques du cancer. Cette technique de diagnostic suggère que l'on connaisse le **gène** déficient et les éventuelles **mutations** qui en sont responsables. Tout comme elle peut aussi être très profitable pour le suivi des thérapies anticancéreuses et de ce fait permettre au médecin de poursuivre ou non le traitement.

Exemple

De nombreuses recherches ont été faites sur le cancer du sein (gènes BRCA1 et BRCA2), de la prostate ou de la thyroïde. Maintenant, il est admis que les mutations touchant le gène BRCA1 sont les plus courantes et prédisposent à la plupart des cancers familiaux du sein et de l'ovaire.

Les **mutations** touchant le **gène** BRCA2 sont plus fréquentes au niveau des populations anglo-saxonnes et nordiques. La recherche de **mutations** dans ces **gènes** se fait dans le cadre du dépistage de cancer dans des familles à risque, sinon cela revient beaucoup trop cher.

En dosant la quantité d'ARNm de la thyroglobuline, **protéine** hautement exprimée par les cellules tumorales de la thyroïde, les médecins peuvent suivre l'état des patients attient du cancer différencié de la thyroïde. Récemment, cette technique a été proposée comme une alternative prometteuse, au lieu d'utiliser directement le dosage de la **protéine** (thyroglobuline).

Remarque

Veuillez noter que la présence de certaines mutations au niveau de certains gènes ne provoque pas nécessairement le cancer. Il faut aussi ajouter à cela deux composantes. Premièrement, chaque personne est différente et ne réagit pas de la même manière aux **mutations**, aux traitements, etc. Deuxièmement, il existe une composante environnementale, vous avez du en entendre parler: l'amiante, certains virus, les ondes électromagnétiques (téléphone, lignes hautes tension),…

- ***Recherche fondamentale***

Dans de multiples applications routinières.

- ***Médecine légale***

Elle est aussi employée pour réaliser des empreintes génétiques, qu'il s'agisse de l'identification génétique d'une personne dans le cadre d'une enquête judiciaire, ou pour un test de paternité.

- ***Agroalimentaire***

Pour identifier des variétés ou des espèces végétales et animales, pour sélectionner de nouvelles variétés de fruits et légumes, comme la tomate, pour le contrôle de la qualité des produits agroalimentaires, par exemple déceler l'existence d'**OGM** dans un aliment donné.

Exemple

Un **OGM** est un organisme génétiquement modifié. C'est un organisme auquel les propriétés génétique ont été changées par addition d'un **gène** ou de plusieurs **gènes** spécifiques, donnant ainsi à l'organisme de nouvelles caractéristiques (par exemple le **gène** de résistance à un parasite ou à un herbicide). Ces **gène**s ajoutés sont appelés des **transgènes**.

Des nombreux laboratoires se sont spécialisés dans la recherche d'**OGM** dans de nombreux produits à la base de notre alimentation (Maïs, soja, farine, semoule, gluten, corn flakes, amidons et dérivés, extrait protéique, sirop de glucose, lait de soja, tourteau, lécithine,...).

Après avoir extrait l'**ADN** des produits, ils font plusieurs **PCR** en utilisant différents couple d'**amorce**s spécifiques pour un transgène connu. S'il n'y a pas us transgénèse dans le produit, la **PCR** sera négative du fait que les **amorces** ne s'hybrident pas sur l'**ADN**. Au contraire, si le transgène est présent, il sera détectable par l'obtention d'un produit d'amplification et la **PCR** est alors positive.

- ***Histoire***

La PCR peut intervenir dans l'étude concernant l'**évolution des espèces**. La divergence des espèces à partir d'un ancêtre commun est reflétée par le degré de divergence entre les séquences nucléotidiques des gènes. La détermination du lien de parenté entre différentes espèces peut se faire par la mesure de la divergence nucléotidique.

La PCR est un bon moyen de procéder à des **études phylogénétiques** en recherchant des liens de parenté entre les individus (Cas les plus célèbres: les enfants du tsar Nicolas II, ou ceux de Louis XVII), sur des squelettes fossiles (ADN fossile),…. Utilisée pour retracer les éventuelles migrations des populations humaines et animales (les indiens d'Amérique, en Islande), pour détecter sur des momies égyptiennes et andines les probables infections bactériennes, virales et parasitaires.

Exemple

Il a été établi par H-C Li et son équipe, la présence dans des momies de la Cordillère des Andes datant de plus de 1500 ans, du virus HTLV-1, à l'origine du **SIDA**.

Ces dernières années, plusieurs études ont porté sur la famille Romanov. Selon l'histoire officielle de l'URSS, le dernier tsar, Nicolas II, aurait péri la nuit du 16 juillet 1918 avec toute sa famille, son épouse et ses 5 enfants, ainsi que quelques serviteurs.

Cependant, en 1920, une jeune femme, connue sous le nom de ***Anna Manahan Anderson***, prétend être la plus jeune des filles du tsar, la grande duchesse Anastasia, qui aurait échappé au massacre. Si bien des gens la soupçonnèrent d'imposture, elle fut reconnue comme Anastasia par des membres de la famille impériale, comme la princesse Irène (La tante d'Anastasia) ou le grand duc Alexandre (Cousin de Nicolas II).

La possibilité de séquencer l'**ADN** ancien allait permettre d'établir la vérité. Anna était morte depuis plusieurs années, mais on possédait une biopsie intestinale d'elle. On put en extraire de l'**ADN** et séquencer des fragments d'***ADN mitochondrial***. Ces séquences furent comparées à celles de membres vivants de la famille de la tsarine (Dont le duc d'Edimbourg). Il fut ainsi prouvé qu'Anna Manahan ne pouvait en aucun cas être la fille de la dernière tsarine.

Selon l'histoire officielle, le jeune roi Louis XVII, fils de Louis XVI et de Marie-Antoinette, est mort au Temple, le 8 juin 1795. Mais la rumeur se propagea rapidement que l'enfant mort au Temple n'était pas Louis XVII, mais une « doublure » qu'on lui avait substituée pour mettre à exécution un plan d'évasion.

Plusieurs prétendants au nom de Louis XVII se firent connaître, immédiatement après la restauration. On en compta même jusqu'à New York, aux Seychelles, aux Açores et en Dalmatie! Le prétendant le plus sérieux fut Naundorff, qui se fit reconnaître comme étant Louis XVII au début de la monarchie de juillet par quelques vieux serviteurs de Versailles ou des Tuileries, notamment par Madame de Rambaut, gouvernante de Louis XVII.

Son cas ne fut tranché que très récemment, grâce à son **ADN** (Prélevé sur son squelette). Et à l'**ADN** extrait du cœur de l'enfant mort au Temple que le médecin responsable avait conservé. La séquence des bases de certaines zones de l'***ADN mitochondrial*** fut établie pour les deux sources et comparée avec la séquence correspondante établie chez des descendantes de Marie-Antoinette.

Il fut alors clairement montré qu'il était impossible que Naundorff soit le fils de Marie-Antoinette, mais que, par contre, l'enfant mort au Temple était très probablement Louis XVII.

La déficience visuelle de John Dalton, découvreur du daltonisme, ont pu être analysés en 1995, cent cinquante ans après sa mort, celle-ci était liée à l'absence du gène MW-opsine.

Manipulation et étude du génome : techniques de base

« Séquençage de l'ADN »

Objectifs spécifiques

Au terme de ce cours qui traite des différentes techniques de séquençage de l'ADN, vous devez être capable de :

- Comprendre ce qu'est un séquençage de la molécule d'ADN ;
- Comprendre quel est le principe de cette technique ;
- Connaitre quelles sont les différents variantes du séquençage de l'ADN ;
- Distinguer les différences entre les procédés et quand les appliquer.

De nombreux génomes, y compris le génome humain, sont à l'heure actuelle entièrement séquencés, et d'avantage de génomes sont en cours de l'être. Un grand nombre de laboratoires, aussi bien publics que privés, participent à cet énorme effort qui révolutionne tant la biologie fondamentale que les biotechnologies.

IV.1. Définition

Le séquençage de l'ADN est une technique qui consiste à déterminer l'ordre d'enchaînement des nucléotides d'un fragment d'ADN donné. C'est aujourd'hui une technique de routine pour les laboratoires de biologie moléculaire. Ce procédé fait appel aux connaissances acquises sur les mécanismes de réplication de la molécule l'ADN au cours des trois dernières décennies.

IV.2. Historique

Cette méthode d'étude des acides nucléiques a révolutionné la biologie moléculaire dans les années 1970, soit seulement une vingtaine d'années après l'établissement de la structure de l'ADN par Watson et Crick.

Au milieu des années 70, les premières méthodes de séquençage ont été parallèlement développées. En 1980, la technique de Sanger (Grande-Bretagne) et celle de Gilbert (Etats-Unis) ont été récompensées toutes deux d'un prix Nobel de chimie. Le premier organisme a été séquencé en 1977. Il s'agissait du bactériophage X174, possédant un ADN simple brin ne nécessitant donc pas l'étape de dénaturation utilisé dans les méthodes de Sanger ou Maxam et Gilbert.

IV.3. Importance du séquençage

Le séquençage d'un nombre impressionnant de génomes (humain ainsi que ceux de nombreux agents infectieux, de mammifères et de plantes) a permis de modifier

considérablement les recherches biomédicales et biologiques en ouvrant de larges horizons dans le domaine de la médecine (prédiction, pronostic, diagnostic, thérapeutique, prévention, ...) et dans de multiples autres disciplines biologiques (agronomie, anthropologie, environnement, ...). Il permet également une meilleure compréhension des cycles de la vie dans leur globalité.

IV.4. Différents acteurs impliqué dans le séquençage

Comme pour la réplication, la transcription, la traduction ou encore la PCR, le séquençage nécessite également différents acteurs afin qu'il puisse s'établir. Les éléments nécessaires à cette technique sont :

- **ADN**

Il provient des organismes dont on souhaite séquencer le génome. L'ADN, le plus souvent sous forme double-brin, est dénaturé afin de séparer les deux brins. Celui qui sera séquencé s'appelle le brin « matrice ».

- **Nucléotides**

Ou plus précisément, les désoxynucléotides, qui forment les briques de l'ADN (A, C, G ou T). Ils sont attachés les uns aux autres grâce à des liaisons chimiques.

- **Didésoxynucléotides**

Ce sont des nucléotides privés du groupement OH en position 3'. Leur incorporation dans une chaîne d'ADN interrompt définitivement la synthèse de l'ADN.

- **Amorce**

C'est un brin d'ADN très court (une vingtaine de nucléotide), qui peut s'hybrider à une séquence complémentaire spécifique.

- **ADN polymérase**

C'est une enzyme qui a pour rôle de copier l'ADN, en synthétisant un brin complémentaire au brin matrice. L'ADN polymérase ne peut fonctionner qu'en additionnant des nucléotides à une amorce déjà existante, en suivant la succession de nucléotides du brin matriciel.

- **Séquenceur**

C'est un appareil qui permet de réaliser les différentes étapes de cette technique de façon automatique.

IV.5. Etapes du séquençage

Comme pour la réaction d'amplification de l'ADN via la PCR, le séquençage de l'ADN passe également par plusieurs étapes qui sont les suivantes :

Pour le séquençage, tout se passe initialement dans un tube à essai, en présence des acteurs de la synthèse d'ADN : de l'ADN à séquencer, des nucléotides, une amorce et de l'ADN polymérase. La présence de chacun de ces acteurs doit être en excès. Du cout, la réaction de séquençage implique plusieurs réactions.

Tant que les bases (A, C, G ou T) sont respectées dans le milieu, l'ADN polymérase utilisera de façon aléatoire les nucléotides présents pour synthétiser un ADN de séquence complémentaire en dupliquant le brin (ou fragment, ou chaîne) matrice.

Quand l'ADN polymérase sélectionne aléatoirement un didésoxynucléotide (ce qui est assez rare du fait qu'il y en a moins que les nucléotides) en l'incorporant à la chaîne en synthèse, celle-ci s'interrompt brusquement. Comme chaque didésoxynucléotide est marqué par un fluorochrome distinct (G jaune, C bleu, A vert et T rouge), la chaîne se terminant par exemple par un G sera jaune.

En raison du grand nombre de réactions de synthèse se déroulant dans le tube réactionnel, statistiquement, il existe des chaînes de différentes tailles ; correspondant à un arrêt de la synthèse à chaque nucléotide ; ainsi que de nombreux fragments de même taille. Ces chaînes débutent toutes à la même position sur l'ADN matriciel (établie par l'amorce utilisée), de sorte que tout brin de même longueur se termine par le même didésoxynucléotide marqué.

Les chaînes d'ADN obtenues peuvent alors être séparées en fonction de leur taille, en présence d'un courant électrique sur un gel d'acrylamide. Tous les fragments partageant la même taille migrent à la même distance et plus ces fragments sont courts, plus ils migrent loin sur le gel. On parvient alors à avoir une succession de bandes colorées, chacune concordant avec le dernier nucléotide ajouté à la chaîne. Afin de connaître l'ordre des nucléotides, c'est-à-dire la séquence de l'ADN traité, il suffit de lire la succession des couleurs, étape assurée automatiquement par les détecteurs du séquenceur.

IV.6. Méthodes de séquençage

En biologie moléculaire, il est souvent utilisé les termes de séquences génomiques ou séquences d'ADN, qui sont données pour des raisons algorithmiques sous forme de chaînes de caractères. Comment ces séquences, ces chaînes de caractères, sont-elles obtenues ?

Depuis 1970, plusieurs méthodes de séquençages peuvent être utilisé afin d'étudier et d'analyser un ADN, un génome donné. Le processus de séquençage repose actuellement sur la technique enzymatique de Sanger.

IV.6.1. Méthode de Sanger

C'est une méthode par synthèse enzymatique inventée en 1977 par **Frederick Sanger** (Angleterre). L'idée principale est qu'en copiant les brins d'ADN et en surveillant quels nucléotides sont ajoutés, l'un après l'autre, on peut retrouver la séquence de nucléotides.

Cette méthode génère des fragments d'ADN terminés par l'une des quatre bases marquées au préalable, car l'élongation de la chaîne s'arrêtera après l'incorporation d'un didésoxynucléotide (ddNTP). C'est le principe sur lequel se base cette méthode.

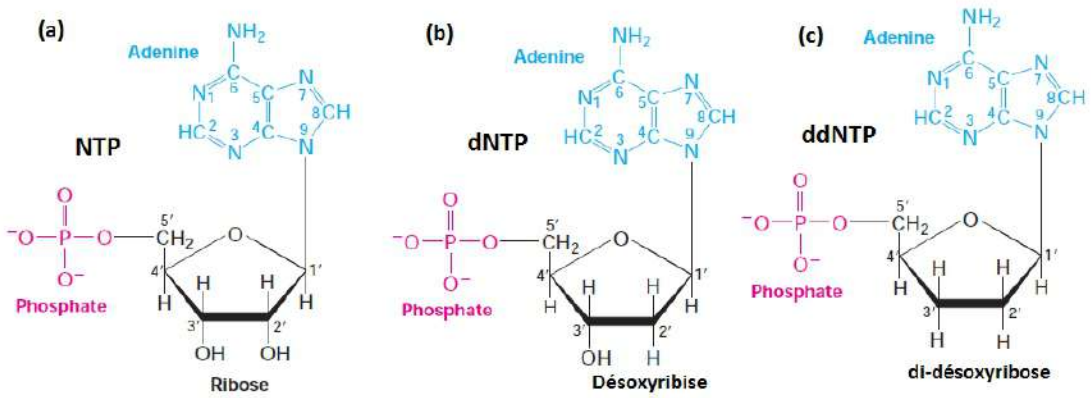


Figure 22 : Différentes formes des nucléotides.

La méthode de Sanger permet de déterminer une séquence inconnue par l'intermédiaire d'une copie d'ADN simple brin. Réalisé grâce à l'ADN polymérase et une amorce spécifique à la séquence cible.

Il y'a quatre différentes réactions de séquençage qui sont effectuées dans quatre tubes distincts. Il est introduit en parallèle dans chaque réaction des brins d'ADN cibles dénaturés en grand nombre, des amorces, la polymérase et les nucléotides normaux (dATP, dTTP, dCTP et dGTP). En plus de tout cela, il y'a l'ajout dans chaque tube d'un nucléotide modifié différent ou didéshydro- nucléotide (ddATP, ddTTP, ddCTP et ddGTP).

Ceux-ci sont identiques aux bases classiques à la différence qu'il leur manque un groupement -OH (fig.22) nécessaire à l'assimilation du nucléotide suivant par la polymérase (liaison phospho-diester) et une fois incorporé, constitue la fin d'une chaîne d'ADN, ce qui permet d'identifier la base finale.

La quantité de nucléotides modifiés incorporés dans chaque tube est bien inférieure au nombre de nucléotides classiques afin de ne pas stopper l'assimilation trop tôt et qu'il y'ait de ce fait toutes les molécules possibles synthétisées.

Ainsi, 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. Il faut savoir que les réactions d'amplification avec les ddNTP sont rapides (moins de 15 minutes).

Par exemple, dans la réaction où on a ajouté du ddGTP, la synthèse s'arrête au niveau des G. La terminaison se fait de manière statistique suivant que l'ADN polymérase utilise l'un ou l'autre de ces nucléotides. Il en résulte un mélange de fragments d'ADN de tailles croissantes, qui se terminent tous au niveau d'un des G dans la séquence.

Le but du protocole est de lire les résultats. Pour cela, il faut placer le contenu des tubes dans quatre puits distincts d'un gel d'électrophorèse, correspondant aux quatre nucléotides possibles. L'ADN étant chargée négativement, plus la taille de la séquence est importante et plus la molécule migre vers le pole positif du gel d'électrophorèse. Dans ce gel après leurs migrations, on peut facilement déduire l'ordre des nucléotides de la séquence concernée. Une simple lecture horizontale de ce gel (une fois révélé) permet de connaître l'ordre des bases de la séquence.

Afin de voir les fragments d'ADN sur le gel d'électrophorèse en polyacrylamide, ils sont exposés à un film photographique ce qui fait que des bandes sombres apparaissent là où se trouvait de l'ADN sur le chromatogramme.

Des avancées ultérieures de cette méthode ont incorporé l'utilisation de **fluorophores**, qui sont de petits composés chimiques dégageant des lumières colorées. Le séquençage peut être réalisé en une seule réaction avec une seule colonne de gel pour représenter les brins d'ADN et cela en ajoutant un fluorophore coloré différent à chaque nucléotide; la couleur de la bande indique la base située à l'extrémité du fragment d'ADN. Cette innovation a permis l'automatisation de la méthode (les ordinateurs et les machines font le travail).

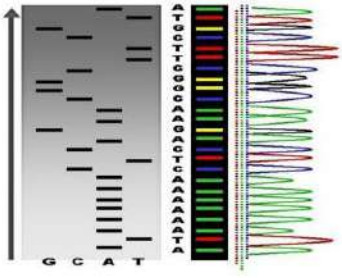


Figure 23 : Electrophorèse en gel standard et d'un séquençage à l'aide de fluorophores.

Remarque

Le séquençage de Sanger est devenu la méthode de choix et elle fait aujourd'hui partie de la routine dans la plupart des laboratoires biologiques.

IV.6.2. Méthode de Maxam et Gilbert

Contrairement à la méthode précédente (enzymatique), celle-ci est basée sur une dégradation chimique de l'ADN permettant de couper le brin suite à un type de base (coupures sélectives). Compte tenu du nombre important de copies présentes dans le tube réactionnel, statistiquement on obtient toutes les possibilités. De ce fait, on peut revenir à la séquence des nucléotides de l'ADN correspondant en reconstituant l'ordre des coupures. On peut décomposer ce séquençage chimique en six étapes successives :

- **Marquage**

Les extrémités des deux brins d'ADN à séquencer sont marquées par un traceur radioactif (^{32}P). Cette réaction s'effectue en général au moyen du dATP radioactif et de polynucléotide kinase.

- **Isolement du fragment d'ADN à séquencer**

A l'aide d'une électrophorèse sur gel de polyacrylamide, ce fragment d'ADN est séparé. Ensuite, celui-ci est découpé du gel et récupéré par diffusion.

- **Séparation de brins**

Une dénaturation thermique est effectuée sur les deux brins de chaque fragment d'ADN afin de les séparés, puis par une nouvelle électrophorèse, ils sont purifiés.

- **Modifications chimiques spécifiques**

Les ADN simple-brin subissent de spécifiques réactions chimiques des divers types de base. **Walter Gilbert** a mis au point plusieurs types de réactions spécifiques, effectuées en parallèle sur une fraction de chaque brin d'ADN marqué : par exemple, une réaction pour les G (alkylation par le sulfate de diméthyle), une réaction pour les G et les A (dépurination), une réaction pour les C, ainsi qu'une réaction pour les C et les T (hydrolyse alcaline). Ces différentes réactions sont réalisées dans des conditions très modérées, de telle sorte que chaque molécule d'ADN ne porte en moyenne que zéro ou une modification.

- **Coupure**

Suite à ces réactions, l'ADN est coupé par réaction avec une base, la pipéridine, au niveau de la modification.

- **Analyse**

Une séparation par électrophorèse en conditions dénaturantes est réalisée pour les produits obtenus des diverses réactions puis, analysés pour reconstituer la séquence de l'ADN et ce pour chaque fragment. Celle-ci est similaire à l'analyse effectuée pour la méthode de Sanger.

Remarque

La méthode de l'électrophorèse suit ensuite celle de Sanger. Moins facile à robotiser que la méthode de Sanger, cette technique est aujourd'hui très peu utilisée dans les milieux industriels.

IV.6.3. Séquençage du génome entier

La connaissance de la structure d'un génome dans sa totalité peut passer par son séquençage. Toutefois, due à l'importante taille des génomes qui peut atteindre plusieurs millions de bases (ou mégabases), il est nécessaire de combiner les méthodes de biologie moléculaire avec celle de l'informatique afin d'être capable de traiter une aussi grande quantité de données.

Deux grands principes de séquençage du génome complet peuvent être employés. Dans les deux cas, l'ADN génomique est préalablement fragmenté par des méthodes enzymatiques (enzymes de restriction) ou physiques (ultrasons) :

IV.6.3.1. Méthode de séquençage par ordonnancement hiérarchique

Elle consiste à classer les fragments génomiques obtenus avant de les séquencer. Suite à l'extraction, l'ADN génomique est découpé en fragments de 50 à 200 kb par sonication, suivie d'un clonage dans un vecteur approprié tel que les chromosomes bactériens artificiels ou BAC.

Le nombre de clones doit couvrir de 5 à 10 fois la totalité de la longueur du génome analysé. Le chevauchement et l'ordonnancement des clones sont réalisés soit par hybridation de sondes spécifiques, soit par analyse des profils de restriction, soit plus fréquemment par un ordonnancement après séquençage et hybridation des extrémités des BAC. Après trie des clones, ces derniers sont fragmentés et séquencés séparément, par la suite assemblés par alignement bio-informatique.

Les avantages de cette méthode sont une plus grande facilité d'assemblage des fragments grâce aux chevauchements des BAC, la possibilité de comparer les fragments aux banques de données disponibles, et l'éventualité de partager le séquençage avec de multiples laboratoires, chacun d'eux en charge d'une région chromosomique.

Le principal inconvénient réside dans la difficulté à cloner des fragments contenant des séquences répétées assez fréquentes dans certains génomes, tel que ceux des mammifères, ce qui rend l'analyse bio-informatique finale difficile.

IV.6.3.2. Méthode globale (ou *whole-genome shotgun*)

Il s'agit d'une méthode de séquençage d'ADN génomique initialement imaginée dans le laboratoire de **Frederick Sanger** à Cambridge à la fin des années 1970 pour séquencer les premiers génomes de virus.

Cette méthode ne fait pas de classement des fragments génomiques obtenus mais elles sont obtenues dans un ordre aléatoire. Un traitement bio-informatique faisant intervenir un assembleur permet de réordonner par chevauchement les séquences communes des fragments génomiques.

La principale variante entre ces deux principes est que l'ordonnancement hiérarchique tente d'aligner et comparer un jeu de clones de grande taille (~ 100 kb), tandis que dans la méthode globale, le génome total est réduit en petits fragments qui sont d'abord séquencés puis alignés.

IV.6.4. Séquençage par hybridation

Le séquençage par hybridation est basé sur l'emploie de puces à ADN disposant de plusieurs centaines (pour les puces de première génération) à plusieurs milliers d'oligonucléotides. Il y'a clivage de l'ADN à analyser en plusieurs fragments, par la suite incubés sur la puce où ils vont s'hybrider par complémentarité avec les oligonucléotides. La détection des oligonucléotides hybridés, à savoir la lecture de la puce, favorise l'obtention du spectre de la séquence d'ADN, c'est-à-dire sa constitution en sous-séquences de n nucléotides, où n représente la taille des sondes sur la puce utilisée. Ensuite, le traitement du spectre par voix informatique permet de reconstituer l'intégralité de la séquence.

IV.6.5. Séquençage haut débit (HTS)

On désigne par séquençage haut débit (HTS pour *High-Throughput Sequencing*) aussi appelé NGS pour *Next-Generation Sequencing*, un ensemble de méthodes apparues à partir de 2005 telles qu'Illumina® (Solexa), Roche 454, Ion torrent : Proton/PGM, SOLiD,… produisant des millions de séquences en un *run* et à faibles coût.

Elles se caractérisent par l'utilisation d'approches massivement parallèles, permettant de séquencer des centaines de milliers de fragments simultanément, beaucoup plus rapidement que les méthodes conventionnelles tel que le séquençage de Sanger et fonctionnent en 3 étapes :

- La première consiste en la préparation et l'amplification des molécules d'ADN à analyser.
- La seconde permet l'incorporation des bases complémentaires du brin à séquencer.
- Enfin la dernière étape comprend la lecture de la séquence proprement dite.

Quasi simultanément, on voix apparaitre trois technologies: le séquençage avec des terminateurs réversibles, le pyroséquençage, et le séquençage par ligation. La technologie Illumina® étant la NGS, la plus usuelle. Cette technologie utilise l'amplification clonale et le séquençage par synthèse. Le processus permet d'identifier simultanément les bases d'ADN lorsqu'elles sont incorporées dans la chaîne d'acide nucléique. Chaque base émet un signal de fluorescence unique lorsqu'elle est ajoutée au brin en cours de synthèse, ceci est utilisé pour déterminer la séquence d'ADN.

La technologie NGS peut être employée afin de séquencer l'ADN de n'importe quel organisme, procurant à quasiment n'importe quelle question biologique, des informations inestimables. Grâce à une variété de méthodes et en tant que procédé hautement évolutive, le séquençage de la molécule d'ADN peut être applicable à de courtes régions ciblées ou à la totalité du génome.

Manipulation et étude du génome : techniques de base

« Méthodes de recherche et de traitement de données »

Objectifs spécifiques

Au terme de ce cours qui traite des différentes méthodes de recherche et de traitement de données, vous devez être capable de :

- Comprendre comment faire une recherche de données en biologie moléculaire ;
- Quels sont les différents moyens pouvant être utilisé et où trouver ces données ;
- Connaitre quelles sont les différents procédés de traitements de données ;
- Comprendre et pouvoir appliquer les différentes étapes de traitement de données ;
- Distinguer les différences entre les procédés et quand les appliquer.

Les biotechnologies et le génie génétique sont des technologies de pointe dont les applications dans les divers domaines des sciences de la nature et de la vie ne sont plus à démontrer. Ces technologies utilisent intensivement les outils issus de la bioinformatique.

V.1. Définition

La bioinformatique est une discipline de recherche formé d'un ensemble de concepts et techniques permettant l'acquisition, l'analyse et l'interprétation des données et informations biologiques contenue dans des molécules biologiques telles que les séquences nucléotidiques et protéiques, au moyen de méthodes informatiques, afin de créer de nouvelles connaissances en biologie. Les deux principales fonctions de la bioinformatique sont donc l'organisation et l'analyse des données biologiques à l'aide de ressources informatiques.

Cette branche a la particularité de regrouper, autour d'un même thème, plusieurs disciplines des sciences ou de la médecine: la biologie (Biochimie, génétique, génomique structurale et fonctionnelle,...), l'informatique, les mathématiques, la physique, la chimie,... Les informaticiens ne sont pas de simples prestataires au service de la biologie mais des interactions fortes entre les disciplines sont nécessaires pour résoudre les problèmes posés par le traitement de l'information biologique. Le bioinformaticien actuel doit savoir jongler avec beaucoup de spécialisations : les probabilités, les propriétés énergétiques des repliements de molécules, l'algorithmique, la génétique, ...

En 2001, elle a été définie par le NCBI comme étant "*Bioinformatics is the field of science in which computer science, biology, and information technology merge into a single discipline*."

V.2. Historique

Face à l'accumulation des données de séquençage des génomes à partir des années 80, l'outil informatique est devenu indispensable, apportant puissance de calcul et possibilité de mise en commun des données, via le développement et l'accessibilité de l'internet ainsi que de l'outil informatique à la communauté scientifique.

La bioinformatique prend ses origines au début des années 1980 lorsque le laboratoire européen de biologie moléculaire (EMBL : European Molecular Biology Laboratory) et le département américain de la santé (NIH : National Institute of Health) ont créé les banques de données EMBL et GENBANK afin de répertorier les séquences d'ADN découvertes par les biologistes. La bioinformatique c'est donc vue progresser et évoluer ces dernières années avec l'évolution des technologies :

- **1955-1965** : Premiers langages informatiques, premier ordinateur commercial ;
- **1970** : 1er programme pour la comparaison de séquences protéiques, alignement optimal entre deux séquences (Needleman & Wunsh) ;
- **1971** : PDB - Protein Data Bank (structures 3D des macromolécules) ;
- **1978** : Matrice de substitution (PAM) (Dayhoff *et al.*) ;
- **1981** : Similarités de séquences dans les banques (Smith & Waterman)
- **1980/1986** : Création de banques de données: EMBL (1980), GenBank (1982), DDBJ (1986), SwissProt (1986) ;
- **1989** : L'internet est diffusé pour le publique ;
- **1990** : BLAST – Simulation de séquence dans les banques ;
- **2000** : - Séquençage du 1er génome de plante, *Arabidopsis thaliana* ;
  - Publication du "draft" de la première carte complète du génome Humain (3000 MB).
- **2001** : Publication des travaux de séquençage du génome humain presque complet.

V.3. Démarche

Cette discipline hybride qui utilise tout le potentiel de traitement de l'informatique : modèles théoriques, algorithmes et programmes, bases de données, ordinateurs, réseau internet, protocoles de communication, langages, ... possède une logique et une démarche scientifique qui peut être subdivisé en trois qui sont :

V.3.1. Compilation et organisation des données biologiques dans des bases de données

On distingue des bases de données généralistes (elles contiennent le plus d'information possible sans expertise très poussée de l'information déposée) ainsi que des bases de données spécialisées autour de thèmes précis.

V.3.2. Traitements systématiques des données

L'un des objectifs est de repérer et de caractériser une fonction et/ou une structure biologique importante. Les résultats de ces analyses forment des données biologiques nouvelles acquises "*in silico*[4]".

V.3.3. Elaboration de stratégies

Elle permet d'apporter des connaissances biologiques supplémentaires en combinant les données biologiques initiales et les données biologiques obtenues "*in silico*". Avoir des connaissances permettant, à leur tour, de développer de nouveaux concepts en biologie et acquérir des concepts qui, pour être validés, peuvent nécessiter le développement de nouvelles théories et outils en mathématiques et en informatique.

V.4. Méthodes de recherche de données

Il est clairement évident qu'à l'heure actuelle, la quantité de données biologiques augmente considérablement et assez rapidement, il est donc essentiel pour un scientifique, un biologiste de savoir accéder à cette information. Des outils de recherches sont mis à disposition pour les différentes disciplines de la biologie ainsi que de la médecine et de nombreuses banques de données notamment en biologie moléculaire ont été crée, qu'elles soient généralistes ou spécialisées.

V.4.1. Banques de données biologiques (Base de données)

Les banques de données ou bases de données biologiques, disponibles (le plus souvent) sur le Web, sont des sorte de bibliothèques en ligne répertoriant des informations et données de séquences biologiques issue de divers champs de recherche tels que la génomique, la protéomique, la métabolomique[5], et la phylogénétique, collectées grâce à des expériences scientifiques et à la littérature largement diffusées par le réseau internet. Elles sont généralement reliées entre elles par des liens.

Toutes les informations connues sont mises à disposition des chercheurs du monde entier le plus rapidement possible. Il s'agit non seulement des séquences nucléiques ou protéiques brutes (successions de caractères) mais également de toutes les annotations des séquences et autres informations connexes. Actuellement, de part le monde, trois grandes

[4] C'est une locution adverbiale latine désignant une recherche ayant été effectuée par des modèles informatiques.

[5] C'est une science très récente qui étudie l'ensemble des métabolites primaires (sucres, acides aminés, acides gras, ...) et des métabolites secondaires dans le cas des plantes (polyphénols, flavonoïdes, alcaloïdes, ...) présents dans une cellule, un organe, un organisme.

banques centralisent l'information. Il s'agit notamment de : EMBL, GenBank et DDBJ (Japon).

Tableau III : Principaux serveurs concernant les bases de données biologiques.

Base de données	Origine	Informations disponible	Site web
EMBL	Allemande	Banques de séquences	https://www.embl.de/services/library/
GenBank	Américaine		https://www.ncbi.nlm.nih.gov/genbank/
DDBJ	Japonaise		https://www.ddbj.nig.ac.jp/index-e.html
FlyBase	États-Unis Angleterre	Séquences génomiques	https://flybase.org/
SGD			https://www.yeastgenome.org/
TAIR	États-Unis		https://www.arabidopsis.org/
VectorBase	États-Unis		https://www.vectorbase.org/
WormBase			https://wormbase.org//#012-34-5
PDB	Européenne Japonaise Américaine	Structure protéique et nucléotidique	https://www.uniprot.org/database/DB-0070
InterPro	Européenne	Séquences protéiques et fonction	https://www.ebi.ac.uk/interpro/
UniProt	Européenne	Séquences protéiques	https://www.uniprot.org/
SWISS-PROT	Suisse		http://www.ebi.ac.uk/swissprot/
PIR	États-Unis		https://proteininformationresource.org/

- **EMBL (European Molecular Biology Librery)**

La bibliothèque de données de l'EMBL fait partie du laboratoire Européen de biologie moléculaire à Heidelberg, en Allemagne. Elle a été créée en 1980 et son rôle principal est de maintenir et de diffuser une base de données de séquences nucléotidiques. Il est également impliqué dans la maintenance d'autres bases de données biologiques telles que la base de données de séquences protéiques SWISS-PROT et distribue d'autres bases de données d'intérêt pour les biologistes moléculaires.

- **GenBank**

C'est une banque de séquences d'ADN américaine, comprenant des séquences nucléotidiques ainsi que leur traduction en protéines à accès libre de plus de 300000 organisme. Elle a été créée au Centre national pour l'information biotechnologique (NCBI) dans le cadre de la collaboration internationale sur le séquençage des nucléotides (INSDC) en 1970. GenBank se construit soit par des dépôts directs en provenance de laboratoires (chercheurs), soit des dépôts en masse des centres de séquençage à grande échelle.

- **DDBJ (DNA Data Bank of Japan)**

C'est une base de données biologique qui rassemble des séquences d'ADN. Situé à l'institut national de génétique de la préfecture de Shizuoka au Japon, elle est également membre de la collaboration internationale de bases de données sur les séquences nucléotidiques (International Nucleotide Sequence Database Collaboration) ou INSDC. Cette base de données en ligne (https://www.ddbj.nig.ac.jp/index-e.html) est accessible à tout scientifique à travers le monde. Elle a commencé ses activités de banque de données en 1986 à l'institue national de génétique (National Institute of Genetics) NIG et reste la seule banque de données de séquences nucléotidiques en Asie.

- **FlyBase**

C'est la principale base de données génétiques et génomiques spécifique intégrées consacrée aux *Drosphilidae* notamment à la mouche du vinaigre '*Drosophila melanogaster*' , l'organisme modèle car la plus étudiée dans les différentes recherches scientifiques ainsi qu'à une douzaine d'autres espèces de diptères (un ordre de la classe des insectes tels que mouches, syrphes, moustiques, taons, moucherons,…) dont le génome est entièrement séquencé. Les informations contenues dans FlyBase proviennent de diverses sources allant des projets de génome à grande échelle à la littérature de recherche primaire.

Le projet FlyBase a été initié en 1989 en impliquant un consortium de chercheurs et de bio-informaticiens de l'université de Harvard, de l'université de New Mexico et de l'université de l'Indiana (États-Unis) ainsi que de l'université de Cambridge en Angleterre. FlyBase donne donc accès à l'ensemble des informations génomiques et bibliographiques concernant *Drosophila melanogaster*, différentes collections de mutants ainsi que de nombreux outils.

- **SGD (Saccharomyces Genome Database)**

SGD est une base de données scientifique de biologie moléculaire et génétique (génomique) consacrée à l'organisme modèle *Saccharomyces cerevisiae* (appelée également levure de boulanger ou levure de bière). Elle donne un libre accès à l'ensemble des informations génomiques, ses caractéristiques chromosomiques, bibliographiques, leurs fonctions et leurs interactions concernant *Saccharomyces cerevisiae* ainsi que ses mutants (http://www.yeastgenome.org).

- **TAIR (The Arabidopsis Information Resource)**

C'est une base de données dédiée à *Arabidopsis thaliana* (ou Arabette des dames). De sa création en 1999 à 2013, TAIR a été principalement financé par la National Science Foundation. Cependant, actuellement elle est gérée par une organisation à but non lucratif, Phoenix Bioinformatics. L'objectif de TAIR est de fournir une base d'information génomique (génome, gènes, produits géniques, variants naturels, allèles mutants et phénotypes) et bibliographique qui répond non seulement aux besoins de la communauté *Arabidopsis* mais de la communauté de recherche biologique dans son ensemble, ce qui nécessite un accès facile aux informations d'*Arabidopsis* ainsi que d'une collection de mutants pour utiliser au maximum cette plante modèle pour résoudre les problèmes de recherche dans d'autres organismes, y compris les espèces végétales économiquement importantes.

- **VectorBase**

Cette banque de données génomique est l'un des cinq centres de ressources bioinformatiques, fondé pour la première fois en juin 2004 financés par l'institut national des allergies et des maladies infectieuses (National Institute of Allergy and Infectious Diseases (NIAID)), une composante des institues national de la santé (National Institutes of Health (NIH)), qui est une agence du département de la santé et des services sociaux des États-Unis. VectorBase se concentre sur les invertébrés vecteurs d'agents pathogènes humains, y compris un certain nombre de non-vecteurs pour l'analyse comparative. Cette base est alimentée en séquences via la collaboration avec les centres de séquençage et la communauté de recherche pour gérer les génomes de vecteurs (principalement l'annotation du génome).

Les données hébergées vont des assemblages génomiques avec des caractéristiques géniques annotées, des données de transcription et d'expression des protéines à la génétique des populations, y compris les phénotypes de variation et de résistance aux insecticides.

- **WormBase**

C'est une base de données biologique en ligne sur la biologie et le génome de l'organisme modèle de nématode '*Caenorhabditis elegans*' et contient des informations sur d'autres nématodes apparentés. WormBase est utilisé par la communauté de recherche de *C. elegans* à la fois comme source d'information et comme lieu de publication et de distribution de leurs résultats. Fondé en 2000, le WormBase consortium est dirigé par Paul Sternberg (CalTech), Matt Berriman (The Wellcome Trust Sanger Institute), Kevin Howe (EBI) et Lincoln Stein (The Ontario Institute for Cancer Research).

- **PDB (Protein Data Bank)**

BDP est une banque de données sur les protéines établie par la recherche collaborative en bioinformatique structurale (*Research Collaboratory for Structural Bioinformatics*). C'est donc principalement une collection mondiale de données sur la structure tridimensionnelle (ou structure 3D) « biologie structurale » de macromolécules biologiques notamment les protéines (d'intérêt pharmaceutique), mais aussi les acides nucléiques. Ces données expérimentales sont déposées dans la PDB par des biologistes et des biochimistes du monde entier et appartiennent au domaine public (accès gratuit via son site internet https://www.uniprot.org/database/DB-0070). Elle a été fondée en 1971 par le laboratoire national de Brookhaven avant d'être transférée en 2003 au projet Worlwide Protein Data Bank (wwPDB), qui se compose de PDBe (Europe), PDBj (Japon), RCSB PDB (Etats-Unis).

- **InterPro**

InterPro est une base de données contenant une ressource de documentation intégrée pour les familles de protéines, les domaines et les sites fonctionnels, dans lesquels des caractéristiques identifiables trouvées dans des protéines connues peuvent être appliquées à de nouvelles séquences protéiques afin de les caractériser fonctionnellement. Cette première regroupe les efforts des projets de bases de données PROSITE, PRINTS, Pfam et ProDom. Elle a été crée en 1999 par l'institue européen de bioinformatique (European Bioinformatics Institute) et comprend donc une description fonctionnelle, une annotation, des références bibliographiques et des liens vers la/ou les bases de données pertinentes des membres, accessible en ligne via son site http://www.ebi.ac.uk/interpro/.

- **UniProt**

Appellation dérivant de la contraction *Universal Protein Resource* (base de données **uni**verselle de **pro**téines). Cette base de données de séquences protéiques est accessible en ligne, elle est issue de la consolidation de l'ensemble des données produites par la communauté scientifique. UniProt est une base de données hiérarchique annotée, où toute séquence est livrée avec un riche ensemble de métadonnées et de liens vers de multiples autres bases de données : nucléotidiques, bibliographiques, phylogénétiques, ... Outre la séquence en acides aminés des protéines, UniProt fournit des informations sur leur fonction et leur structure.

- **SWISS-PROT**

SWISS-PROT est une base de données annotée de séquences protéiques fiables et organisée qui s'efforce de fournir un haut niveau d'annotation (comme la description de la fonction d'une protéine, la structure de ses domaines, les modifications post-traductionnelles, les variantes à savoir la similarité à d'autres protéines, les maladies associées, ...), limitant à un niveau minimal les redondance et possédant de nombreux liens avec d'autres banques. Swiss-Prot a été créé en 1986 par Amos Bairoch au département de biochimie médicale de l'Université de Genève et est le fruit d'un effort collaboratif du département et du laboratoire européen de biologie moléculaire (EMBL) depuis 1987. SWISS-PROT est disponible sur: http://www.expasy.ch/sprot/ et http://www.ebi.ac.uk/swissprot/.

- **PIR (Protein Information Resource)**

Le PIR est une banque de donnée de séquences protéiques et d'outils d'analyse librement accessible à la communauté scientifique (https://proteininformationresource.org/). Localisé à l'université de Georgetown aux États-Unis et établi en 1984, par le National Biomedical Research Foundation, il permet d'aider les chercheurs dans l'interprétation et l'identification des séquences protéiques obtenus dans leurs expériences.

V.5. Techniques d'analyse des données

V.5.1. Recherche d'homologie/divergence entre séquences

La comparaison de séquences génomiques et protéiques est l'une des premières tâches informatiques couramment exécutée par les biologistes afin de pouvoir analyser les données brutes de ces expériences. Il s'agit de déterminer dans quelle mesure ou proportion deux séquences ou plus, génomiques ou protéiques, se ressemblent.

Pour ce faire des algorithmes ont été mis en œuvre afin de calculer les meilleurs alignements entre deux ou plusieurs séquences.

***Rappelle :* Qu'est-ce qu'une séquence ?**

Une séquence nucléotidique/génomique est l'enchaînement des nucléotides le long d'une macromolécule d'ADN/ARN. Elle est généralement représentée par une succession des caractères alphabétique représentant les bases azotés (A, C, G, U et T) qui distinguent les cinq types de nucléotides. La succession des différents nucléotides au niveau des régions codantes constituant les gènes détermine la séquence en acides aminés qui compose un polypeptide, et son repliement ainsi que les diverses modifications chimiques aboutissant à une protéine fonctionnelle.

Le long d'un polypeptide, une séquence protéique est une succession des vingt types d'acides aminés ; cette séquence est généralement représentée par une chaîne utilisant un alphabet de vingt lettres.

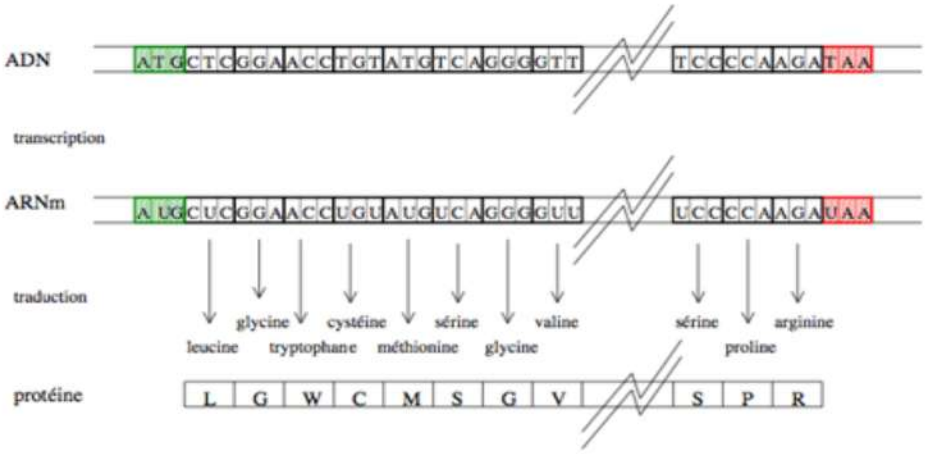


Figure 24 : D'une séquence génomique à une séquence protéique.

Le but principal à la comparaison de séquences est de déduire des informations sur la/les séquence(s) traité à partir des connaissances attachées à une autre préalablement étudié et préexistante dans une base de données et ainsi aider à identifier les membres d'une même famille de gènes, de genre ou d'espèce d'organisme. Par conséquent, si deux séquences génomiques sont assez semblables, et que l'une d'elles est connue pour être codante, alors on peut émettre l'hypothèse que la seconde le soit également. De même, on suppose généralement que les protéines correspondantes remplissent des fonctions similaires ; si la fonction de l'une est connue, la fonction de la seconde peut en être déduite, si deux séquences protéiques sont similaires. Ce principe d'inférence se justifie par des considérations sur le processus d'évolution.

Afin de pouvoir donc comparer entre deux séquences voir plusieurs, il est nécessaire d'avoir recoure à des logiciels ou des bases de données qui permettent justement de comparer nos séquences avec d'autres présentent au niveau de ces bases de données, entre autre :

- **BLAST**

Le Blast pour **B**asic **L**ocal **A**lignment **S**earch **T**ool, est un programme bioinformatique conçu pour trouver des régions similaires entre des séquences biologiques (nucléiques ou protéiques) en réalisant un alignement de ces régions homologues. Ce programme, disponible en ligne, est généralement rattaché aux banques de données génomiques et protéomiques générales ou spécifiques. Le plus répondu est celui du centre national de l'information en biotechnologie (National Center for Biotechnology Information (NCBI)).

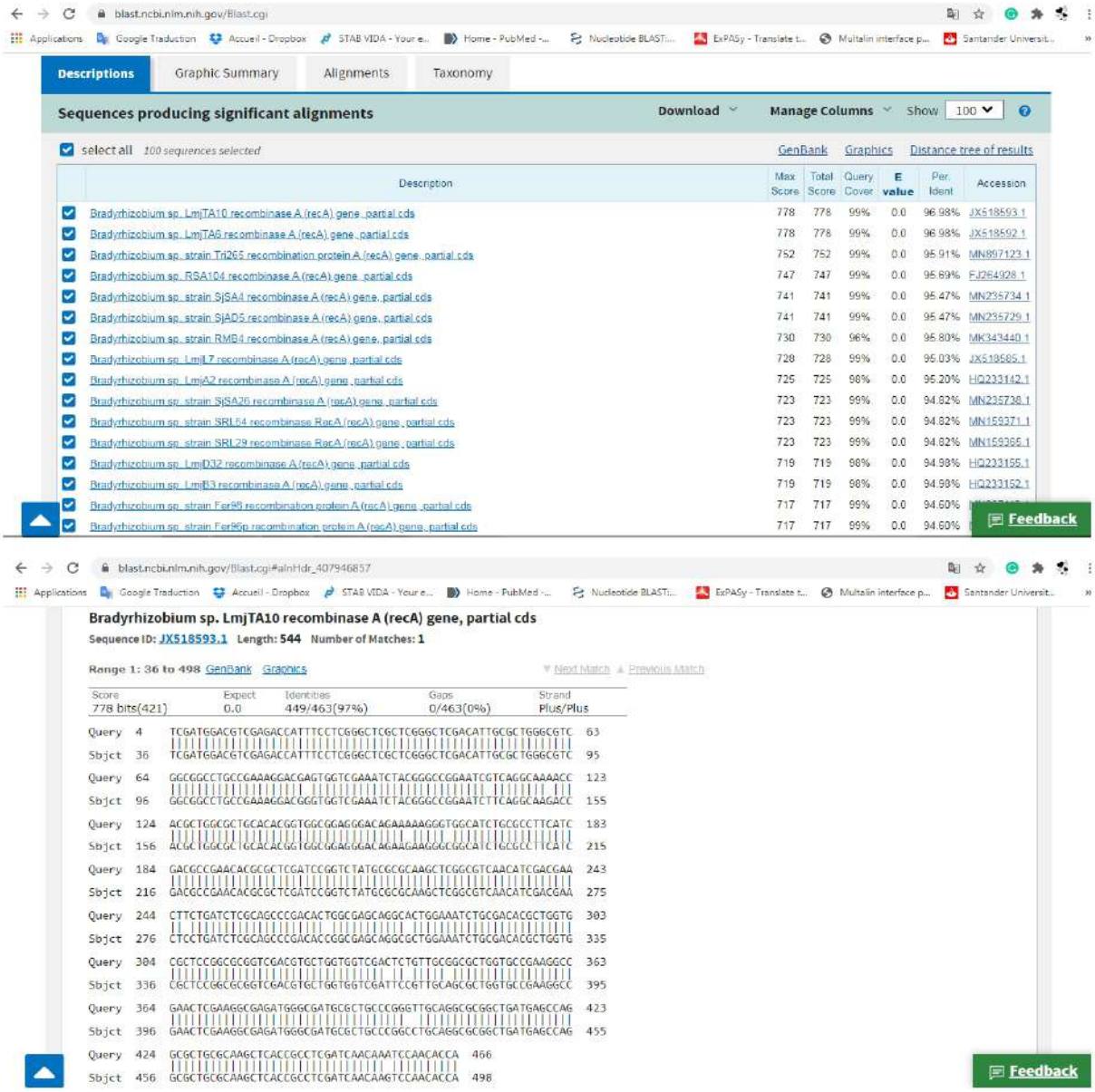


Figure 25 : Exemple de résultats de similarité entres séquences nucléotidiques.

Son principe de fonctionnement est que BLAST recherche dans une base de données de séquence des segments qui sont localement homologues à une séquence-test fournie par le scientifique. Une matrice de similarité est utilisée par BLAST afin de calculer des scores d'alignement. Il procure un score pour chacun des alignements trouvés et utilise ce score pour octroyer une estimation statistique de la pertinence de cet alignement (probabilité dû au hasard).

V.5.2. Alignement de séquences

L'alignement de séquences / alignement séquentiel est un processus par lequel deux (ou n) séquences sont **comparées** afin d'obtenir le plus de correspondances (identités ou

substitutions conservatives) possibles entre les bases qui les composent et montrer s'ils dérivent d'un ancêtre commun, et leur identité par estimation de la fraction de résidus identiques. Les séquences nucléotidiques proches ont très probablement des rôles proches et les conclusions fonctionnelles peuvent également être tirées suite à la comparaison de séquences nucléotidiques ou même protéiques.

La similarité entre séquences est considérée comme étant le pourcentage d'identités et/ou de substitutions conservatives. La quantification du degré de similarité est donnée par un score. Le résultat de la recherche d'une similarité peut être utilisé pour inférer l'homologie de séquences. On considère deux séquences comme similaire si leur score de substitution est supérieur à 0.

Exemple

Si l'on compare deux séquences A et B et que l'on trouve que la séquence du gène *A* ressemble fortement à la séquence du gène *B*, alors on peut déduire que la séquence de la protéine codée par *A* sera très proche de la séquence de la protéine codée par *B*. Les deux protéines adopteront donc une structure tridimensionnelle proche et auront donc une fonction semblable.

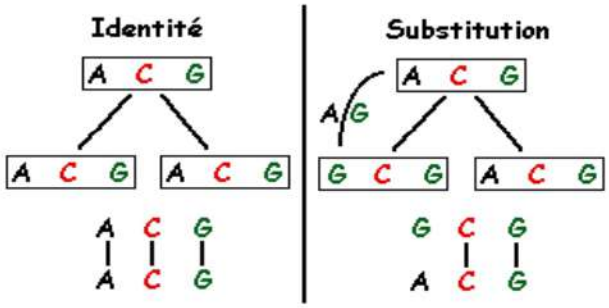


Figure 26 : Représentation de similarité et de dis-similarité.

On dit que deux séquences sont **homologues** si elles ont un ancêtre commun. L'homologie se mesure par la similarité : une similarité significative est signe d'homologie sauf si les séquences présentent une faible complexité à savoir une région contenant peu de caractères différents.

V.5.2.1. Types d'alignement

L'alignement ou la comparaison de séquences nucléotidiques ou protéique peut être considéré comme différent en fonction de la région prise comme cible de comparaison. On distingue :

- **Alignement local**

C'est l'alignement des séquences sur une partie de leur longueur.

- **Alignement global**

C'est un alignement des séquences sur toute leur longueur.

- **Alignement optimal**

On considère l'alignement des séquences comme optimal, celui qui produit le plus haut score possible de similarité.

- **Alignement multiple**

Est un alignement global de trois séquences ou plus.

Lors d'un alignement on peut voir apparaitre des brèches ou ce qu'on appelle plus communément des "***gap***" qui est un espace artificiel introduit dans une séquence donnée pour contrebalancer et matérialiser une insertion dans une autre séquence. Il permet d'optimiser l'alignement entre les séquences.

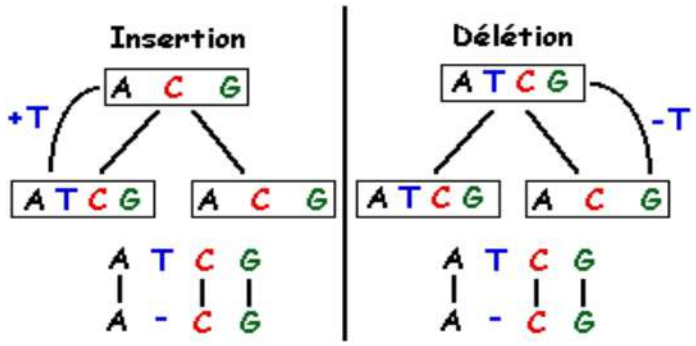


Figure 27 : Représentation d'un alignement entre deux séquences avec la formation de gap.

V.5.2.2. Programmes d'alignement

Les méthodes de programmation permettent de calculer, sous un système de scores donné, l'alignement optimal, global ou local, entre deux séquences voir plus en un temps proportionnel au produit des longueurs des séquences cibles. Lorsque ces programmes sont appliqués à une banque de séquences, le temps de calculs de ces méthodes augmente linéairement avec la taille de la banque. On peut définit deux spécificités pour une méthode de comparaison de séquences :

- ***La sensibilité***

C'est l'aptitude à détecter toutes les similarités considérées comme significatives et donc à générer le minimum de faux-négatifs.

- ***La sélectivité***

C'est l'aptitude à ne sélectionner que des similarités considérées comme significatives et donc à générer le minimum de faux-positifs.

Les programmes des familles *Fasta* et *BLAST* permettent de réduire le facteur temps en sacrifiant un peu de sensibilité. L'un et l'autre simplifient le problème en présélectionnant les séquences de la banque susceptibles de présenter une similarité significative avec la séquence requête et en localisant les régions potentiellement similaires dans les séquences.

Ces étapes sélectives permettent de n'appliquer les méthodes de comparaison, coûteuses en temps, qu'à un sous-ensemble des séquences de la banque et de restreindre le calcul de l'alignement optimal à des parties des séquences. Cette logique de recherche plus rapide dans son exécution, comporte donc le risque d'éliminer des séquences qui ont une similarité plus difficile à détecter ou d'aboutir à des alignements sub-optimaux.

- **Programme FASTA**

Ce programme ne considère que les séquences présentant une région de forte similitude avec la séquence recherchée. Il applique ensuite localement à chacune de ces meilleures zones de ressemblance un algorithme d'alignement optimal. La fragmentation de la séquence en courts motifs (nommés *uplets*) transcodés en entiers à savoir la codification numérique des séquences, donne à l'algorithme l'essentiel de sa rapidité.

- **Programme BLAST**

Le programme BLAST (*Basic Local Alignment Search Tool*) s'appuie sur une méthode heuristique qui utilise la méthode de Smith et Waterman. C'est un programme qui effectue un alignement local entre deux séquences nucléiques ou protéiques. La recherche des homologies entre une séquence à analyser et toutes les séquences d'une base de données est permit grâce à la rapidité du ***BLAST***.

- **Programme d'alignement multiple progressif (*Clustal*)**

Le programme *ClustalW* utilise un algorithme d'alignement multiple progressif. La similarité de chaque séquence est évaluée par rapport à **toutes les séquences**, puis 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 et on obtient ainsi une **matrice de distances**.

Un dendrogramme ("*guide tree*") est construit : il s'agit d'un arrangement traduisant les relations globales de **parenté** entre les séquences. Cet arbre phylogénique est construit selon

la méthode "*Neighbor-Joining*". Il indique donc l'ordre à partir duquel l'alignement multiple graduel sera établi.

Le programme *ClustalW* comporte des particularités qui minimisent ce risque d'un alignement erroné :

- Le poids des séquences est ajusté ;
- Selon l'étape de l'alignement et la divergence des séquences, des matrices de substitution appropriées sont utilisées ;
- L'introduction de *gap* est favorisée à des endroits spécifiques.

V.5.3. Correction manuelle des séquences brutes

Les instruments de séquençage (Sanger, illumina, et autre) modernes génèrent des fichiers de trace de chromatogramme «bruts» qui nécessitent un traitement préalable pour obtenir des séquences de qualité suffisante pour des analyses en aval. La présence d'erreurs de séquençage de type substitution et insertion/délétion dans les séquences (*reads*) introduites à deux niveaux, pendant l'amplification ou au moment du séquençage peuvent arriver, ce qui complexifie les analyses sous-jacentes. En effet, ces erreurs peuvent impliquer des temps d'alignement plus importants, voir fausser les résultats d'un alignement, dû à la dis-similarité qu'elles induisent entre les séquences ce qui conduit forcément à des molécules inexactes. Ainsi, des variations séquentielles, peuvent être amenées à être erronément confondues avec des erreurs de séquençage et *vis-versa*.

La fréquence de ces erreurs est dépendante des protocoles propres à chaque technologie de séquençage et aux polymérases utilisées. Les erreurs de séquençage *sensus stricto* sont des erreurs de lecture par le séquenceur. Il est donc nécessaire et systématique de procéder à l'inspection de chaque fichier de trace (chromatogramme) pour identifier les exécutions de séquençage problématiques, supprimer les base non fiables et découper les extrémités de la séquence dont les bases n'ont pas été identifié et mentionné comme 'NNN'.

Remarque

Comme vous le savez l'ADN est double brin et lors du séquençage chaque brin est obtenu séparément (individuellement). Il est donc nécessaire de procéder à l'assemblage (apparier les séquences) et l'alignement des deux séquences d'ADN (directe et inverse) afin de déterminer une seule séquence consensus à partir de la paire puis procéder à la correction et garantir de ce fait la qualité de la séquence finale.

V.5.3.1. Programmes informatiques utilisés pour la correction

La détection ainsi que la correction des erreurs représente la deuxième étape de la plupart des expériences d'analyse des données de séquençage après l'identification (Blast). La

correction d'erreurs repose sur différentes méthodologies, dépendant principalement du taux et du profil d'erreurs des *reads*. De ce fait, des développements algorithmiques différents ont été développés. Ainsi, de nombreuses méthodes visant à corriger ces erreurs, appelées correcteurs, ont été mis au point. Plusieurs applications/logiciels sont disponible et mise à disposition du biologiste afin de pouvoir donc corriger facilement ses séquences. Relativement peu d'outils logiciels non propriétaires (payant) sont disponibles pour faciliter ce processus.

- **Sequencher**

C'est un logiciel de bioinformatique, destiné au traitement artificiel de données biologique à savoir l'assemblage et l'analyse de séquences d'ADN. Pour étudier les mutations génétiques, le séquençage ainsi que la détection d'anomalies, cette application est largement utilisée par les différents laboratoires. Ses capacités comprennent des algorithmes d'assemblage d'ADN multiples et configurables, la présence d'outils d'édition de séquence d'ADN complets, la possibilité d'introduire des séquences de référence afin de comparer les séquences et détecter assez rapidement et facilement les SNP, cartographier les sites de restrictions d'une séquence X, la capacités d'importation et d'exportation de données, y compris la gestion personnalisable des fonctionnalités de GenBank.

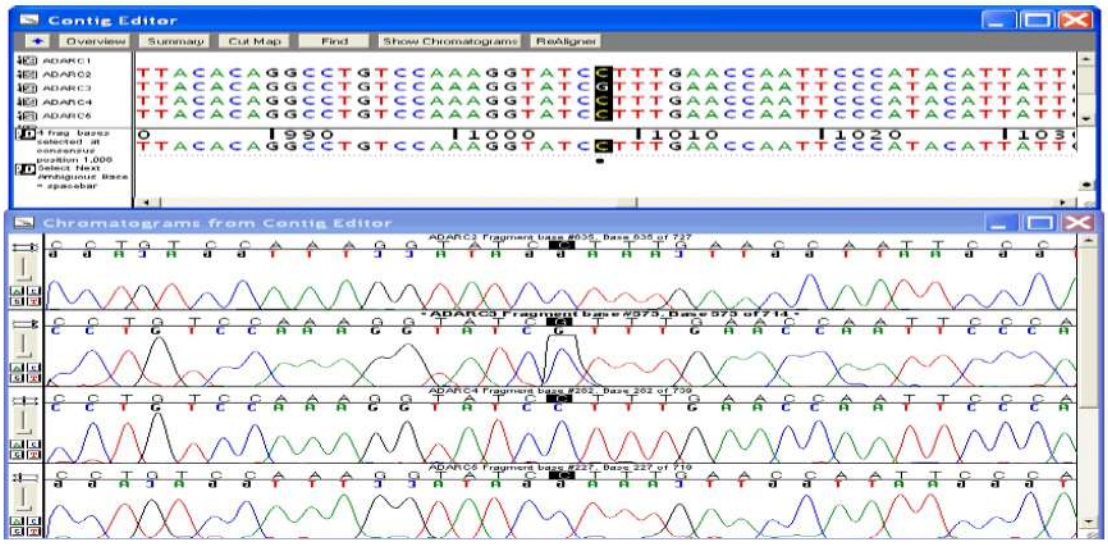


Figure 28 : Interface du Sequensher montrant la sélection d'une base du chromatogramme.

C'est un logiciel pouvant fonctionner avec les différents systèmes d'exploitation. Il peut être téléchargé sur son site officiel http://www.genecodes.com/. La version d'essai est gratuite et d'une validité de 15 jours. Après ces 15 jours, afin d'avoir toutes les fonctionnalités de ce logiciel il est impératif de procéder à l'achat de la licence.

- **Geneious**

Geneious Basic est un logiciel de bureau pour l'organisation et l'alignement, l'assemblage et l'analyse de séquences d'ADN, d'ARN et de protéines, intégrant de nombreux outils de bioinformatique et de biologie moléculaire dans une interface simple, facile à utiliser et flexible.

Figure 29 : Arbre phylogénétique montrant la relation d'une séquence de source inconnue avec des séquences publique similaire effectuées avec Geneious Basic.

- **SeqTrace**

SeqTrace est une application gratuite et open source conçue pour automatiser l'ensemble du travail en facilitant le traitement par lots d'un grand nombre de fichiers de trace. SeqTrace peut identifier, aligner et calculer des séquences consensus à partir de la correspondance des traces avant et arrière, filtrer les appels de base de faible qualité et terminer les séquences finies. Le logiciel dispose d'une interface graphique qui comprend un visualiseur de chromatogramme complet et un éditeur de séquence. SeqTrace fonctionne sur les systèmes d'exploitation les plus courants et est disponible, avec la documentation de support, à http://seqtrace.googlecode.com/.

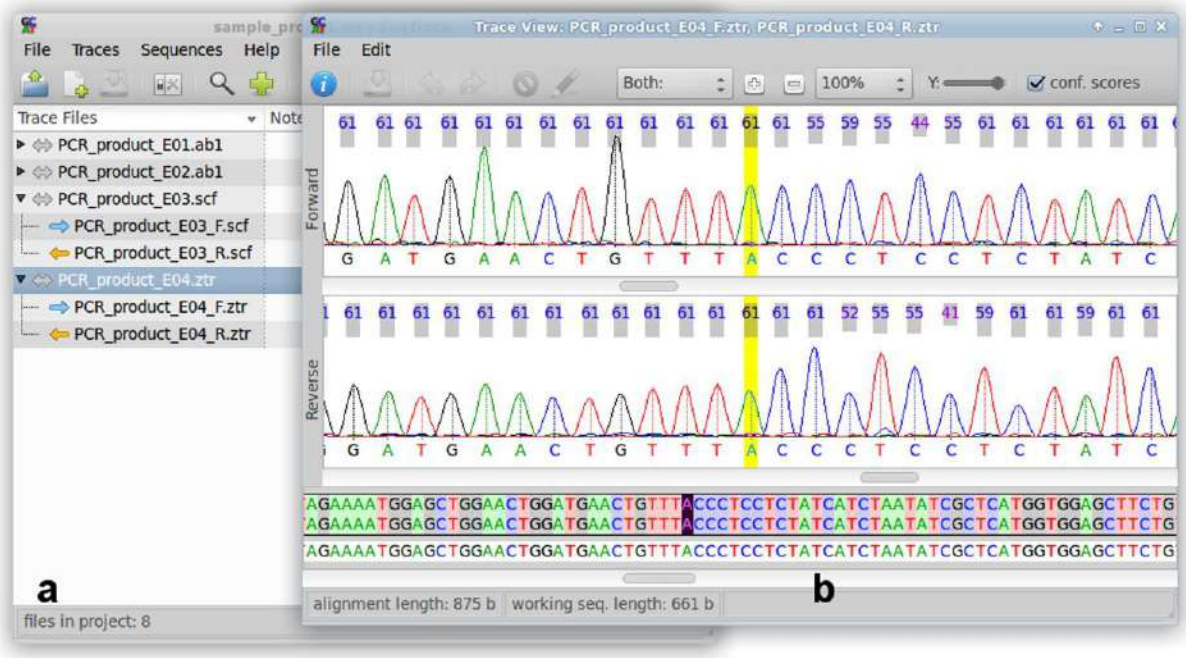


Figure 30 : Interface de SeqTrace, y compris la fenêtre du projet (a) et la fenêtre d'affichage du chromatogramme (b).

V.6. Construction d'arbres phylogénétiques

La phylogénie est une pratique couramment employé par les scientifiques, il s'agit de l'étude de l'évolution des organismes vivants en vue d'établir leur parenté. Le but de la phylogénie est de comprendre les relations de filiation, de retracer l'historique évolutif d'une espèce ou d'un groupe taxonomique supérieur d'organismes. Les arbres phylogénétiques basés sur certains gènes marqueurs permettent de schématiser et d'appréhender ces relations rapidement.

Un **arbre phylogénétique** (arbre de parenté, arbre d'évolution ou encore cladogramme) est un arbre schématique permettant de retracer simplement les liens de parenté entre des groupes d'êtres vivants. Chaque arbre est constitué de plusieurs éléments indispensables à la compréhension de la filiation qui sont les nœuds, les branches ainsi que les racines.

- La racine symbolise le dernier ancêtre commun. Un arbre phylogénétique peut être enraciné ou pas, en fonction de la possibilité d'identifier l'ancêtre commun à toutes les feuilles.
- Les nœuds de l'arbre représente l'ancêtre commun de ses descendants, ce sont les unités taxonomiques évolutives (OTUs : Operational Taxonomic Unit).

- Les branches représentent les relations de parentés (ancêtre/descendants) entre les unités taxonomiques et on peut distinguer des branches internes et des branches externes. La longueur de la branche ne reflète pas le temps ou le taux de divergence.
- Le clade est représenté par les branches descendantes reliées à un ancêtre commun, non celui de l'ancêtre qui reste impossible à déterminer.

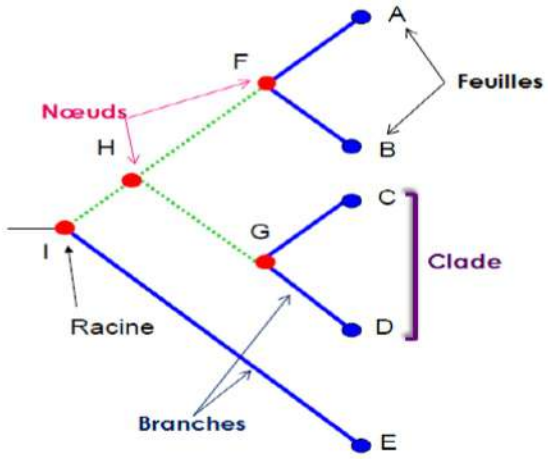


Figure 31 : Représentation graphique d'un arbre phylogénétique ainsi que ces constituants.

V.6.1. Méthodes de construction d'arbres phylogénétiques

Plusieurs méthodes peuvent être utilisées afin de construire cet arbre, plus ou moins rapides et plus ou moins fiables : Distance (phénétique), Maximum de parcimonie et Maximum de Vraisemblance.

a) Méthodes de distance

Ce sont des méthodes de construction d'arbre phylogénétique sans racine, rapides et bien adaptées aux séquences présentant un degré de similarité élevé. Plusieurs méthodes ont été développées pour construire un arbre phylogénétique à partir d'une matrice de distance.

- UPGMA « Unweighted Pair Group Method with Arithmetic mean » : très simple, basée sur le groupement des séquences les plus similaires, indépendamment de leur vitesse d'évolution et de leurs parentés phylogénétiques, sans déterminer d'ancêtre commun. Elle a vite été délaissée au profit d'autres méthodes plus avancées.

- Neighbor-Joining (NJ) : développée par Saitou et Nei (1987), elle a l'avantage d'être vraiment rapide. En général, elle est utilisée pour faire des arbres de plusieurs milliers de séquences et de similarité très élevé. C'est actuellement la plus utilisée pour reconstruire des phylogénies par méthode de distance.

b) Méthode du maximum de vraisemblance ou « *Maximum Likelihood*-ML»

C'est une méthode dite de caractères, elle repose sur un ou plusieurs caractères à évaluer. Il s'agit d'une méthode probabiliste, qui nécessite d'appliquer différents modèles d'évolution et le choix de ce modèle est crucial pour la qualité de l'arbre obtenu. On dit qu'il convient de l'utiliser à partir du moment où le nombre de caractères analysés est supérieur à la moitié du nombre de séquences analysées, sinon la reconstruction est considérée comme incorrecte. Elle est couramment décrite comme le meilleur moyen de trouver l'arbre le plus proche de la réalité, c'est-à-dire le moyen le plus efficace. Son désavantage se situe au niveau des temps de calculs qui sont extrêmement longs.

c) Méthodes de maximum de parcimonie (*Maximum Parcimony*)

Cette méthode est très appréciée car rapide en temps de calcul, mais pas aussi précise que la précédente (ML). Elle permet de construire des arbres de classification hiérarchique après enracinement, qui permettent de refléter la structure de parenté d'un ensemble de taxons. Cette méthode utilise des algorithmes basés sur les caractères plutôt que de distances. Plusieurs programmes utilisent la méthode de maximum de parcimonie pour la reconstruction d'un arbre, dont MEGA version 5, 6 et X.

V.6.2. Programmes de construction d'arbres phylogénétiques

Plusieurs logiciels informatiques ont été mise au point afin de pouvoir construire des arbres phylogénétiques retraçant les lignées de parenté entre les organismes. Ils sont disponibles en ligne en libre accès. On citera donc les programmes les plus utilisés à savoir le PAUP, MEGA, Phylowin, Phylip, ARB, DAMBE, bosque, EMBOSS, Crux,…. Le site suivant, http://evolution.genetics.washington.edu/phylip/software.html, contient tout les logiciels pouvant être utilisé pour la construction d'arbres phylogénétique. Chaque programme à ces propres spécificités et accompagné d'un manuel explicatif.

Généralement, afin de pouvoir construire un arbre phylogénétique on a recoure à une matrice constitué de toutes les séquences nucléotidiques ou protéiques nécessaire à l'étude en question. La méthode de distance est choisie en fonction du critère de distance entre les futures feuilles de l'arbre. Par exemple, si on dispose de séquences d'ADN, on peut choisir comme distance entre deux d'entre elles le nombre de nucléotides qui diffèrent. Pour déterminer cette valeur, on est systématiquement amené à en effectuer un alignement. Puis on peut utiliser la méthode UPGMA ou celle du Neighbor-Joining pour en déduire l'arbre.

Références

Achachi A., El Fahime M., El Alaoui M.A., Alaoui S., Ait Barka E., El Guilli M., Soulaymani A., Ibriz M. (2012) Méthode améliorée d'extraction des ARN totaux et mise au point de la technique RT-PCR pour la détection du *Citrus psorosis virus* au Maroc. *10e Conférence internationale sur les maladies des plantes tours*, 3, 4 et 5 Décembre 2012. 1-10.

Algorithmes et programmes de comparaison de séquences. Interprétation des résultats : E-value, P-value. http://biochimej.univ-angers.fr/Page2/BIOINFORMATIQUE/7ModuleBioInfoJMGE/9AlgoProgramme/1ProgAlgo.htm. Consulté en ligne le 10.09.2020.

Altschul S.F. Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 25, 3389 – 3402.

Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* 215(3), 403-410. doi: 10.1016/S0022-2836(05)80360-2.

Ameziane, N., Bogard, M., Lamoril, J. (2006) Principes de biologie moléculaire en biologie clinique. Elsevier SAS.

Apweiler R., Attwood T.K., Bairoch A., Bateman A., Birney E., Biswas M., Bucher P., Cerutti L., Corpet F., Croning M.D.R., Durbin R., Falquet L., Fleischmann W., Gouzy J., Hermjakob H., Hulo N., Jonassen I., Kahn D., Kanapin A., Karavidopoulou Y., Lopez R., Marx B., Mulder N.J., Oinn T.M., Pagni M., Servant F., Sigrist C.J.A., Zdobnov E.M. (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res*. 29(1), 37–40. doi: 10.1093/nar/29.1.37.

Bairoch A., Apweiler R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res*. 28(1), 45–48. doi: 10.1093/nar/28.1.45.

Bairoch A., Apweiler R. (1996) The SWISS-PROT Protein Sequence Data Bank and Its New Supplement TREMBL. *Nucleic Acids Res*. 24(1), 21–25. https://doi.org/10.1093/nar/24.1.21.

Benson D.A., Karsch-Mizrachi I., Lipman D.J., Ostell J., Sayers E.W. (2010) GenBank. *Nucleic Acids Res*. 38, D46–D51. doi:10.1093/nar/gkp1024.

Berger S.L., Kimmel A.R. (1987) Guide to molecular cloning techniques, Methods in Enzymology, *Academic Press*, 152.

Bernstein F.C., Koetzle TF., Williams G.J.B., MeyerJr E.F., Brice M.D., Rodgers J.R., Kennard O., Shimanouchi T., Tasumi M. (1977) The protein data bank: A computer-based archival file for macromolecular structures. *J. mol. biol.* 112(3), 535-542. https://doi.org/10.1016/S0022-2836(77)80200-3.

Beroud C., Antignac C., Jeanpierre C., Junien C. (1990) Un programme informatique pour la recherche d'amorces pour l'amplification par PCR. *médecine/sciences*. 6, 901-903.

Berthet N. (2013) La puce à ADN de reséquençage : un outil rapide pour mieux identifier et comprendre une émergence virale et bactérienne. *Bull. Acad. Natle Méd.*, 197(9), 1669-1682.

Cherry J.M., Adler C., Ball C., Chervitz S.A., Dwight S.S., Hester E.T., Jia Y., Juvik G., Roe T.Y., Schroeder M., Weng S., Botstein D (1998). SGD: *Saccharomyces* Genome Database. *Nucleic Acids Res*. 26(1), 73–79. https://doi.org/10.1093/nar/26.1.73.

Cherry J.M., Hong E.L., Amundsen C., Balakrishnan R., Binkley G., Chan E.T., Christie K.R., Costanzo M.C., Dwight S.S., Engel S.R., Fisk D.G., Hirschman J.E., Hitz B.C., Karra K., Krieger C.J., Miyasato S.R., Nash R.S., Park J., Skrzypek M.S., Simison M., Weng S., Wong E.D. (2012) *Saccharomyces* Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res*. 40, D700–D705. doi: 10.1093/nar/gkr1029.

Christie K.R., Weng S., Balakrishnan R., Costanzo M.C., Dolinski K., Dwight S.S., Engel S.R., Feierbach B., Fisk D.G., Hirschman J.E., Hong E.L., Issel-Tarver L., Nash R., Sethuraman A., Starr B., Theesfeld C.L., Andrada R., Binkley G., Dong Q., Lane C., Schroeder M., Botstein D., Cherry J.M. (2004) *Saccharomyces* Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Res.* 32(1), D311-D314. https://doi.org/10.1093/nar/gkh033.

Corpet F., Chevalet C. (2000) Analyse informatique des données moléculaires. *INRA Prod. Anim.* 191-195.

Corpet F. (1988) Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res*. 16, 10881 – 10890.

Delgrange O. (2000) Bioinformatique : un domaine pluridisciplinaire. *élément*, 27-30.

Drysdale R. (2008) FlyBase: a database for the Drosophila research community. *Meth. Mol. Biol.* 420, 45–59. DOI 10.1007/978-1-59745-583-1.

Falentin H., Auer L., Mariadassou M., Pascal G., Rué O., Dugat-Bony E., Delbès C., Nicolas A., Rifa E., Mondy S., Boulch M.L., Cauquil L., Hernandez-Raquet G., Terrat S., Abraham A.L. (2019) Guide pratique à destination des biologistes, bioinformaticiens et statisticiens qui souhaitent s'initier aux analyses métabarcoding. *Le Cahier des Techniques de l'Inra*. 97, 1-13.

Fouque B. (1989) Sondes à acides nucléiques, marquage radioactif - Institut National des Sciences et Techniques Nucléaires, CEA - Ministère de !'Education Nationale.

Gaillardin C., Tinsley C.R. (2007) Génétique moléculaire. *AgroParisTech*. 181p.

Gilbert W., Maxam A. (1973) The nucleotide sequence of the lac operator. *Proc. Natl. Acad. SciUSA*. 70, 3581-3584.

Giraldo-Calderón G.I., Emrich S.J., MacCallum R.M., Maslen G., Dialynas E., Topalis P., Ho N., Gesing S., The VectorBase Consortium, Madey G., Collins F.H., Lawson D. (2015) VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. *Nucleic Acids Res.* 43, D707–D713. doi: 10.1093/nar/gku1117.

Goodfellow M. (1971) Numerical taxonomy of some nocardioform bacteria. *J. Gen. Microbiol*. 69, 33-80.

Guesdon J.L. (1989) Principe de l'utilisation des sondes nucléiques. *Ann. Biochim. Clin-Qué*., 28(1), 4-10.

Higgins D.G., Fuchs R., Stoehr P.J., Graham N.C. (1992) The EMBL Data Library. *Nucleic Acids Res*. 20, 2071–2074. https://doi.org/10.1093/nar/20.suppl.2071.

Housset, C., Raisonnier, A. (2010) Biologie Moléculaire. Université Pierre et Marie Curie. 1-207.

Howe K., Davis P., Paulini M., Tuli M.A., Williams G., Yook K., Durbin R., Kersey P., Sternberg P.W. (2012) WormBase: Annotating many nematode genomes. *Worm*. 1(1), 15–21. doi: 10.4161/worm.19574.

Huala E., Dickerman A.W., Garcia-Hernandez M., Weems D., Reiser L., LaFond F., Hanley D., Kiphart D., Zhuang M., Huang W., Mueller L.A., Bhattacharyya D., Bhaya D., Sobral B.W., Beavis W., Meinke D.W., Town C.D., Somerville C., Rhee S.Y. (2001) The *Arabidopsis* Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res*. 29(1), 102–105. doi: 10.1093/nar/29.1.102.

Kaminuma E., Kosuge T., Kodama Y., Aono H., Mashima J., Gojobori T., Sugawara H., Ogasawara O., Takagi T., Okubo K., Nakamura Y. (2011) DDBJ progress report. *Nucleic Acids Res.* 39, D22-27. doi:10.1093/nar/gkq1041. PMC 3013661. PMID 21062814.

Kearse M., Moir R., Wilson A., Stones-Havas S., Cheung M., Sturrock S., Buxton S., Cooper A., Markowitz S., Duran C., Thierer T., Ashton B., Meintjes P., Drummond A. (2012) Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics Applications Note*. 28(12), 1647–1649. doi:10.1093/bioinformatics/bts199.

Kerdelhue C., Rasplus J.Y. (20) Le séquençage des acides nucléiques et les méthodes d'analyse des données moléculaires en phylogénie. *Bulletin de la Société Entomologique de France*. 1-37.

Lamoril J., Ameziane N., Deybach J.C., Bouizegarène P., Bogard M. (2008) DNA sequencing technologies: A revolution in motion. *Immuno-analyse & Biologie Spécialisée*. 23(5), 260-279. DOI : 10.1016/j.immbio.2008.07.016.

Land, A.H., Doig, A.G. (1960) An automatic method of solving discrete programming problems. *Econometrica*. 28, 497-520.

Lee P., Hudson T.J. (2000) La puce à ADN en médecine et en science. *médecine/sciences,* 16(1), 43-9.

Maftah A., Petit J.-M., Raymond J. 2018. Mini manuel de biologie moléculaire. 4e édition, Dunod, Paris, 2007, 2011, 2015, 2018, ISBN 978-2-10-077368-8.

Mulder N.J., Apweiler R., Attwood T.K., Bairoch A., Bateman A., Binns D., Bork P., Buillard V., Cerutti L., Copley R., Courcelle E., Das U., Daugherty L., Dibley M., Finn R., Fleischmann W., Gough J., Haft D., Hulo N., Hunter S., Kahn D., Kanapin A., Kejariwal A., Labarga A., Langendijk-Genevaux P.S., Lonsdale D., Lopez R., Letunic I., Madera M., Maslen J., McAnulla C., McDowall J., Mistry J., Mitchell A., Nikolskaya A.N., Orchard S., Orengo C., Petryszak R., Selengut J.D., Sigrist C.J.A., Thomas P.D., Valentin F., Wilson D., Wu C.H., Yeats C. (2007) New developments in the InterPro database. *Nucleic Acids Res.* 35, D224–D228. https://doi.org/10.1093/nar/gkl841.

Mulder N.J., Apweiler R., Attwood T.K., Bairoch A., Bateman A., Binns D., Bradley P., Bork P., Bucher P., Cerutti L., Copley R., Courcelle E., Das U., Durbin R., Fleischmann W., Gough J., Haft D., Harte N., Hulo N., Kahn D., Kanapin A., Krestyaninova M., Lonsdale D., Lopez

R., Letunic I., Madera M., Maslen J., McDowall J., Mitchell A., Nikolskaya A.N., Orchard S., Pagni M., Ponting C.P., Quevillon E., Selengut J., Sigrist C.J.A., Silventoinen V., Studholme D.J., Vaughan R., Wu C.H. (2005) InterPro, progress and status in 2005. *Nucleic Acids Res*.1(33), D201-D205. doi: 10.1093/nar/gki106.

Naumoff D.G., Carreras M. (2009) PSI Protein Classifier: a new program automating PSI-BLAST search results. *Mol. Biol. (Engl Transl).* 43, 652 – 664.

Needleman S.B., Wunsch C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48, 443-453.

Pearson W.R., Lipman D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA.* 85, 2444 – 244.

Peccoud J. (1993) La PCR quantitative : un nouvel outil pour l'analyse médicale. *médecine/sciences*. 9, 1378-1385.

Poitras E., Houde A. (2002) La PCR en temps réel: principes et applications. *Rev. Biol. Biotech*. 2(2), 2-11.

Radman M., Taddei F., Halliday J. (1994) Correction des erreurs dans l'ADN : de la génétique bactérienne aux mécanismes de prédisposition héréditaire aux cancers chez l'homme. *médecine/sciences*. 10, 1024-1030.

Riquet J., Pitel F. (2000) Les techniques de base de la génétique moléculaire. *INRA Prod. Anim.* 29-35.

Rosuel Y. (1994) Apport de la PCR dans les méthodes de biologie moléculaire appliquées au diagnostic des adénomes hypophysaires. Sciences pharmaceutiques. ffdumas-02156744f.

Saitou N., Nei M. (1987) The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees. *Mol. Biol. Evol.* 4, 406-425.

Sambrook J., Fritsch E.F., Maniatis T. (1989) Molecular Cloning: A laboratory Manual - Second Edition, Cold Spring Harbor. Laboratory Press.

Sanger F., Nicklen S. Coulson A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA*. 74, 5463—5467.

Smith, T., Waterman M. (1981) Identification of common molecular subsequences. *J. Mol. Biol.* 147, 195-197.

Strachan T., Read A. (2012) Génétique moléculaire humaine. 4^{e} édition, Médecine sciences publication, Lavoisier, France, 781p.

Stucky B.J. (2012) SeqTrace: A Graphical Tool for Rapidly Processing DNA Sequencing Chromatograms. *J. Biomol. Tech*. 23(3), 90–93. doi: 10.7171/jbt.12-2303-004.

Tagu D. Moussard C. (2003) Principes des techniques de biologie moléculaire. 2^{e} édition revue et augmentée. INRA, Paris. 184p.

Tamura K., Stecher G., Peterson D., Filipski A., Kumar S. (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol*. 30, 2725-2729.

Tateno Y., Imanishi T., Miyazaki S., Fukami-Kobayashi K., Saitou N., Sugawara H., Gojobori T. (2002) DNA Data Bank of Japan (DDBJ) for genome scale research in life science. *Nucleic Acids Res.* 30(1), 27–30. *doi:10.1093/nar/30.1.27. PMC 99140. PMID 11752245.*

The FlyBase Consortium. (1996) FlyBase: The *Drosophila* Database. *Nucleic Acids Research*. 24(1), 53–56. https://doi.org/10.1093/nar/24.1.53.

Thompson J.D., Higgins D.G., Gibson T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*. 22, 4673 – 4680.

Tse C., Capeau J. (2003) Quantification des acides nucléiques par PCR quantitative en temps réel. *Ann. Biol. Clin. 61, 279–293.*

Uhel F., Zafrani L., Commission de la recherche translationnelle de la Société de réanimation de langue française. (2019) Nouvelles techniques de biologie moléculaire (New Techniques in Molecular Biology). *Méd. Intensive Réa.* 1-9. DOI 10.3166/rea-2019-0119

Uniprot Consortium. (2007) The Universal Protein Resource (UniProt), *Nucleic Acids Res.* 35, D193-197. DOI 10.1093/nar/gkl929.

UniProt Consortium. (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res*. 47, D506–D515. doi: 10.1093/nar/gky1049.

Université de Tours (2010) Les outils de génétique moléculaire. *Réseau GÉNET.* http://genet.univ-tours.fr/gen001300_fichiers/CHAP5D/GEN05D1EC22.HTM

Védy S., Valois A., Budzilawski D., Perez P., Puyhardy J.M. (2013) Bilan d'utilisation d'une PCR « maison » qualitative : une validation acceptable en regard de la norme NF EN ISO 15189 ? *Ann Biol Clin.* 71(3), 363-72.

Wu C.H., Yeh L.S.L., Huang H., Arminski L., Castro-Alvear J., Chen Y., Hu Z., Kourtesis P., Ledley R.S., Suzek B.E., Vinayaka C.R., Zhang J., Barker W.C. (2003) The Protein Information Resource. *Nucleic Acids Res*. 31(1), 345–347. doi: 10.1093/nar/gkg040.

Yang Z. (1996) Maximum-likelihood models for combined analyses of multiple sequence data. *J. Mol. Evol*. 42, 587-596.

Yarfitz S., Ketchell D.S. (2000) A library-based bioinformatics services program. *Bull. Med. Libr. Assoc*. 88(1), 36–48.