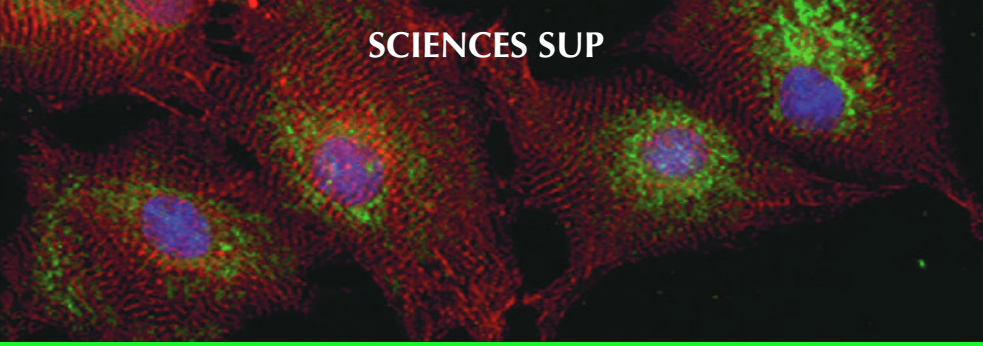


SCIENCES SUP

Aide-mémoire

L1/L2 • PCEM 1 • PH1

AIDE-MÉMOIRE
BIOLOGIE ET GÉNÉTIQUE MOLÉCULAIRES

Bernard Swynghedauw

avec la collaboration de Jean-Sébastien Silvestre

DUNOD

AIDE-MÉMOIRE

BIOLOGIE ET GÉNÉTIQUE MOLÉCULAIRES

Bernard Swynghedauw
Directeur de recherche émérite à l'INSERM (hôpital Lariboisière)

Jean-Sébastien Silvestre
Directeur de recherche à l'INSERM (hôpital Lariboisière)

3e édition

DUNOD

Les auteurs remercient Ludovic Dénard et Benjamin Peylet pour leurs précieux conseils.

ISBN 978-2-10-053798-3

Table des matières

➤ Note

Les gènes sont en italiques, comme c'est la règle, par contre « *les mots anglais sont en italique entre guillemets* ». On peut indifféremment écrire « des ARNs », « des ADNs » comme le font les anglophones (« *RNAs, DNAs* ») ou « des ARN », « des ADN » puisqu'en français beaucoup n'admettent pas de pluriel sur les abréviations. « Des ARN », « des ADN » a été choisi, bien que ce choix puisse être contesté.

Chapitre **1**

Introduction

« Biologie moléculaire » est un terme consacré par l'usage. *Stricto sensu,* « biologie moléculaire » devrait inclure tous les aspects moléculaires des études portant sur la vie. Il en va en pratique un peu différemment et l'on entend par là tout ce qui concerne les gènes, les produits des gènes, les aspects moléculaires de l'hérédité, c'est dire que « génétique moléculaire » (Clark 2005) ou « génomique » (Gibson 2004) seraient à la limite plus appropriés. La biologie moléculaire comprend aussi, et depuis peu, tout ce qui concerne les techniques dérivées de l'étude et de l'analyse de l'ensemble du génome, c'est-à-dire la génomique. D'une manière générale, la génomique inclut l'étude de la structure, du contenu et de l'évolution du génome, et recouvre de fait la génétique moléculaire, mais en pratique « génomique » recouvre également tout ce qui concerne l'expression génique aussi bien au niveau des ARN messagers (le transcriptome) qu'à celui des protéines (la protéomique), et de leur fonction (le physiome). Le point commun à toutes ces techniques est qu'elles mesurent la totalité des gènes exprimés ou

des protéines, et pas seulement quelques éléments sélectionnés *a priori*, comme c'était le cas jusqu'à maintenant. La métabolomique met potentiellement en évidence les effets nets des réactions enzymatiques en mesurant les produits (Gibson 2004). L'étape finale (pour l'instant) est la mise au point d'un outil permettant d'intégrer l'ensemble de ces données dans des réseaux interconnectés. Comme on le voit la définition est vaste et très empirique, surtout si on la compare par exemple aux définitions beaucoup moins ambiguës qui sont celles de la biologie cellulaire et de la biochimie.

Ce qui définit peut-être le mieux la biologie moléculaire telle qu'on la pratique actuellement, c'est qu'elle permet de mieux considérer la vie comme ce qu'elle est fondamentalement c'est-à-dire un processus unique, commun aux plantes, aux bactéries, aux insectes et aux animaux, ayant un ancêtre commun et que l'incroyable panoplie de techniques diverses qui constitue actuellement la Biotechnologie s'applique à toutes les formes de la vie (Clark 2005). C'est dire que l'Évolution, au sens darwinien du terme, doit toujours rester au cœur de la réflexion biologique (Darwin 1859, Ridley 2004, Stearns 2005).

Après un rappel historique, ce livre comportera quatre parties : structure et la fonction de l'ADN, données concernant le polymorphisme génotypique et phénotypique, en y incluant les bases de l'évolution, les principales biotechnologies et, pour finir, un certain nombre d'exemples issus de la génétique médicale, de la pharmacogénétique et de la biologie médicale. Les termes anglais figurent partout où ils ont semblé nécessaires, car en pratique l'étudiant et le chercheur les rencontreront tous les jours dans leurs lectures ; les traductions ne sont souvent pas rentrées dans les mœurs.

Je souhaite par cet aide-mémoire venir en aide aux seniors, cultivés, désireux d'avoir une idée rapide de la révolution contemporaine en biologie, et aider les plus jeunes à se situer par rapport aux deux révolutions scientifiques majeures contemporaines, la biologie moléculaire et l'informatique, la première étant d'ailleurs très dépendante de la seconde. Il n'y a pas de livre de ce type qui ne reflète l'expérience professionnelle de l'auteur, le présent ouvrage ne fait pas exception à la

règle, et le lecteur ne sera pas étonné de le voir illustré par des exemples pris surtout chez l'homme et en cardiologie.

Résumer l'essentiel de la biologie moléculaire en quelques pages était relativement facile il y a quelques années, les connaissances en biologie ont pris en ce moment une croissance exponentielle qui impose des choix, voire de véritables paris sur l'avenir. La biologie moléculaire et la génétique sont des disciplines en plein développement, presque chaque année apparaissent de nouvelles techniques, certaines représentent des avancées si considérables qu'elles rendent caduques les techniques encore routinières l'année d'avant. Nous avons tenu compte de ces avancées (exemple « *genome-wide association* », Ch. 5.2.).

L'avenir de la discipline et les sources d'emploi pour le jeune Docteur ès sciences sont certes dans l'enseignement et la recherche publique, mais surtout dans les applications, qu'il s'agisse des biotechnologies, de la Recherche/Développement (R/D) voire du marketing. Les biotechnologies offrent des possibilités immenses en R/D dans l'industrie pharmaceutique –mais aussi dans la quête de nouveaux biomarqueurs, dans la recherche biométrique, agronomique et dans l'industrie alimentaire. Il ne faut pas oublier, voire mépriser le marketing, ce secteur est souvent demandeur de scientifiques formés et à l'esprit ouvert. Ayant été confronté à ce problème pendant toute ma carrière, l'avenir concret du jeune scientifique ou du jeune médecin est une arrière-pensée, très terre à terre, qui fut un fil conducteur durant toute la rédaction de ce petit aide-mémoire.

Chapitre 2

Rappel historique

La révolution biologique à laquelle nous assistons n'aurait bien évidemment pas été possible sans l'ensemble du mouvement scientifique et technologique –à la limite, on pourrait dire que l'invention de l'imprimerie par Gutenberg ou celle de l'électricité ont été décisives –mais on peut être plus précis. Les racines de la biologie contemporaine deviennent identifiables au milieu du XIX^{e} siècle. Quatre noms ont été décisifs, chacun d'eux symbolisant une des idées-piliers de la biologie contemporaine : Charles Darwin et la théorie de l'évolution, Gregor Mendel et les lois de l'hérédité, Claude Bernard qui fit de la physiologie une science expérimentale, enfin Louis Pasteur et la microbiologie, outil premier des biotechnologies. Le milieu du XIX^{e} siècle n'a, bien entendu, pas été une période décisive qu'en biologie, il l'a été aussi dans presque toutes les autres sciences exactes, mais les racines ont bien commencé à pousser à cette époque et il y a bien, en biologie, aux environs de 1855-1865, un avant et un après.

2.1 AVANT LE MILIEU DU XIXe SIÈCLE

C'est J.-B. Lamarck qui a découvert l'évolution biologique. Lamarck était d'abord un savant du XVIIIe siècle, héritier de Buffon, qui avait réalisé un travail gigantesque de systématique, mais c'est sa « Philosophie Zoologique » qui reste la première théorie cohérente de l'évolution. Cette théorie était basée sur l'évolution des caractères acquis sous l'action directe du milieu et de l'usage. On sait maintenant que les caractères acquis ne sont pas transmissibles, mais c'est indiscutablement à Lamarck que l'on doit la notion d'évolution, ce qui ne l'a pas empêché de mourir en 1829 aveugle et oublié.

Au début du XIXe, l'hérédité était encore une notion confuse, en partie théologique, loin de toute approche scientifique. Les deux ténors de l'Académie des Sciences de l'époque, Cuvier et Geoffroy Saint-Hilaire, s'opposaient dans des joutes oratoires célèbres sans la moindre allusion à une quelconque hérédité. Pour Cuvier, fixiste convaincu, forte personnalité, il existe un plan fixe de création hiérarchisée du règne animale. Pour Geoffroy Saint-Hilaire, transformiste, tous les animaux sont bâtis sur le même plan grâce à un nombre limité d'éléments organiques, le tout dirigé par un fluide vital. Mais pour les deux protagonistes, surtout pour Cuvier, la diversité du monde vivant est un acquis.

Avant Claude Bernard, la physiologie était chimie physiologique. Lavoisier et Laplace, avec des moyens essentiellement chimiques, démontrent les bases essentielles des échanges gazeux. Ce sont aussi des chimistes allemands qui ont permis le dosage de l'azote et la synthèse de l'urée. La physiologie, c'est aussi Magendie, le patron de Claude Bernard. Magendie avait fait de l'expérimentation et du scepticisme un dogme absolu, dogme qu'il transmit à Bernard. Ce faisant il s'opposait aux « vitalistes » et tout particulièrement à Bichat qui admettait l'existence de « forces vitales » tout à fait différentes des forces physico-chimiques. « D'abord les faits », résumait tout Magendie, « mais ensuite posez la bonne question » répliquait Bernard. Vieux débat, encore bien présent dans beaucoup de laboratoires.

C'est de la microbiologie que sont nées les biotechnologies. Avant Pasteur, il y avait un consensus, la possibilité de voir apparaître des éléments vivants spontanément, même après ébullition, dans un bouillon de viande par exemple. Les expériences de Needham, réfutées par Spallanzani, sont restées dans l'anonymat, dans l'ignorance où l'on était alors de l'existence des bactéries existant dans l'air ambiant. Cette même ignorance est à l'origine d'erreurs catastrophiques. Le grand Broussais, par exemple, avait fait de l'inflammation le *nec plus ultra* de la thérapie : « il faut suppurer pour guérir ». C'est Ignace Semmelwweis, obstétricien viennois, qui le premier eut l'intuition de l'asepsie, avant même que Pasteur ne découvre les bactéries; tous deux s'ignoraient d'ailleurs totalement.

2.2 1855-1865

Les quatre fondateurs de la biologie contemporaine sont nés presqu'en même temps : 1809 (à Shrewsbury, GB) pour Charles Darwin, l'inventeur de l'évolution des espèces, 1822 (à Heinendorf en Moravie) pour Gregor Johann Mendel, le fondateur de la génétique contemporaine, 1818 (à Saint-Julien, près de Villefranche-sur-Saône) pour Claude Bernard, l'inventeur de la physiologie, 1822 (à Dole, Jura) pour Louis Pasteur, l'inventeur de l'outil microbiologique. Tous ont publié leur « chef-d'œuvre » presqu'en même temps : Darwin, *L'origine des espèces*, 1859, un gros succès de librairie; c'est le 8 février et le 3 mars 1865 que Mendel publie son mémoire princeps sur l'hybridation des petits pois, et établit les bases théoriques de la génétique devant le Naturforschender de Brno (maintenant République Tchèque); Claude Bernard, malade, malheureux en ménage, se retire en 1865 dans ses vignes, à Saint-Julien, d'où il écrit et publie *L'introduction à la médecine expérimentale*; il est plus difficile de dater le « chef-d'œuvre » de Pasteur, il y en a tant, 1857, l'acte de naissance de la microbiologie est habituellement attribué à son mémoire sur la fermentation lactique présenté à la Société des Sciences de Lille, 1865 est la période où il découvre les fondements biologiques de la vinification et celle où il découvre la cause de la maladie du ver à soie, la première maladie infectieuse identifiée.

Darwin, a vécu longtemps retiré à Down dans le Kent, célèbre, mais discret (il n'en fut pas moins enterré à Westminster). Par contre, Mendel, trop en avance pour son temps finit ses jours comme supérieur (en 1868) de son couvent où il mourut en 1884. Il fallut attendre 1900 pour que les lois de l'hybridation végétale soient redécouvertes par les travaux indépendants d'un hollandais, Hugo de Vries, d'un allemand, Carl Correns et d'un Autrichien, Erich von Tschermak, et pour que, fort honnêtement, de Vries en attribue la paternité au moine de Brno. C'est également à ce moment qu'un anglais, William Bateson (qui propose le mot « génétique »), et un français, Lucien Cuénot, démontrent l'applicabilité des lois de Mendel au règne animal. Ce fut le début d'un déluge de publications confirmant les observations faites dans le Naturforschender. Claude Bernard, Pasteur surtout, sont tous deux morts couverts de gloire, primés (le prix de physiologie expérimentale a été décerné quatre fois à Claude Bernard), sénateurs impériaux (mais Pasteur, nommé en 1870, ne put jamais siéger). Tous deux académiciens français, ils se connaissaient, s'estimaient, et Bernard, l'aîné, fut pour Pasteur un soutien indéfectible.

2.3 LA FILIATION

2.3.1 Le néo-darwinisme et l'après darwinisme

Que Darwin représente une véritable césure dans l'histoire de la biologie ne souffre guère de contestation. Au cours de l'histoire du vivant, la pression évolutive sélectionne parmi les variants les espèces les plus résistantes, et crée ainsi de nouvelles espèces. La sélection naturelle, dit darwinienne, n'agit que par accumulation des variations (ou mutations) héréditaires. Ces changements sont héréditaires, variables et confèrent à ceux qui les portent un avantage transmissible. *L'origine des espèces* fut dès sa parution un « *best-seller* », mais aussi une source de polémiques acharnées, pas toujours de nature scientifique, lesquelles sont loin d'être achevées. Les polémiques de nature théologique n'ont pas leur place ici, mais il est un autre type de débat plus intéressant et qui porte sur le fait que le darwinisme n'explique pas toute l'évolution.

Le néo-darwinisme a consisté à introduire, 50 ans avant la découverte de l'acide désoxyribonucléique, ADN, la notion de mutations dans le « plasma germinatif ». Plus perverses seront les conclusions sociales que d'aucuns voudront en tirer, eugénisme rationnel de Francis Galton, déviations prônées par Trofim Lyssenko en support du régime stalinien. Il existe un continuum entre la première description de Darwin et la découverte de l'ADN, des gènes, et des bases de la génétique. L'intuition géniale de Darwin reçoit progressivement confirmation après la découverte des mécanismes qui président au brassage des gènes lors de la méiose et celle du polymorphisme de l'ADN ; elle est devenue une loi biologique à laquelle toute nouvelle donnée biologique doit se confronter. La physiopathologie elle-même se doit de tenir compte des affections résultant de conflits entre notre patrimoine génétique, adapté depuis toujours à un certain type d'environnement et le fait que depuis peu notre environnement a été radicalement modifié. Les séquençages effectués récemment ont démontré l'unicité du monde vivant, on retrouve en effet une majorité d'éléments communs aux génomes bactériens et au génome humain. L'unité historique du monde vivant est maintenant une évidence et plus que la pression évolutive c'est probablement là que se situe l'apport majeur de Darwin.

2.3.2 Naissance de la génétique

L'intuition de Mendel était comparable à celle de Darwin. Mendel en décrivant ses deux lois ne pouvait pas ne pas penser que l'hérédité avait un support chimique. Les étapes qui jalonnent le parcours entre Mendel et la séquence du génome humain sont bien connues :

– La découverte et le décompte des chromosomes[1], identification des différentes étapes de la division cellulaire par l'école allemande. Cette étape a été déterminée par les progrès en matière de microscopie.

1. Les bâtonnets nucléaires supports de l'hérédité furent baptisés chromosomes par Wilhem Waldmeyer en 1888, mais la découverte des chromosomes est généralement attribuée à Walter Flemming. Gène a été introduit par le danois Wilhem Johannsen.

– Thomas Morgan (1866-1945) de l'université de Columbia, est une étape à lui tout seul. C'est lui qui a apporté les preuves les plus décisives concernant le support matériel de l'hérédité en développant un modèle animal, toujours en usage, la Drosophile. C'est à lui que l'on doit la mise en évidence du brassage génétique à l'origine de la variabilité des individus et des espèces, et aussi des mutations pathogènes[1].
– À la suite de Morgan, la biologie moléculaire a connu une explosion qui a abouti à la découverte de la structure en double hélice de l'ADN par Crick et Watson, et à celle du code génétique et des mécanismes de la transcription et de la traduction des protéines. L'intuition de Mendel a été définitivement confirmée par le séquençage du génome et par les données de la génétique moderne. Les retombées en sont considérables.

2.3.3 Le développement scientifique de la physiologie

Claude Bernard était d'abord un expérimentateur, et on lui doit la découverte de plusieurs fonctions comme la fonction glycogénique du foie. Mais il a surtout été celui qui a établi les principes de la physiologie et de la médecine expérimentales et a fondé la physiologie. La notion de milieu intérieur restera le premier élément d'intégration à partir duquel seront découvertes la bioénergétique, les hormones (un concept, plus encore que des molécules), les vitamines et la respiration cellulaire. Parallèlement, la méthodologie prônée et expérimentée par Claude Bernard devient de routine dans tous les laboratoires du monde, et permet d'établir les fondements de la physiologie rénale, cardiaque, vasculaire, digestive et de la neurophysiologie. À la fin du XX^{e} siècle, la discipline a souffert de l'ambiance réductionniste et de la compétition avec la biologie moléculaire et la génomique. L'approche intégrée chère à Claude Bernard n'a plus été à la mode pendant une bonne trentaine d'années, dans le monde entier. Mais, au fur et à mesure de l'achèvement du programme génome et du développement

1. Morgan est devenu une unité, le centiMorgan, qui mesure, de façon statistique, les espaces séparant les gènes sur un chromosome.

de la bioinformatique et de la technologie transgénique, on a assisté à un retour en force des approches intégrées aux dépens des approches réductionnistes. Ce mouvement de balancier est une vieille tradition dans l'histoire des sciences.

2.3.4 La microbiologie et la naissance de la biotechnologie

La découverte des « microbes » et peut-être plus encore la mise sur pied des Instituts qui portent son nom, font de Pasteur le fondateur de cette discipline majeure qui est à l'origine de l'infectiologie et de la biotechnologie contemporaines[1]. Ce n'est pas le seul apport de Pasteur à la biologie. La « génération Pasteur » comprend des microbiologistes comme Calmette et Guérin (le BCG), Yersin (le bacille de la peste) ou Charles Nicolle (le typhus), mais aussi des biologistes moléculaires comme François Jacob et Jacques Monod qui utilisèrent le modèle microbiologique pour décrire la transcription. Les « microbes » sont devenus les modèles préférés d'étude du vivant. On connaît l'aphorisme de Monod selon lequel « ce qui est vrai pour *Escherichia Coli* est vrai pour l'éléphant ». Les Instituts Pasteur - l'un des premiers à être créé outre-mer, le fut en Australie vers 1890 par un élève de Pasteur, Adrien Loir - furent, à travers le monde, l'outil au moyen duquel se dissémina la méthode pastorienne. Il est évident que toutes les grandes avancées biologiques conceptuelles ou techniques ne sont pas nées dans ces Instituts[2], mais il est non moins discutable que c'est de Pasteur que datent les débuts de la recherche biotechnologie contemporaine.

1. On doit « microbes » à Emile Littré, l'auteur du célèbre Dictionnaire. Littré, qui avait fait sa médecine mais sans passer sa thèse, inventa spécialement le mot pour Pasteur en 1878.
2. Les trois découvertes sans lesquels le séquençage du génome humain n'aurait pas été possible sont les transcriptases inverses découvertes par Baltimore, Temin et Dulbecco, les enzymes de restriction découverts par Arber, Nathans et Smith et la PCR inventée par K. Mullis. Tous ont reçu le Prix Nobel, tous sont anglo-saxons, aucun n'est pastorien.

Chapitre **3**

Données de base et sémantique élémentaire

3.1 LES UNITÉS DU VIVANT

3.1.1 Définir la vie

La vie est programmée, l'être vivant est un concentré d'énergie, face à un extérieur inorganisé, faible en énergie. Définir la vie ou en résumer les propriétés fondamentales n'est pas simple, Claude Bernard, en avait déjà souligné la complexité, et qu'il y fallait, au moins, cinq éléments : l'organisation, la reproduction, la nutrition, l'évolution et le développement et enfin le trépied vieillissement – maladie – mort.

Cette définition recoupe assez bien celle, plus moderne, de Daniel Koshland (Koshland 2003), ancien éditeur de Science (tab. 1). D'autres ont repris ces données, et considèrent comme déterminants :

– la capacité de se reproduire, la vie est un automate auto-entretenu par des molécules informationnelles (ce qui regroupe les « piliers » (i), (ii) et (v) de Koshland, tableau 1) ;

– la nutrition et le métabolisme, la vie est un système chimique auto-entretenu par de petites molécules chimiques (pilier (iv));
– la complexité de l'organisation macromoléculaire, la vie utilise le pouvoir structurant des liaisons faibles permises par les caractéristiques de la matière;
– tous les organismes reposent enfin sur un coefficient de sécurité énorme comme en témoignent les quantités phénoménales de graines, œufs, spermatozoïdes produits par tous les organismes et dilapidés dans la nature.

Les Américains, qui misent actuellement beaucoup sur leurs programmes spatiaux et sur l'astrobiologie, ont une réelle demande pour une définition de la vie qui en permette l'identification; la plus concise serait « la vie est un système chimique auto-entretenu, capable de subir une évolution darwinienne ».

3.1.2 Cellules et organismes vivants

Si diverse soit-elle, la vie possède une unité fondamentale, la cellule, qui est délimitée par une membrane et contient tous les composants permettant à la cellule d'être vivante, au sens défini plus haut, c'est-à-dire, au minimum, outre la membrane externe faite de phospholipides et de protéines membranaires, un génome et un cytoplasme où se trouvent les enzymes responsables du métabolisme énergétique ainsi qu'une machinerie complexe qui sert à fabriquer les protéines, les ribosomes. Les cellules sont toujours filles d'autres cellules et ne proviennent jamais de l'assemblage de leurs constituants. Il y a deux types de cellules vivantes : les procaryotes[1] et les eucaryotes. À la différence des procaryotes, les eucaryotes possèdent un noyau délimité par une membrane et qui contient la majeure partie du matériel génétique (fig. 1).

1. Il y a deux types de procaryotes, les eubactéries qui sont les bactéries les plus connues et les plus ordinaires, et les archéobactéries qui ont une structure membranaire très différente de celle des eubactéries leur permettant de pousser dans des milieux extrêmes (méthane, hautes températures).

Tableau 1 LES SEPT PILIERS DE LA VIE (KOSHLAND 2003).

1	La programmation à la fois des ingrédients et de leur mode d'emploi, celle-ci est présente dans l'ADN
2	La possibilité d'improvisation, sur un long court, grâce à l'outil qu'est l'évolution darwinienne
3	La compartimentalisation soit dans la peau, soit dans les membranes, qui permet le maintien de concentrations efficaces à l'intérieur de l'être vivant
4	Sur un plan énergétique, les êtres vivants sont des systèmes ouverts en déséquilibre et les changements d'entropie sont compensés par l'énergie solaire
5	Les systèmes vivants peuvent se régénérer totalement, la naissance d'un nouvel enfant est un des aspects de cette restitution *ad integrum*
6	Il y a adaptabilité, la faim, les cals au pied en sont des exemples, ce type d'adaptabilité est rapide, épigènétique et complète l'évolution darwinienne
7	La séclusion, les cascades métaboliques sont isolées tout comme des conducteurs électriques, ce qui leur permet de fonctionner collectivement dans des conditions de concentration optima

La forme de vie la plus simple est la bactérie qui est un être monocellulaire possédant une seule copie (les bactéries sont haploïdes) d'un nombre peu élevé de gènes (≈ 4 000) et se reproduisant de façon non sexuée. Il y a 2,7-1,9 milliards d'années le monde était entièrement

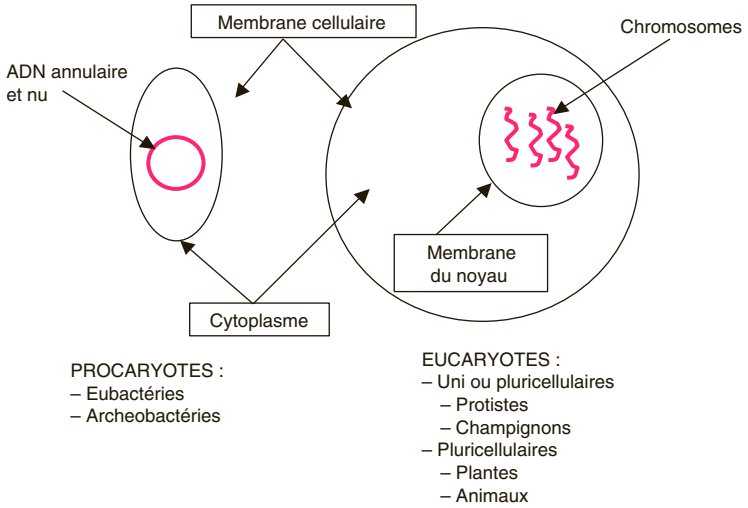


Figure 1 La cellule, unité élémentaire du vivant.
Les deux différences majeures entre procaryotes et eucaryotes sont l'absence de noyau et de membrane nucléaire chez les procaryotes ainsi que le fait que l'ADN des procaryotes est annulaire.

bactérien. Les bactéries ont eu une évolution très lente qui s'est faite souvent par transferts de gènes d'une espèce à l'autre (dit de façon horizontale par opposition au transfert vertical chez les eucaryotes chez qui le changement de structure du génome passe d'une génération à l'autre). La rapidité de leur croissance, les facilités que l'on a à les conserver en ont fait un outil idéal pour les biotechnologies. Les bactéries peuvent être source d'infections pathogènes, mais elles sont plus souvent, chez l'homme, des parasites utiles (à la digestion par exemple). Il faut savoir enfin que les bactéries ont été la source de l'oxygène de l'air et de l'eau.

Les eucaryotes peuvent être uni- ou multicellulaires. La levure est un eucaryote monocellulaire qui possède un noyau et 16 chromosomes, c'est un modèle très courant en biologie moléculaire. Il existe un certain nombre de modèles pluricellulaires utilisés dans la plupart des laboratoires pour chacun des règnes animaux ou végétaux, la *Drosophile* pour les insectes, *Danio rerio* (le poisson zèbre) pour les poissons, *Arabidopsis thaliana* (le cresson) pour les plantes, la souris pour les mammifères.

La cellule est l'unité la plus élémentaire du vivant, mais il existe des formes intermédiaires entre les cellules et la matière inerte à la frontière du vivant (fig. 2).

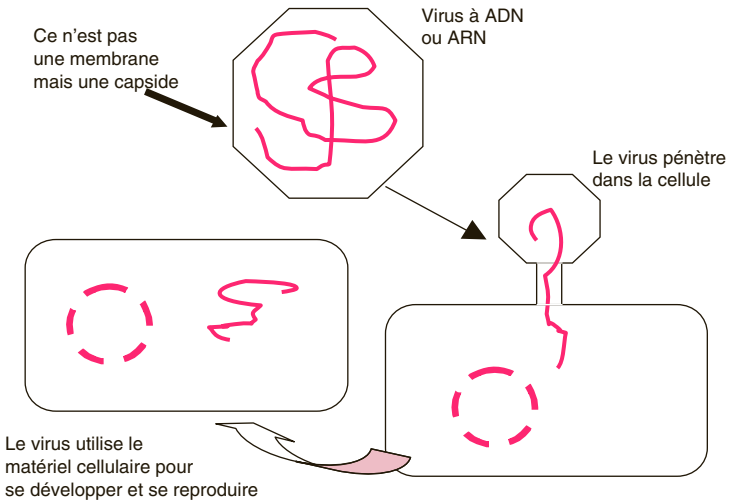


Figure 2 Virus.

Les virus sont à la limite du monde des vivants en ce qu'ils ne possèdent pas de membrane externe, ils sont délimités par une enveloppe protéique ou une capside. Il y a des virus à ARN et des virus à ADN. Les virus sont nécessairement des parasites, et leurs hôtes leur servent à se reproduire. En rouge et en trait d'union l'ADN de l'hôte.

(1) Les **virus** sont les plus importants, ce ne sont pas des cellules vivantes, car ils n'ont pas de membrane externe et sont délimités par une capside ou une enveloppe protéique. Ils ont un matériel génétique qui peut être de l'ADN ou de l'ARN, mais n'ont ni ribosomes, ni métabolisme propre et sont obligatoirement parasites et demandent à être hébergés par un hôte cellulaire dont ils utilisent les ressources pour se reproduire. Les plus connus des virus sont les rétrovirus qui sont des virus à ARN qui utilisent l'ADN de l'hôte pour se répliquer, le plus célèbre des rétrovirus est le virus du SIDA. L'hôte du virus peut être une bactérie, le virus s'appelle alors bactériophage ou phage.

(2) Les **viroïdes** sont des virus à ARN sans revêtement protéique qui n'infectent que des plantes.

(3) Les **plasmides** sont des molécules d'ADN capables de se répliquer qui n'ont pas de phase extracellulaire et ils vivent en permanence dans leurs hôtes.

(4) Les **transposons** ne peuvent se répliquer et sont des molécules d'ADN incorporées dans l'ADN de leur hôte, ils doivent sauter d'un hôte à l'autre pour survivre.

(5) Les **prions** sont des protéines dont la structure spatiale est anomale, ils peuvent infecter le tissu nerveux et y développer la maladie de la vache folle et ne contiennent pas d'acides nucléiques.

3.2 STRUCTURE DE L'APPAREIL GÉNÉTIQUE

L'unité de base de la biologie moléculaire est la molécule d'acide désoxyribonucléique, ADN, seul support chimique connu de l'hérédité. L'ADN est l'outil de base, véritable fondement de la biotechnologie. Le génome est l'ensemble du matériel génétique d'un individu (ou d'une cellule) dont il constitue le génotype, le génotypage représente l'acte technique qui permet de déterminer un génotype donné.

3.2.1 Structure de l'ADN

La molécule d'ADN est la plus grosse molécule de l'organisme, sa masse moléculaire variant de 3,3. 10^9 chez l'homme à 10^5 chez les bactéries[1]. C'est une double hélice faite de deux monomères enroulés les uns sur les autres (fig. 3a & b). Cette molécule est composée d'une succession répétitive de nucléotides composés eux-mêmes, dans l'ordre, d'une base purique ou pyrimidique, d'un sucre, le Désoxyribose, et d'un acide phosphorique. La séquence sucre-acide phosphorique est la séquence invariable de l'ADN. Les nucléotides ne se distinguent les uns des autres que par leur base. Il y en a, dans l'ADN, quatre types : l'Adénine, A, et la Guanine, G, qui sont des purines, la Cytosine, C, et la Thymine, T, qui sont des pyrimidines. La Thymine est la seule base spécifique de l'ADN, elle ne se retrouve pas dans l'Acide RiboNucléique, ARN, où elle est remplacée par l'Uridine, U. Ces bases ont une structure particulière qui confère à la molécule ses propriétés :

– (1) Elles ont les unes pour les autres des affinités spécifiques et s'apparient d'une manière invariable T-A et C-G, c'est-à-dire que les bases puriques s'apparient aux bases pyrimidiques, mais ne peuvent contracter de liaisons entre elles, et *vis versa*. Cette propriété donne à la molécule sa stabilité et permet l'établissement de liaisons d'une hélice (ou monomère) à l'autre. Ce sont des liaisons hydrogène, mais leur nombre varie selon le couple de bases concerné (3 liaisons pour C-G ; 2 pour T-A). Ces liaisons sont très spécifiques et puisque G ne peut se lier qu'à C et T qu'à A, les deux brins d'ADN ont une structure en image, et sont dits complémentaires, ce qui veut dire que la structure de l'un permet de prédire celle de l'autre.

1. La taille du génome croit assez régulièrement au fur et à mesure que l'on avance dans l'échelle de complexité dans l'évolution des eucaryotes, mais ceci est loin d'être linéaire et il y a de très nombreuses exceptions à la règle. Le nombre de chromosomes et le nombre de gènes croissent également sur cette échelle, mais les exceptions sont ici encore plus nombreuses. Le génome du riz est plus petit (430 mega paires de base) que celui de l'homme 3 300 Mpb), mais il contient environ deux fois plus de gènes (21 000 chez l'homme, 45 000 pour le riz).

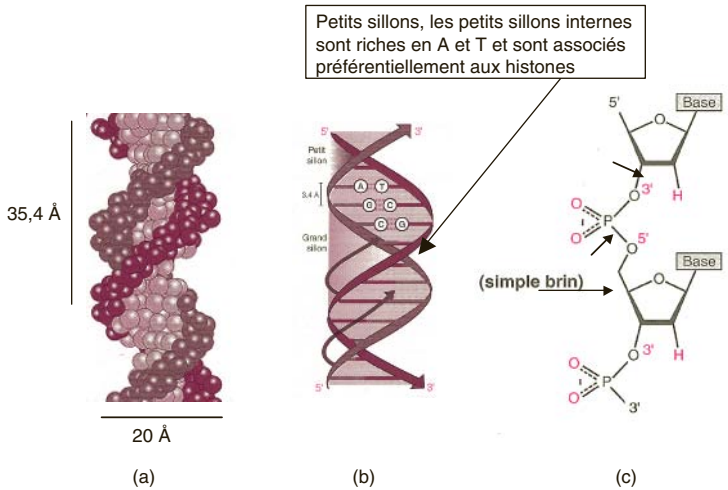


Figure 3 Molécule d'ADN.

(a) Modèle moléculaire. (b) Schéma montrant les liaisons hydrogène Adénine = Thymine, A = T ou Guanine = Cytosine, G = C qui maintiennent la structure hélicoïdale. En (a) et (b) on identifie la double hélice polynucléotique avec les bases puriques ou pyrimidiques à l'intérieur de la molécule et les deux sillons, le grand (12 Å de large) et le petit (6 Å de large). La structure se répète tout le long de la molécule tous les 10,5 nucléotides, c'est-à-dire tous les 35,4 Å. (c) Structure chimique des nucléotides d'un brin d'ADN montrant les liaisons phosphodiesters entre le désoxyribose et le phosphate (flèches) et la manière conventionnelle par laquelle on oriente la molécule, on dit 5'-3' ou l'inverse en fonction de la position d'un des carbones du sucre. Les deux hélices sont en effet antiparallèles, l'une étant orientée en 3'-5' et l'autre en 5'-3'. Figures reprises de Weinman S. et Méhul P. Toute la biochimie. Dunod éd. 2004.

– (2) Le nucléotide représente l'unité de longueur utilisée pour mesurer un fragment d'ADN. Un nucléotide contenant par définition une base, on dira qu'un fragment monobrin d'ADN de 100 nucléotides a

une longueur de 100 bases, ou 1 kb, et l'on dira que le fragment apparié double brin est composé de 100 paires de bases, pb.

– (3) L'enchaînement désoxyribose-acide phosphorique-base est asymétrique. Sur un plan strictement chimique, la molécule peut donc être orientée. Il faut dès lors qualifier cette orientation. On utilise pour cela la structure de l'anneau ribose qui possède deux radicaux hydroxyles libres, l'un en position 3', l'autre en 5' (fig. 3c). Par convention un fragment d'ADN commence en 5' et se termine en 3'. On dira qu'un fragment, monobrin, est en 5'-3'. Ce fragment sera également dit sens ou codant parce que sa séquence ressemble à celle de la protéine, mais il est important de savoir que le fragment codant n'est pas celui qui sert de matrice à la synthèse des transcrits, on y reviendra.

Sur un plan fonctionnel, l'ADN se divise en deux parties, l'ADN non codant ou anonyme qui représente la majorité de l'ADN, et l'ADN des gènes. La distinction entre les deux n'est ni claire ni définitive. L'ADN codant est relativement bien délimité, mais il y a dans l'ADN anonyme à la fois de nombreuses séquences régulant l'activité des gènes et faisant de ce fait partie de la définition du gène, et des séquences codantes non encore identifiées. On reviendra sur ces ambiguïtés plus loin.

3.2.2 Les ARNs

La cellule synthétise plusieurs types d'ARN : l'ARN messager, ARNm, quantitativement le moins important mais qui va reproduire le code génétique sous une forme telle que l'information pourra être transmise hors du noyau dans le cytoplasme où elle servira à synthétiser les protéines; l'ARN de transfert qui servira à transporter spécifiquement chacun des acides aminés vers le lieu où se fabriqueront les protéines; l'ARN ribosomal, le plus abondant, constituant essentiel des ribosomes, il sert avec les protéines ribosomales à catalyser et à diriger la synthèse protéique; les microARNs de découverte toute récente (ARN interférence) et qui jouent un rôle important dans la régulation de la transcription et de la traduction.

Une polymérase spécifique correspond à plusieurs de ces ARN, RNA polymérase I pour l'ARN ribosomale, RNA polymérase II pour l'ARN messager, RNA polymérase III pour l'ARN transfert.

L'ARN a la même orientation 5'-3' que le brin antisens de l'ADN, il en a la même structure à une exception près, l'ARN ne possède pas de thymidine, laquelle est remplacée par une uridine (U). L'ARN naissant a la forme d'un simple brin qui peut se replier sur lui-même du fait des appariements dus aux bases complémentaires qui peuvent se faire face.

3.2.3 Les chromosomes

La représentation des paires de chromosomes s'appelle caryotype, les chromosomes y sont rangés par ordre de taille (fig. 4). Établir un caryo-

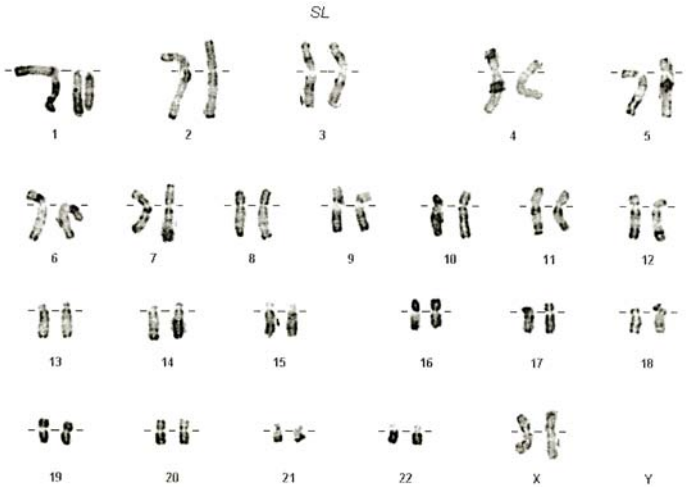


Figure 4 Caryotype humain (dû à l'obligeance du Professeur Stanislas Lyonnet que nous remercions bien vivement).

type consiste à isoler les chromosomes et ensuite à les classer selon leur taille, la position de leur centromère et leur profil de bandes spécifiques. Chaque espèce vivante possède un caryotype qui lui est propre.

a) Chez les procaryotes

Les procaryotes sont des petites (2-3 µm) cellules dépourvues de membrane nucléaire, et le matériel génétique est situé dans le même compartiment que le cytoplasme. Il y a deux catégories de procaryotes, les eubactéries et les archéobactéries. Ces deux domaines d'être vivants sont aussi distincts l'un de l'autre qu'ils le sont du troisième domaine, les eucaryotes. Eu- et archéobactéries se distinguent formellement par la structure de leur membrane externe et par leur génome. Les procaryotes se multiplient plus vite que les autres êtres vivants et leur génome est circulaire. Ils ne possèdent qu'un seul chromosome par cellule et n'ont donc qu'une seule copie des gènes par cellule. Cependant les bactéries peuvent contenir un matériel extra-génétique circulaire, les plasmides. Le génome d'*Escherichia Coli*, par exemple, contient 10 fois moins de gènes que le génome humain.

b) Chez les eucaryotes

Les chromosomes sont des structures intranucléaires en bâtonnet qui comprennent tout le matériel nécessaire à la transmission de l'hérédité. La cellule humaine somatique est diploïde (2n), ce qui veut dire que tous les chromosomes qui la composent vont par paire. Il y a 22 paires de chromosomes autosomaux, c'est-à-dire non sexuels, et une paire qui détermine le sexe, XX pour le sexe féminin, XY pour le sexe masculin, ce qui veut dire que cette dernière paire est la seule paire qui soit asymétrique et que le chromosome Y est un marqueur du sexe masculin, en tout 46 chromosomes. Les gamètes ou cellules germinales, c'est-à-dire soit les spermatozoïdes, soit les ovocytes, sont haploïdes, c'est-à-dire 1n, leurs chromosomes sont tous uniques.

Les chromosomes sont des structures visibles à certains moments du cycle mitotique. Ils ont une taille de l'ordre du µm. Ils sont composés d'un centromère d'où partent quatre bras, deux longs, q, deux courts, p.

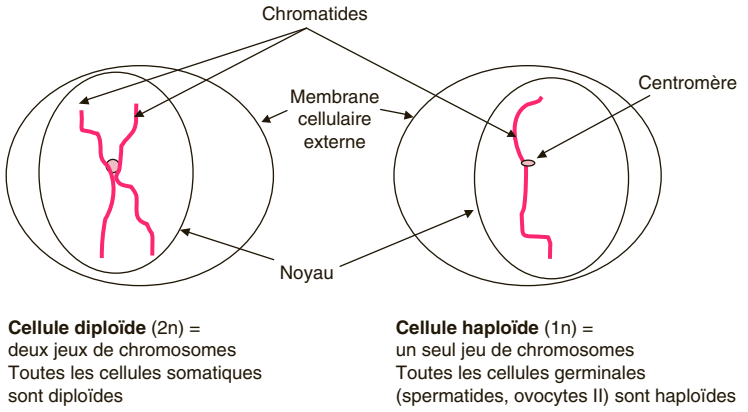


Figure 5 Cellule diploïde (2n) et cellule haploïde (1n).
Le « n » signifie le nombre de chromosomes présents dans la cellule. Les êtres humains possèdent 46 chromosomes dont 22 paires (2 × 22) de chromosomes autosomaux et une paire de chromosomes sexuels, XX chez les femmes XY chez les hommes, soit, au total 23 paires de chromosomes. Les chromosomes des cellules diploïdes sont formés par deux copies d'un chromatide et donc deux copies de chaque gène. Dans les cellules germinales haploïdes les chromosomes sont formés par un chromatide unique.

En utilisant certaines colorations, on peut mettre en évidence des bandes, spécifiques du colorant utilisé (bande Q pour la Quinacrine par exemple), très reproductibles, et qui permettent d'établir de véritables cartes. Les chromosomes sont divisés en secteur et ces secteurs sont numérotés après coloration. On établit en routine des cartes chromosomiques chez l'homme, appelées caryotypes, afin de détecter certaines maladies héréditaires qui sont accompagnées d'anomalies chromosomiques visibles, comme le mongolisme par exemple. Le caryotype de l'homme est un examen essentiel en génétique médicale, il permet, entre autre, de détecter des anomalies chromosomiques comme la présence de trois chromosomes homologues qui définit les trisomies.

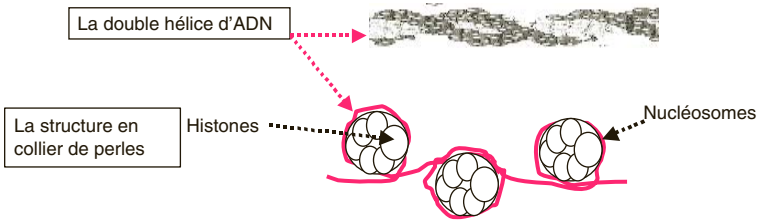


Figure 6 Ruban de chromatine.

Un nucléosome contient 200 pb d'ADN, et 9 histones (une paire de H2A, H2B H3 et H4 et un seul H1 lequel possède deux bras qui lui permettent de se lier aux nucléosomes voisins lors du compactage). La chromatine est un matériau nucléaire composite fait d'ADN et de protéines régulant la transcription. Elle constitue les chromosomes. La chromatine a une structure spatiale complexe, variable selon le stade où se trouve la cellule. Cette structure spatiale joue un rôle physiologique considérable. C'est un ruban composé de sous-unités, les nucléosomes composés eux-mêmes d'une molécule d'ADN et de protéines. Les nucléosomes sont des cylindres dont le cœur est un ensemble d'histones autour duquel s'enroule l'ADN en deux tours. L'histone H1 se lie à certaines séquences de l'ADN. La chromatine peut être compactée (hétérochromatine) ou désenroulée (euchromatine), c'est dans ce dernier cas que les gènes sont activés

Les télomères sont des structures particulières, très anciennes sur un plan évolutionniste, situées aux extrémités des chromosomes. Le maintien de cette structure est assuré après chaque mitose grâce à un enzyme, la télomérase qui est une réverse transcriptase associée à une matrice ADN qui fournit le modèle sur lequel la transcriptase assurera la reconstitution *ad integrum* de l'extrémité ADN selon une séquence toujours la même pour une espèce animale donnée. La régulation de cette réaction est complexe, elle est assurée par tout un groupe de protéines dites « *telomere-associated proteins* ». Le maintien d'un télomère est important pour empêcher les extrémités des différents chromosomes de fusionner. L'érosion du télomère est liée au vieillissement et en

particulier au vieillissement replicatif. L'activité télomérase est liée en pathologie à une activité proliférative élevée, dans certains cancers en particulier (fig. 5 à 7).

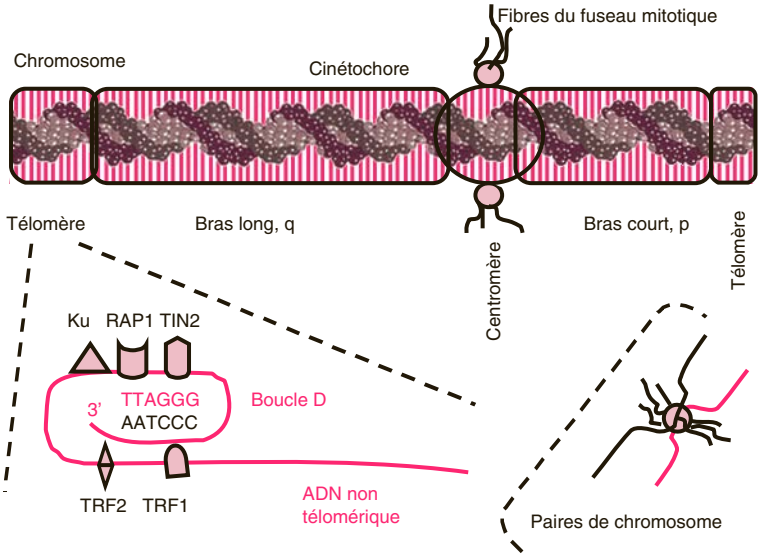


Figure 7 Structure d'un chromatide et d'un chromosome.

Télomère. En haut les deux bras (q et p) d'un chromatide séparés par le centromère. L'ensemble comprend une molécule d'ADN (schématisée sur la figure) et toute une série de protéines (dont les histones et les facteurs de transcription, voir texte). Certaines colorations permettent d'identifier des bandes schématisées ici en rouge, ces bandes servent à la nomenclature (numérotée) topographique des gènes. Les régions situées aux extrémités des chromatides sont les télomères. Cette structure est maintenue grâce à l'activité de la télomérase qui est une transcriptase réverse associée à une matrice (un modèle) ARN. Protéines de liaison modulant la télomérase : les deux « *Telomeric repeat binding factor* », TRF-1 et -2 se lient à l'ADN, et plusieurs autres protéines, Ku, Rap1, Tin2 qui régulent l'activité des TRF. Le maintien d'un télomère normal est indispensable à la structure du chromosome.

3.2.4 La chromatine

L'un des problèmes majeurs que la cellule doit résoudre est de ranger la chromatine, car cette structure complexe fait un mètre de long et il lui faut pour se ranger dans un chromosome se replier sous une forme très compacte, le rapport de compaction de l'ADN est de l'ordre de 10 000. La chromatine existe sous une forme compacte, au cours de l'interphase, et dans ces cas les chromosomes ne sont pas visibles, c'est l'hétérochromatine. L'organisation spatiale de la chromatine est un élément majeur dans la régulation de la transcription, et la transcription ne peut pas avoir lieu dans l'hétérochromatine.

Ce matériel comprend non seulement l'ADN (une molécule par chromosome), mais aussi toutes les protéines qui en contrôlent l'activité, l'organisation spatiale et l'intégrité. L'ensemble ADN-protéines chromosomales s'appelle chromatine. On ne peut pas envisager de créer la vie avec le seul ADN, et il faut également disposer des protéines responsables de la transcription. Pour se mettre en route la machinerie héréditaire doit disposer de toutes les protéines régulatrices, fragiles, présentes dans les chromosomes. La chromatine est faite d'ADN et de protéines, ces dernières pouvant être des protéines naissantes, des facteurs de transcription régulant l'expression génique, et les histones dont le degré d'acétylation est responsable de la structure spatiale de la chromatine (voir plus loin « Les grands mécanismes. Transcription »). L'unité élémentaire de la chromatine est le nucléosome qui est une structure cylindrique de quelques nm dont le centre est composé d'histones et autour duquel s'enroule la molécule d'ADN. Les histones ne sont pas les seules protéines à être associées à l'ADN, il y en a d'autres qui servent essentiellement à réguler la transcription (fig. 6).

3.2.5 Quelques notions sémantiques

a) Génotype et phénotype

On connaît depuis de nombreuses années les bases chimiques de l'hérédité. Elles sont résumées dans le schéma de la figure 8. Je ressemble à mes parents et à un être humain parce que les molécules chimi-

ques avec lesquelles je suis fait (ou faite) ressemblent à celles dont mes parents sont faits et à celles dont les autres humains sont faits, et cette ressemblance m'a été transmise, lors de la conception. La morphologie d'un être et ses principales fonctions vitales sont l'expression de sa composition en protéines et la fonction des protéines dépend de leur arrangement spatial. La structure chimique des protéines, c'est-à-dire leur composition en acides aminés, ne détermine pas directement leur arrangement spatial (c'est-à-dire leur structure tertiaire). Le repliement de la chaîne des protéines ne se fait pas en fait au hasard, par essais successifs, comme l'avait postulé Anfinsen à propos de la ribonucléase pancréatique. Cela est vrai dans les conditions définies dans l'expérience d'Anfinsen, mais ce n'est pas vrai d'une façon générale *in vivo* dans le cytoplasme, ou… dans le canal biliaire, dont la composition n'est pas la même que celle de l'eau ou du tampon utilisé par Anfinsen.

L'arrangement spatial des protéines se fait selon un programme qui dicte la façon dont le repliement de la séquence primaire doit se faire pour aboutir à la structure tertiaire finale, et donc à la fonction. Ce processus est guidé à la fois par l'environnement où se trouve la protéine (protéines voisines, structure cellulaire) et par des protéines auxiliaires – les protéines-chaperons – qui indiquent la façon dont ce repliement doit se faire. Ces protéines-chaperons[1] agissent un peu à la façon d'un métier à tisser, mais le processus exact par lequel elles fonctionnent est encore mal connu. Les protéines-chaperons évitent la perte de temps, et d'énergie, que représente l'exploration de toutes les conformations rendues possibles par la structure primaire, leur présence explique pourquoi les macroprotéines d'un organisme donné ont un certain nombre de points en commun.

La séquence primaire se transmet héréditairement, sous forme d'un code génétique. Le code génétique est lui-même constitué par un arrangement chimique particulier de l'Acide Désoxyribonucléique, ADN. Le Génotype c'est ce code, c'est la constitution génétique d'un indi-

1. Les principales protéines chaperons, ou chaperonines, sont les *heat-shock proteins*, HSP, HSP 70 en particulier, et le complexe GroE (ou HSP60/HSP10).

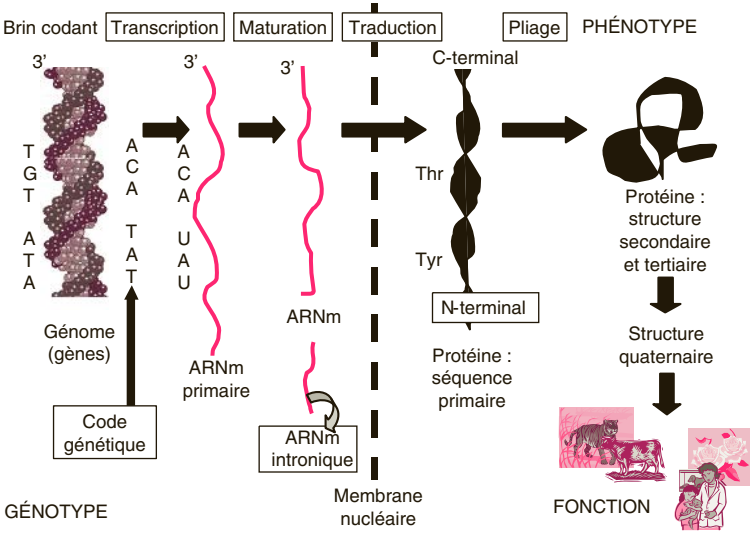


Figure 8 Génotype et phénotype.

Le génotype, c'est la constitution génétique d'un organisme, le génotype est par définition héréditaire et transmissible, il est présent dans la molécule d'ADN. Seule une petite portion de l'ADN, les exons et les introns des gènes, va être transcrite, c'est-à-dire transformée en information ARN. L'information primaire inclut les introns. Par la suite, dans le noyau, l'ARN primaire, ou pré-ARN messager, ARNm, va perdre la portion intronique transcrite et donner naissance à l'ARNm. Ce dernier subira en même temps un processus de maturation avant d'être exporté hors du noyau et traduit en protéine. Cette dernière reflète quatre niveaux d'organisation : la séquence primaire reste inchangée et reflétera très exactement la séquence du code génétique; la structure secondaire est spatiale et dépend des liaisons hydrogène existant entre les groupes peptides, il y a deux types de structure secondaire, les alpha-hélices et les structures en feuillet bêta antiparallèles; la structure tertiaire dépend du caractère hydrophilique ou hydrophobique des résidus amino acides, les premiers ont tendance à se trouver à l'extérieur de la molécule, au contraire des seconds, cette tendance structurelle est inversée lorsque la protéine est intra-membranaire; la structure quaternaire enfin consiste en l'assemblage des sous-unités. C'est la protéine qui créera la fonction, le trait, le signe clinique, mais elle ne le fera qu'assemblée à d'autres protéines. On parle alors de réseau fonctionnel.

vidu, telle qu'elle est dans le code supporté par l'ADN. Le Phénotype, c'est la manifestation apparente de ce génotype, ce peut être une fonction cellulaire, une fonction physiologique, ou la fonction d'une protéine. Il existe quatre groupes de protéines : les protéines de structure (ex. les protéines contractiles), les enzymes qui agissent sur des substrats (on entend par substrat une molécule qui est modifiée par l'action d'un enzyme), les protéines régulatrices (ex. les facteurs de transcription) et enfin les protéines de transport (ex. l'hémoglobine qui transporte l'oxygène).

b) Comment s'orienter en génétique ?

Dans la pratique courante les problèmes d'orientation d'une séquence nucléotidique sont souvent un obstacle majeur à la compréhension. Comme nous l'avons énoncé précédemment (fig. 3) l'orientation des deux brins d'ADN est donnée par la position des radicaux hydroxyles du désoxyribose. Le brin 5'-3' est dit codant parce qu'il a la même séquence et la même orientation que l'ARNm. Le décodage par l'ARN polymérase se fera physiquement sur le brin non codant 3'-5' complémentaire. On peut réaliser des constructions artificielles sur lesquelles on place un promoteur très actif qui va transcrire le brin codant, l'anti-ARN messager obtenu est appelé anti-sens, il peut servir de témoin négatif, mais aussi s'hybrider et « neutraliser » l'ARNm actif (fig. 9). C'est un des outils potentiels de la thérapie génique.

c) Cis- et trans-régulation

La transrégulation est la régulation contrôlée par des facteurs diffusibles qui inter-réagissent avec les séquences régulatrices d'ADN situées en amont du point d'initiation de la transcription du gène, les facteurs en question agissent en trans. La cisrégulation est l'opération ADN-ADN qui fait suite, et par laquelle la séquence amplificatrice activée (en plus ou en moins) va réguler l'activité du promoteur et ensuite contrôler la transcription. Elle a lieu sur le même chromosome (fig. 10).

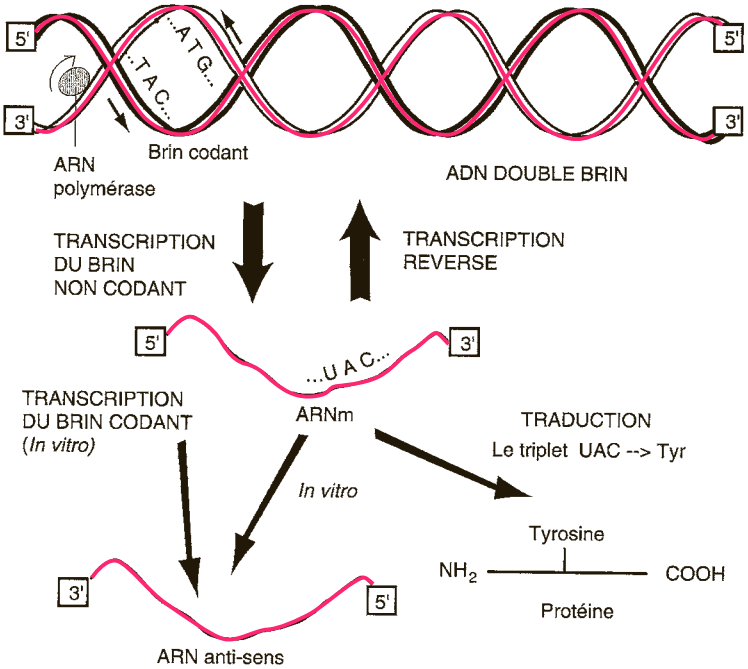


Figure 9 Comment s'orienter en génétique ?

L'ADN possède deux brins organisés en miroir. Le brin 5'-3' est dit codant parce qu'il possède le code qui sera copié, mais ce n'est pas sur lui que se fera la transcription (voir aussi figure 1). Par contre le brin antiparallèle non codant est celui sur lequel se fera la transcription par l'ARN polymérase. L'ARN messager, ARNm, aura la même séquence que le brin codant et la séquence en miroir du brin non codant, à une exception près, T n'existe pas dans les ARN et doit être remplacé par U. La traduction se fait dans le cytoplasme. Il faut savoir qu'une protéine est également orientée et possède une extrémité N terminale et une extrémité C terminale. L'acide aminé situé en N terminal est codé par un triplet lui-même en 5'. On peut artificiellement fabriquer un ARN antisens soit à partir de l'ARNm, soit en transcrivant le brin non codant de l'ADN. Un tel ARN peut être utilisé comme témoin négatif. Il peut aussi exister *in vivo*, tout au moins pour certains gènes.

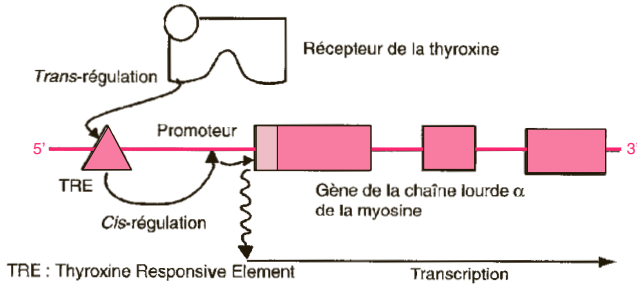


Figure 10 Régulation de la transcription en cis ou en trans.
Une régulation en trans est le fait d'un facteur de transcription, qui est une protéine, qui inter-réagit avec une séquence ADN nucléotidique située en amont du gène, en 5', et qui régule la transcription. Cis régulation ADN-ADN, se dit de la régulation de l'activité du promoteur sur la transcription par les séquences amplificatrices non codantes situées sur le même chromosome.

3.2.6 Le gène

a) Définitions

Le code génétique ne peut être transcrit que lorsqu'il est présent dans un fragment très particulier de l'ADN, appelé gène. Classiquement, on appelle gène un ensemble de nucléotides qui contient toute l'information nécessaire pour transcrire un fragment d'ADN en ARN. Certains de ces ARN, les ARNm, seront traduits en protéines (fig. 11), d'autres resteront ARN ribosomaux, ARNr, transferts, ARNt[1] ou servant à certains mécanismes de régulation (microARN, miRNA), ces derniers ARN serviront néanmoins indirectement à la traduction en protéines. Le gène peut de définir de deux manières soit en le considérant comme une entité fonctionnelle, c'est-à-dire l'unité de base de l'hérédité, soit en le considérant comme une entité physique avec une position fixe sur le chromosome, le locus, position que le séquençage du génome a permis de déterminer avec précision. En fait c'est surtout la première définition

1. Les ARNr et les ARNt sont synthétisés dans le nucléole.

qui est la bonne, un gène se définit par sa fonction c'est-à-dire le phénotype dont il est porteur. Il faut donc considérer le gène comme étant la séquence ADN qui contient toutes les informations nécessaires à la fabrication d'une protéine ou d'un ARN (Lodish 2000). La longueur d'un gène est donc très variable, et il y a de nombreux gènes qui se chevauchent dans un même locus (Kleinjan DJ, *in* Wright 2007).

Le gène comprend trois groupes de séquences ADN : (1) les séquences codantes ou exons, celles qui possèdent le code génétique (il existe néanmoins dans ces séquences une portion non codante située dans le premier codon) ; (2) les éléments qui vont réguler la transcription, c'est la zone régulatrice située généralement, mais pas exclusivement, en amont au niveau de l'extrémité 5' (il peut y avoir des éléments régulateurs introniques), certains de ces éléments peuvent se situer physiquement très loin de la séquence codante ; les facteurs de transcription, ont la capacité de se lier spécifiquement à certaines de ces séquences et les miRNA ; (3) enfin la structure fonctionnelle d'un gène inclut des nucléotides qui n'ont aucun rôle connu particulier et qui sont simplement inclus dans la structure probablement par hasard (les introns) (fig. 11 & 12).

La définition physique un peu simpliste du gène est remise en cause pour de multiples raisons (qui sont énumérées dans l'Addendum 1), comme le fait que, sur un plan structurel, les gènes peuvent se chevaucher et que le phénomène d'épissage se produit également pour les protéines. Les définitions récentes insistent donc toutes sur la fonction, plus que sur la structure. Celle de Gerstein (2007) se base sur les résultats du programme ENCODE[1], et définit le gène comme étant « ensemble de séquences

1. Ce programme tout récent est en train de bouleverser notre vue traditionnelle des choses. Il se propose d'établir la carte de l'activité transcriptionnelle et de sa régulation en utilisant la technique des *tiling arrays* (littéralement alignement de carrelages). C'est ce programme qui a découvert les éléments nouveaux qui ont remis en cause la définition du gène à savoir : il y a dans la portion anonyme du génome des séquences, qui n'étaient pas identifiées auparavant comme des gènes, et qui sont de fait transcrites en ARN mais pas en protéines, dans de nombreux gènes le site d'initiation de la transcription peut se trouver très loin des exons, dans certaines conditions il y a transcription à l'envers, c'est-à-dire transcription du brin 3'-5' dit non-codant, enfin la proximité physique ne suffit pas pour authentifier le gène ciblé par une séquence régulatrice (Gerstein 2007).

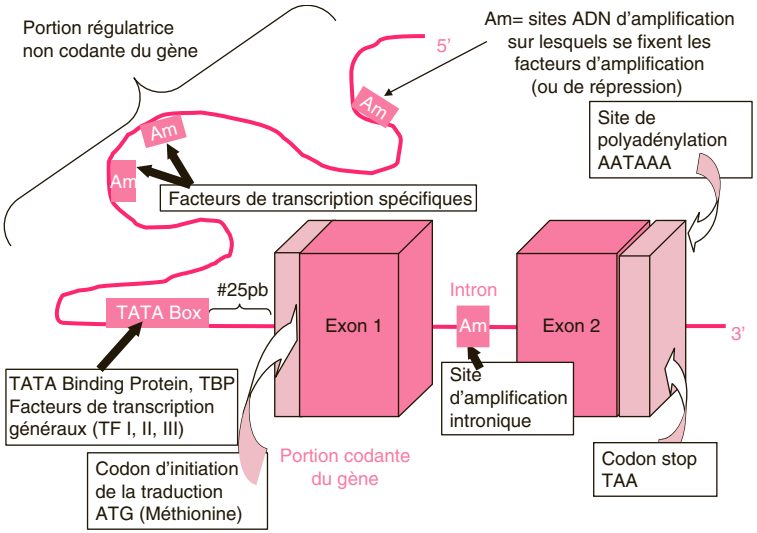


Figure 11 Structure d'un gène.

La portion codante du gène comprend le code génétique qui sera l'information ADN qui sera transformée en séquence d'acides aminés. Ce code est situé sur la portion ADN appelée exon. Chez les eucaryotes, la plupart des gènes sont faits de plusieurs exons, séparés par des introns. Le gène comprend par ailleurs une portion régulatrice non codante laquelle se subdivise en deux groupes de sites : (i) les sites de transcription généraux dont la TATA box (TATA est la séquence caractéristique de ce site) lequel existe dans la grande majorité des gènes. La TATA box est régulée par un grand nombre de protéines et c'est sur elle que se fixe l'ARN polymérase qui sera l'enzyme qui régule la transcription. (ii) Les sites dits amplificateurs (Am) qui vont amplifier la transcription et sont sensibles à des facteurs de transcription spécifiques comme par exemple celui qui est activé par l'AMP cyclique et est responsable des effets génomiques de l'adrénaline. La transcription commence au niveau du codon d'initiation et se termine au niveau du codon stop.

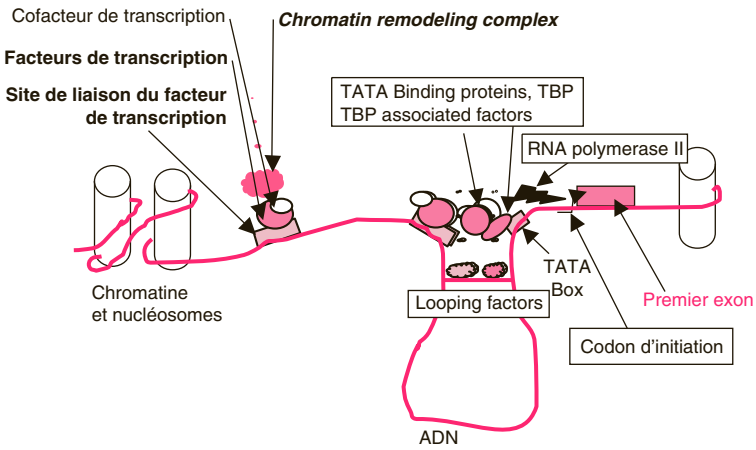


Figure 12 Promoteur d'un gène.

L'initiation de la transcription nécessite plusieurs douzaines de protéines qui inter-réagissent. La RNA polymérase II est un complexe de 15 protéines, la « *TATA binding protein* », les « *TBP associated factors* » sont au nombre de 8, il y a encore les facteurs de transcription et leurs très nombreux cofacteurs ainsi que plusieurs facteurs qui remodèlent la chromatine. La transcription demande également une réorganisation spatiale du complexe, par les « *looping factors* » entre autre (redessiné d'après Wray 2003).

ADN codant pour un jeu cohérent de produits fonctionnels potentiellement capables de se chevaucher » (Gerstein 2007). Cette définition a l'avantage de souligner deux points centraux : (1) un gène code pour des produits fonctionnels qui peuvent être des protéines et les ARNs ribosomaux et transferts, comme on le sait depuis longtemps, mais aussi des ARNs, comme les microARN régulateurs; (2) il existe des séquences ADN, les *transcriptionally active regions*, TAR, qui ne sont pas des gènes connus, et qui sont transcrits en ARN; (3) les sites d'initiation de la transcription sont plus nombreux qu'on ne le croyait, la trans-

cription de certains gènes pouvant dépendre de sites situés très loin de leurs séquences codantes. La nomenclature simplifiée des gènes (*TRAFD1; PTPN1; CD69*; *CLEC…*) se trouve sur « genecards.org ». Par convention les gènes sont toujours en italique.

b) Portion codante

La portion codante d'un gène d'eucaryote est complexe et polymorphe. Chez les eucaryotes, elle comprend une succession d'exons et d'introns. Les exons sont les seules portions du gène qui sont représentées dans l'ARN transcrit mature et qui sont porteuses du code génétique (Tab. 2). Ils peuvent correspondre à des séquences traduites ou non traduites en protéines. Les introns sont situés entre les exons, ils sont d'abord transcrits, puis au cours de la maturation des ARN ils sont excisés : ne subsistent que les séquences correspondant aux exons et qui forment l'ARN mature. À titre d'exemple, le gène du récepteur des lipoprotéines comprend 18 exons, celui de la chaîne lourde de la myosine en contient plus de 40. Les procaryotes, comme les bactéries, ont une portion codante unique sans introns. Chez les eucaryotes, certains gènes, comme ceux des récepteurs de l'adrénaline, ou les gènes mitochondriaux, sont dépourvus d'introns, ce qui suggère qu'il s'agit de vestiges ancestraux de gènes bactériens ou viraux incorporés dans le génome des eucaryotes au cours de l'évolution. Les exons du gène sont disposés dans l'ordre où vont se trouver les acides aminés de la protéine. Les divers domaines fonctionnels de la protéine correspondent rarement, et quand c'est le cas seulement par hasard, à un seul exon (comme le site de liaison de la troponine I à l'actine). Ces domaines sont généralement à cheval sur un ou plusieurs exons. C'est le cas des récepteurs des lipoprotéines, des sites ATPasique et de fixation de l'actine pour la chaîne lourde de la myosine et du site de liaison de la troponine I à la troponine C. Cette non-coïncidence entre cet aspect particulier de la structure et la fonction tire son origine de l'évolution même des protéines. Une protéine active en 2007 est le résultat de millions d'années d'évolution, période au cours de laquelle un certain nombre de mutations ponctuelles ou non, voire de duplications, se

Tableau 2 LE CODE GÉNÉTIQUE
(PAR ORDRE DE DÉGÉNÉRESCENCE CROISSANTE DES CODONS).

Sur ce tableau les bases sont données, comme c'est l'habitude, en ribonucléotides, c'est-à-dire que U (seul présent dans l'ARN) apparaît à la place de T. Il faudrait pour parler de code génétique *stricto sensu* utiliser T, seul présent dans l'ADN. Noter qu'il existe plusieurs triplets pour chacun des acides aminés, ces triplets qui codent pour le même acide aminé sont dits synonymes. Une mutation est dite synonyme lorsqu'elle porte sur un tel triplet (par exemple pour la cystéine une mutation UGC/UGU), elle ne change pas l'expression de la protéine, et, comme c'est la protéine qui est responsable de la fonction physiologique, c'est elle qui, au cours de l'évolution, est soumise à la pression sélective, une mutation à ce niveau n'aura pas de signification évolutive. À l'inverse une mutation qui se produit sur un codon non synonyme (exemple UCU qui code pour une sérine versus UGU qui code pour une cystéine) va modifier l'expression de la protéine et sera donc susceptible d'être soumise à pression sélective. Cette donnée a également un intérêt médical en cancérogenèse (voir plus loin). Il y a des exceptions au code génétique universel, en particulier chez certains protozoaires et dans les mitochondries (UGA qui est le codon stop universel code pour le tryptophane dans les mitochondries).

Acide Aminé	Triplets correspondant
Methionine	AUG [Codon initiateur]
Tryptophane	UGG
Phénylalanine	UUU UUC
Histidine	CAU CAC
Glutamine	CAA CAG
Asparagine	AAU AAC
Lysine	AAA AAG
Acide aspartique	GAU GAC
Acide glutamique	GAA GAG
Cystéine	UGU UGC
Tyrosine	UAU UAC
Isoleucine	AUU AUC AUA
Valine	GUU GUA GUC GUG
Proline	CCU CCA CCC CCG
Thréonine	ACU ACA ACC ACG
Alanine	GCU GCA GCC GCG
Glycine	GGU GGA GGC GGG
Sérine	UCU UCA UCC UCG AGU AGC
Leucine	CUU CUA CUC CUG UUA UUG
Arginine	CGU CGA CGC CGG AGA AGG
[Codons "Stops"]	UAA UAG UGA

sont produites; se sont transmises celles qui étaient efficaces sur un plan fonctionnel, même si elles résultent de ce que l'on a pu appeler un véritable bricolage dû au hasard.

Les processus de recombinaison génétique opérant au cours des méioses pendant les millions d'années de l'évolution ont conduit à multiplier les gènes avec de légères différences sans toutefois altérer la fonction codante du gène. C'est ce qui fait par exemple que le gène codant pour tel enzyme a exactement la même fonction enzymatique chez l'homme et le rat, mais pas tout à fait la même structure (les anticorps correspondants ne reconnaissent pas les protéines de l'autre espèce). On retrouve également des séquences encore très semblables à celle du gène ancestral, mais ne codant pas pour une protéine, les pseudo-gènes. On a par exemple montré qu'une sonde codant pour une des isoformes de l'actine de souris, pouvait s'hybrider complètement avec cinq séquences ayant moins de 5 % de divergences dans leur composition avec la séquence de la sonde, avec 15 à 20 séquences ayant 5 à 20 % de divergence, c'est-à-dire encore très proches de la sonde originelle, et avec prêt de 20 à 50 séquences ayant plus de 20 % de divergences avec la sonde, ces dernières séquences correspondant à des pseudo-gènes. Ces familles multigéniques sont particulièrement fréquentes lorsqu'il s'agit de gènes codant pour des protéines très conservées au cours de l'évolution et jouant des rôles essentiels en l'absence duquel il n'y a pratiquement pas de vie possible. C'est le cas de l'actine qui est une protéine contractile non, seulement responsable du mouvement musculaire, mais aussi essentielle pour tout ce qui concerne la motilité aussi bien dans des êtres monocellulaires primitifs, que pour le mouvement membranaire des cellules non contractiles.

c) Portion régulatrice

La portion régulatrice d'un gène comprend : (1) le promoteur généralement situé juste en amont de l'extrémité 5' du premier exon (il est activé par des facteurs de transcription dits généraux parce que retrouvés dans pratiquement tous les êtres vivants), (2) des séquences régulatrices du niveau de transcription du gène considéré, ces séquences et les

facteurs de transcription qui en contrôlent l'activité sont dites spécifiques parce qu'elles ne se retrouvent que sur certains gènes dont elles spécifient la transcription (on les appelle amplificateurs ou « *enhancers* »[1]) (fig. 11 & 12) et (3) des séquences situées à l'extrémité 3' contenant des signaux régulateurs de la terminaison de la transcription, ces dernières séquences peuvent également comprendre des amplificateurs.

Le promoteur est la portion du gène à laquelle se lie l'ARN polymérase qui est l'enzyme qui va catalyser la transcription en roulant en quelque sorte sur la portion non codante du gène (fig. 9). La polymérase se fixe d'abord de façon assez complexe, par l'intermédiaire de protéines spécifiques sur des segments consensus du promoteur appelés « TATA box » et « CAAT ». L'amplificateur (en plus ou en moins, activateur ou inhibiteur) est très variable d'un gène à l'autre, c'est la portion régulatrice où l'on trouve diverses séquences consensus d'ADN spécifiques de protéines régulant la transcription et activées, directement ou non, par les hormones (et peut-être les contraintes mécaniques).

d) Types de gènes

Beaucoup de gènes codent pour plusieurs protéines. La majorité des gènes sont uniques, non répétitifs. Certains peuvent coder pour plusieurs protéines, en général des isoprotéines, c'est-à-dire des protéines qui ont des fonctions analogues (comme les isoenzymes) mais des structures légèrement différentes. Ces gènes possèdent en général un ou plusieurs exons qui codent pour une portion de la protéine qui est commune aux diverses isoprotéines, et plusieurs autres exons spécifiques de chacune des isoprotéines.

Il existe par ailleurs des familles de gènes, c'est-à-dire des gènes issus d'un gène ancestral commun, codant pour des protéines possédant différents degrés d'homologie, ces gènes peuvent être dispersés sur un ou

1. Les effets d'un amplificateur peuvent être limités par la mise boucle d'un fragment d'ADN au moyen de « *looping factors* » qui rapprochent des séquences ADN éloignées (illustré fig. 12).

plusieurs chromosomes. C'est par exemple le cas pour deux protéines responsables de la contraction musculaire : l'actine et la myosine. Il existe environ 64 gènes ou pseudo-gènes pour l'actine, ces gènes ne codent que pour environ une dizaine d'isoformes différentes. La myosine est le principal responsable de la contraction musculaire. C'est un polymère possédant deux paires de sous-unités légères et une paire de sous-unité lourdes. Six gènes au minimum codent pour la sous-unité lourde, ils permettent à la myosine d'exister sous forme de différents isoenzymes qui rendent compte des particularités contractiles des différents types de muscles.

Par superfamille de gènes on entend des familles moins étroitement unies, ce qui veut dire que les analogies de structure sont moins étroites. Un exemple important est constitué par les superfamilles de récepteurs membranaires à 7 motifs hydrophobes (par lesquels le récepteur peut s'ancrer dans la membrane) ayant une affinité pour les G protéines (récepteurs R7G), ces récepteurs ont la particularité de ne pas avoir d'introns et d'avoir tous des fonctions très comparables. La superfamille des récepteurs nucléaires ayant une affinité pour l'ADN est un autre exemple, elle inclut les récepteurs des hormones sexuelles, de l'aldostérone et du cortisol, ces récepteurs sont en fait des facteurs de transcription puisqu'ils peuvent se lier directement avec l'ADN dans le noyau.

Il faut souligner le fait que les gènes responsables en commun d'une fonction particulière ne se trouvent pas nécessairement, et de loin, sur le même chromosome. Chez la souris par exemple les gènes codant pour les protéines de la contraction musculaire se trouvent dispersés entre, au moins, sept chromosomes différents : chromosome 1 pour deux des chaînes légères de la myosine de muscle squelettique rapide, 2 pour l'actine de muscle squelettique, 7 pour une autre des chaînes légères de la myosine de muscle squelettique rapide, 9 pour la chaîne légère de myosine cardiaque, 11 pour la chaîne légère spécifique des oreillettes et pour les chaînes lourdes des myosines squelettiques rapide, embryonnaires, néonatales et extra-oculaires, 14 pour les deux isoformes des chaînes lourdes de la myosine cardiaque (ces deux formes sont en tandem à 4 kb l'une de l'autre), 17 pour l'isoactine cardiaque.

3.2.7 Génome mitochondrial

L'ADN n'existe pas que dans les chromosomes, il existe également une petite proportion de l'ADN cellulaire qui est extra-nucléaire et qui se trouve dans les mitochondries. Cet ADN est très particulier :

- il est très peu abondant et donc difficile à isoler au sein de la masse d'ADN cellulaire qui est essentiellement nucléaire;
- il est très certainement d'origine bactérienne;
- il est défini par un seul type d'ADN circulaire double-brin (un brin lourd riche en guanine et un brin léger riche en cytosine) dont la séquence complète est maintenant établie;
- sa longueur, chez l'homme, est de 16 569 paires de base, il contient 37 gènes, 13 codent pour des enzymes mitochondriaux responsables de l'oxydation phosphorylante, les autres codent pour des ARN ribosomaux ou de transfert en charge de la synthèse mitochondriale ce qui veut dire que la majorité des protéines des mitochondries ont une origine nucléaire, et sont importées dans les mitochondries;
- l'ADN des mitochondries est très compact et ne contient que 7 % d'ADN non codant, par ailleurs les gènes ne contiennent pas d'introns;
- il est transmis par la mère (voir plus loin figure 22), c'est une des particularités des maladies génétique mitochondriales, il est le siège de nombreuses mutations elles-mêmes à l'origine de maladies génétiques mitochondriales.

3.3 LES GRANDS MÉCANISMES

La synthèse protéique peut être régulée à différents niveaux : au niveau de la transcription du gène en précurseur de l'ARNm, lors de la maturation du transcrit primitif dans le noyau, ou encore au niveau de la traduction du code génétique dans le cytoplasme (fig. 13). On peut ainsi lorsque l'on étudie un mécanisme de régulation de la synthèse protéique mesurer, en même temps que la protéine nouvellement formée, l'ARNm correspondant, et si la concentration de ce dernier augmente

avant celle de la protéine, on peut prédire que la régulation sera prétraductionnelle; s'il n'y a aucun lien chronologique entre les deux, la régulation est traductionnelle.

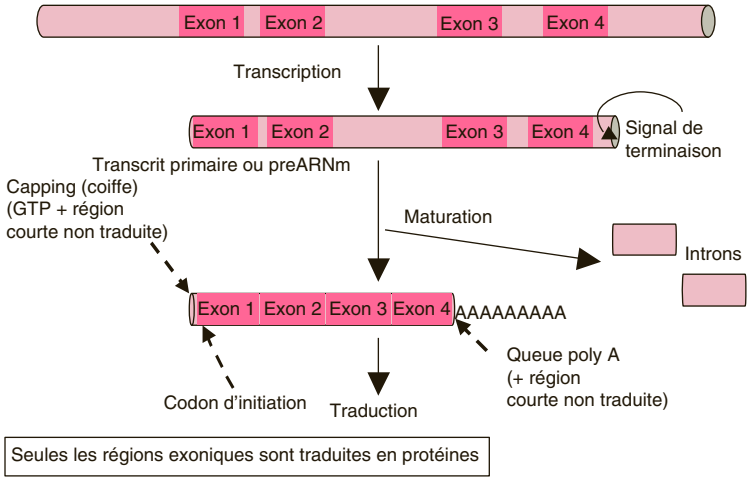


Figure 13 Transcription, maturation, traduction chez les eucaryotes. La transcription aboutit à la formation d'un transcrit primaire lequel inclut les introns. La maturation (1) élimine les séquences ARN introniques, (2) fabrique une coiffe Guanosine Tri Phoshate, GTP = quelques séquences non traduites à l'extrémité 5', (3) synthétise une queue poly AA à l'extrémité 3'.

3.3.1 Transcription

Transcription qualifie le processus par lequel l'information contenue dans l'ADN est convertie en équivalent ARN. Une partie de ces ARN, les ARN messager, ARNm, servira à la synthèse des protéines. La régulation de la transcription est plus complexe chez les eucaryotes que chez les procaryotes, ceci est dû en particulier à la compartimentalisation qui existe chez les eucaryotes, l'élément régulateur majeur est constitué par des protéines, les facteurs de transcription lesquels sont syn-

thétisés dans le cytoplasme, ils doivent pénétrer dans le noyau pour agir. Par ailleurs l'ADN des eucaryotes est beaucoup plus condensé en hétérochromatine que ne l'est celui des procaryotes. Cette condensation gêne l'accès des facteurs de transcription et de l'ARN polymérase et empêche la transcription. Sa levée participe donc indirectement à la régulation de la transcription (voir précédemment Chromatine).

a) Régulation au niveau de l'ARN polymérase

La transcription est régulée par les ARN polymérases (Tab. 3). Deux sont responsables de la synthèse des ARN transferts et ribosomaux. L'ARN polymérase II s'occupe de la synthèse des ARNm[1]. Au niveau transcriptionnel proprement dit, la régulation est le fait des facteurs de transcription qui sont des protéines capables de se lier à la fois à des séquences ADN consensus et à l'ARN polymérase par l'intermédiaire d'un complexe (fait d'environ 20 sous-unités protéiques), appelés médiateur.

Tableau 3 LES ARN POLYMÉRASES

ARN polymérase I : transcrit les gènes codant pour les deux molécules d'ARN ribosomial
ARN polymérase III : transcrit les gènes de l'ARN transfert, de l'ARN 5S, et quelques autres petites molécules d'ARN
ARN polymérase II : transcrit la plupart des gènes des eucaryotes qui codent pour des protéines, leur régulation est la plus complexe

La transcription aboutit d'abord à la synthèse d'ARN dits pré-messager ou ARN nucléaire ou transcrit primaire qui sont la copie conforme des introns et des exons du gène. Ensuite l'ARN pré-messager va subir une maturation, c'est-à-dire qu'il va à la fois perdre ses introns qui seront excisés et libérés en formant des sortes de lasso, alors que les exons restant seront religaturés entre eux, c'est l'épissage. Ce processus est, on l'a

1. Il y a avant la traduction plusieurs niveaux de régulation possible, tous situés dans le noyau. On peut, pour les disséquer, travailler sur des noyaux purifiés (techniques de « *run-on* » ou de « *run-off* »).

déjà vu, un des moyens utilisés par la cellule pour former des isoformes à partir d'un gène unique.

Les transcrits vont en même temps subir deux autres processus de maturation : (1) en 5', une séquence GTP méthylée va se fixer en 5', c'est le « *capping* » (ou coiffage) qui pourrait, entre autre protéger l'ARN contre l'attaque par des nucléases; (2) en 3', la transcription va s'arrêter lorsque la polymérase va rencontrer un site de terminaison, l'un des codons-stop; mais une fois synthétisés les ARN seront clivés une vingtaine de bases en aval d'une séquence de reconnaissance, la séquence AAUAAA, après cette coupure une polyA polymérase va catalyser la greffe en 3' d'une séquence polyadénylée AAAAAA de longueur variable. Cette séquence semble jouer un rôle dans la stabilité de l'ARNm, peut-être en se fixant à une protéine « protectrice ». La RNA polymérase sait donc quand elle doit s'arrêter.

La figure 14 fournit quelques exemples de ces types de régulation. Les hormones hydrosolubles, comme les catécholamines, ne peuvent traverser la couche lipidique membranaire et disposent d'un système en deux temps : dans un premier temps elles se fixent sur un récepteur situé sur la membrane externe qui par l'intermédiaire d'un système de transduction, contrôlé par des G protéines, va transmettre le signal et activer la libération d'un second messager, l'adénosine mono-phosphate cyclique, ou AMPc, qui va à son tour phosphoryler un facteur de transcription spécifique. Ce facteur agit en trans sur une séquence nucléotidique, qui va à son tour agir en cis sur la transcription. Les hormones liposolubles traversent la membrane externe et inter-réagissent directement avec un récepteur nucléaire (qui, on l'a vu plus haut, appartient à une superfamille de gènes) qui est aussi un facteur de transcription, certains sont d'ailleurs des oncoprotéines. Le récepteur activé agira directement en trans. Certains cas particuliers existent, comme l'insuline qui active un facteur de transcription par l'intermédiaire d'une kinase.

Certains gènes peuvent être à l'origine de plusieurs protéines par épissage (« *splicing* ») alternatif, ce processus est multiforme. Il est utilisé chez les eucaryotes et intervient au moment de la maturation de l'ARNm soit par sélection du promoteur (fig. 15), soit par sélection de la queue, soit par sélection alternative d'une cassette.

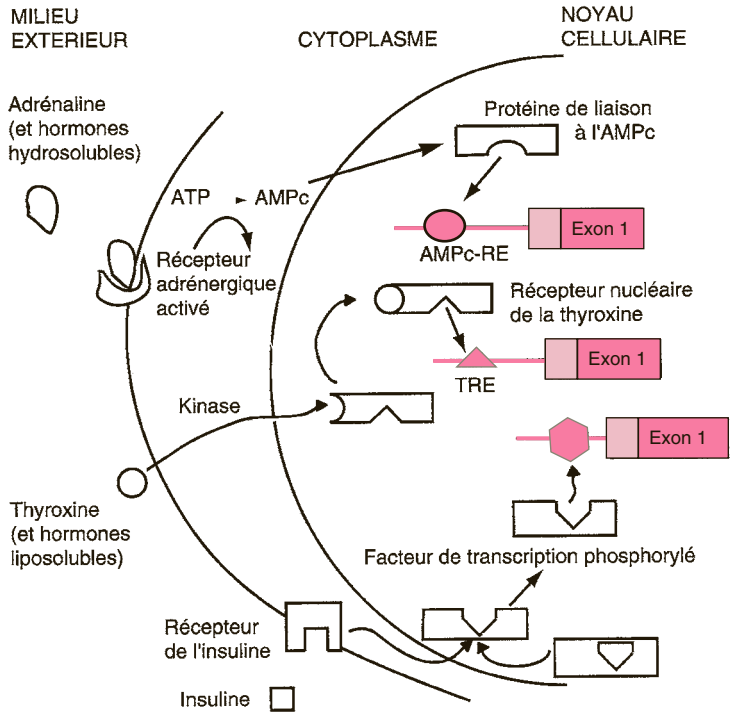


Figure 14 Effets d'un certain nombre d'hormones sur la transcription.

(1) Les hormones hydrosolubles ne peuvent diffuser à travers la membrane et ont besoin d'un récepteur qui transmette le signal et active la synthèse d'un second messager, ici l'AMPcyclique. Ce dernier va à son tour phosphoryler une protéine, un facteur de transcription, qui ira se fixer en *trans* sur une séquence ADN spécifique consensus et activer ainsi en *cis* la transcription.

(2) Les hormones liposolubles, comme la thyroxine, diffusent à travers la membrane et se fixent sur des récepteurs nucléaires capables de se fixer eux-mêmes sur des séquences consensus d'ADN.

(3) Enfin, certaines hormones agissent de façon plus complexe, en catalysant, comme l'insuline, l'autophosphorylation de leurs récepteurs, puis la phosphorylation d'un facteur de transcription.

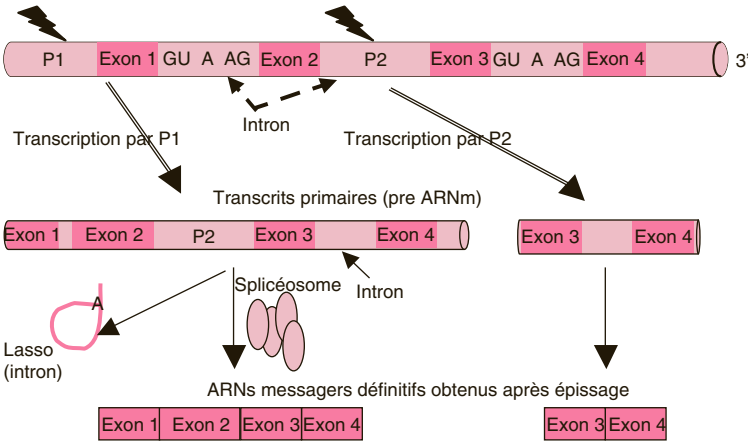


Figure 15 Épissage alternatif.

Certains gènes eucaryotes peuvent transcrire plusieurs protéines. Dans cet exemple, un même gène possède plusieurs promoteurs, P1 et P2. L'activation de P1 aboutit à la transcription primaire de tout le gène, P2 compris. L'activation de P2 ne permet que la transcription des séquences situées en 3' par rapport à P2. Les résidus GU et AG situés, en haut, sur les introns indiquent les points d'épissage. Le résidu A situé au milieu des introns indique le point de fermeture du lasso formé par les introns après leur excision. Le splicéosome est un ensemble de ribonucléoprotéines nucléaires qui forme l'unité fonctionnelle d'épissage.

Récemment on a découvert l'existence d'un processus analogue à l'excision des introns, mais qui porte sur les protéines. Chez certaines bactéries, peut-être chez d'autres espèces (chez les eucaryotes il existe des séquences ADN codant pour les intéines, mais elles sont probablement excisées), il existe des protéines appelées intéines qui sont transcrites en ARNm et qui seront excisées lors de la traduction, les régions qui resteront et formeront la protéine finale seront composées d'extéines.

Il existe des exemples, rares, au cours desquels la séquence de l'ARNm lui-même peut être modifié par ARN « *editing* ». Cette réaction inter-

vient dans la synthèse d'une apolipoprotéine, l'apo B48, à partir de la séquence de l'apolipoprotéine B100 et qui fait intervenir une désaminase.

Chez les procaryotes, la transcription est plus simple et fait intervenir la formation d'un bulbe de transcription (fig. 16), par ailleurs une seule molécule d'ARNm peut contenir les informations provenant soit d'une seul soit de plusieurs gènes proches les uns des autres.

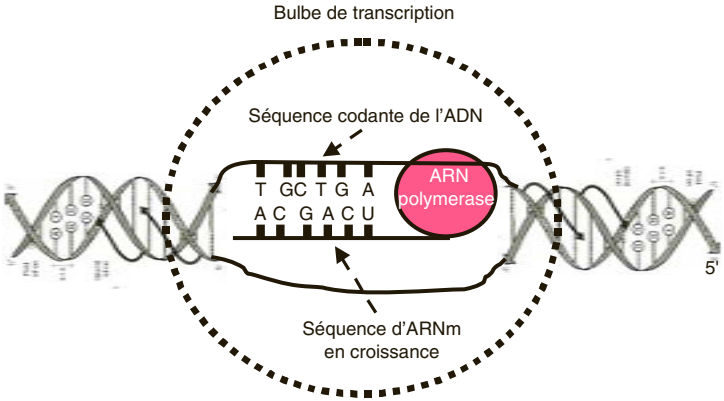


Figure 16 Initiation de la transcription chez les procaryotes.
La réaction est ici plus simple et caractérisée par la formation d'un bulbe de translation.

b) Facteurs de transcription

Les facteurs généraux non spécifiques, communs à la plupart de gènes (TFI, II et III) agissent en cis, les facteurs plus spécifiques, sensibles aux hormones par exemple, agissent en trans (fig. 12). Les facteurs de transcription partagent un certain nombre de propriétés : ils répondent à un stimulus qui indique la nécessité d'activer ou d'inactiver l'expression d'un gène donné, ils sont capables de pénétrer dans le noyau, ils ont deux domaines indépendants, l'un qui reconnaît une séquence ADN spécifique, l'autre qui se lie à l'appareil de la transcription, qui comprend, entre autre, l'ARN polymérase. Les sites ADN promoteurs peuvent être situés

à des emplacements très variables par rapport à la séquence codante : le locus *Shh* par exemple est à 800 kb du site de transcription (son activité dépend de facteurs qui forment et déforment les boucles d'ADN et rapprochent ainsi le promoteur de la séquence codante), certains promoteurs peuvent être introniques (*CCR5* chez l'homme), voir même exonique (*keratin 18* chez l'homme).

La connaissance de la séquence du génome fournit, on l'oublie souvent, deux types d'information concernant les gènes, la séquence codante, bien entendu, mais aussi les séquences promotrices en cis ou en trans, c'est-à-dire les séquences qui régulent l'expression des séquences codantes. La transcription d'un gène peut être modifiée par des mutations au niveau de la séquence codante, mais elle peut aussi être altérée par des mutations au niveau de ces promoteurs (lesquels sont très polymorphes, il existe, par exemple, chez l'homme de nombreux allèles des promoteurs des cytokines). Ces facteurs sont très sensibles à la pression sélective au cours de l'évolution.

Les familles de facteurs de transcription sont relativement peu nombreuses et ces facteurs eux-mêmes sont excessivement pleiotropes, un seul facteur de transcription peut en effet modifier l'expression de plusieurs protéines (certaines pouvant elles-mêmes être des facteurs de transcription) à la fois. L'abondance d'une protéine est héréditaire et transmissible, elle varie d'un tissu à l'autre, mais aussi d'une espèce vivante à l'autre, cette abondance est contrôlée par les séquences cis et les facteurs de transcription qui les activent (Wray 2003).

c) Régulation au niveau de la chromatine

La transcription ne peut pas avoir lieu lorsque la chromatine est sous forme d'hétérochromatine. Cette structure compacte est maintenue par les histones des nucléosomes. Les histones sont des protéines basiques dimériques qui sont presque inchangées au cours de l'évolution. Ils possèdent une queue, celle-ci peut être acétylée par une transférase spécifique et désacétylée par une désacétylase. Quand les histones sont désacétylés, les nucléosomes sont compactés en hétérochromatine, l'acétylation désagrège ces structures et les rend perméables aux facteurs de transcription. C'est donc un préalable nécessaire à la transcrip-

tion chez les eucaryotes, pas chez les bactéries. L'un des schémas proposés inclut donc comme étape première dans la transcription une liaison entre ces transférases et les facteurs de transcription.

L'inactivation d'un gène, d'un groupe de gènes ou d'un chromosome ou « *silencing* » est due à la méthylation de cytosines sur des séquences CG. Le rôle physiologique de ce processus n'est connu qu'au cours du développement ou dans le cas des deux chromosomes X, chez la femme.

3.3.2 Traduction

Traduction (« *translation* ») qualifie le processus qui aboutit à la synthèse d'une protéine à partir des informations contenues dans l'ARNm.

La synthèse protéique a lieu au niveau de formations particulaires intracellulaires, les ribosomes. Les ribosomes sont composés d'ARN et de protéines ribosomales (environ une cinquantaine) et comportent deux sous-unités (60S et 40S chez les eucaryotes, 50S et 30S chez les procaryotes). Ces dernières contiennent les enzymes nécessaires à la traduction (fig. 17).

Les acides aminés nécessaires à la synthèse d'une protéine ne peuvent inter-réagir avec les ribosomes directement, ils sont véhiculés jusqu'aux ribosomes grâce aux ARN transfert, lesquels possèdent un site d'attachement pour un acide aminé donné (il y a autant d'ARN de transfert qu'il y a d'acides aminés, un ARN transfert sera dit acylé lorsqu'il porte l'acide aminé qui lui est propre) et une région appelée anticodon parce qu'elle reconnaît un codon donné. Ces ARN sont de petite taille, des séquences de bases complémentaires internes peuvent former des régions à double brin séparées les une des autres par des régions à simple brin (fig. 17 & 18), l'ensemble formant, en structure plane, une feuille de trèfle. Aucun ARN transfert ne peut s'apparier avec un codon-stop. Il existe 64 codons (Tab. 2), mais seulement 30 ARN transfert dans le noyau (et 22 dans les mitochondries). La sélection de l'acide aminé approprié lors de l'assemblage sera donc conditionnée par l'appariement très spécifique qui va se produire entre les bases de l'anticodon, situées sur l'ARN de transfert, et celles du codon, situées

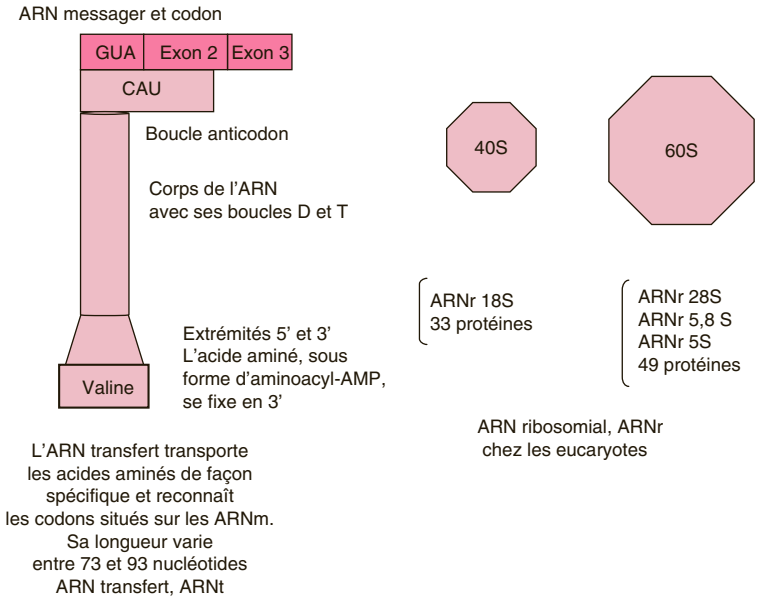


Figure 17 Structure de l'ARN transfert, ARNt, et de l'ARN ribosomal, ARNr.

La structure tertiaire de l'ARNt a une forme de L avec deux boucles D et T.
L'acide aminé se fixe en 3'. L'anticodon est au milieu de la séquence. Les ARNr des eucaryotes (figure) sont différents des ARNr des procaryotes lequel consiste en deux sous unités respectivement 50S (5S ARNr, 23S ARNr et 32 protéines) et 30S (16S ARNr et 21 protéines).

sur l'ARN messager. Cet appariement est catalysé par un enzyme spécifique, une aminoacyl-ARN transfert synthétase (il y a une synthétase par acide aminé, certains acides aminés, codés par plusieurs codons, possèdent plusieurs synthétases). Les erreurs d'appariement existent, elles sont en règle corrigées (comme les erreurs d'appariement de l'ADN) par un mécanisme d'édition. Il va y avoir au cours de cette synthèse

glissement, translation de la chaîne d'acides aminés en cours de formation sur l'ARN messager et les ribosomes – la traduction se dit souvent translation (c'est le terme anglais) pour cette raison. La traduction est un processus intra cytoplasmique chez les eucaryotes, chez les procaryotes, elle est fréquemment contemporaine de la transcription, ce qui fait que, chez les procaryotes, la chaîne d'acides aminés se trouve souvent en cours de formation alors que l'ARNm lui-même n'est pas encore complètement synthétisé.

L'initiation de la traduction se fait grâce aux facteurs d'initiation (eIF1 et eIF3), qui sont des protéines. Le signal de début de traduction est la séquence AUG, qui code pour la méthionine, l'ARN transfert

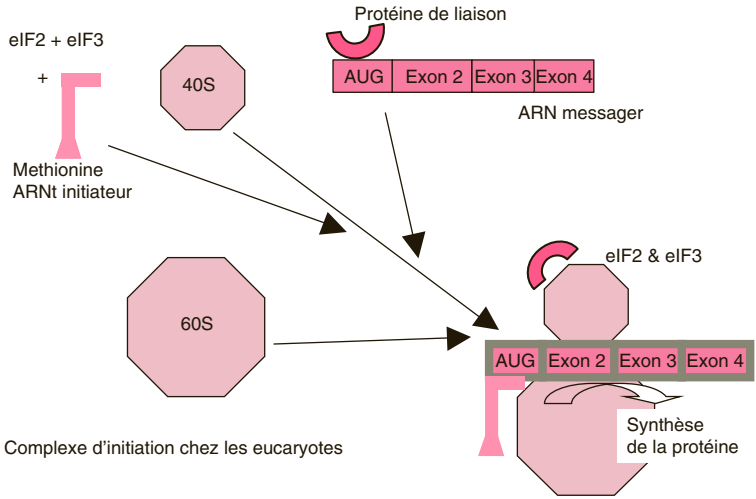


Figure 18 Initiation de la traduction chez les eucaryotes.
La traduction de l'ARNm en protéines débute par une série de réactions consistant à assembler deux facteurs d'initiation, eIF, les deux sous unités d'ARNr et le messager. La fixation de l'ARNt initiateur chargé de son acide aminé, la méthionine, va déclencher la traduction, c'est-à-dire l'accrochage des ARNt chargés de leurs acides aminés.

correspondant est dit ARNt initiateur, il est différent de l'ARN qui introduit, quand il le faut, une méthionine dans la chaîne protéique. Le ribosome est ancré à l'ARNm grâce à une séquence présente sur l'ARNm (AGGAGGU) appelée site de fixation du ribosome, laquelle est complémentaire d'une séquence portée par le ribosome. L'ensemble forme un complexe, visible au microscope électronique.

L'élongation se poursuit sous l'effet des facteurs d'élongation, la lecture de l'ARNm se faisant dans le sens 5'-3'. L'assemblage des acides aminés situés sur les ARNt aboutira à la synthèse de nouvelles protéines grâce à des réactions de condensation créant les liaisons peptidiques. Ces réactions sont catalysées par une peptidyl transférase située dans la grande sous-unité ribosomale et sont consommatrices d'énergie (sous forme de GTP).

La fin de la traduction intervient au moment où le glissement atteint les codons stop puisqu'il n'existe pas d'ARN transfert capable de s'associer aux codons stop. Elle nécessite l'intervention d'un facteur de dissociation (RF pour « *release factor* ») qui permet la libération de la chaîne. Un polyribosome, structure hélicoïdale visible au microscope électronique, représente une unité de traduction, il est fait d'un ARN messager couvert par une multitude de ribosomes (environ tous les 80 nucléotides). Un signal particulier dit d'adressage est destiné à diriger la protéine là où elle est nécessaire.

a) Régulation au niveau des ARN

L'ARN interférence, ARNi, est un mécanisme nouveau découvert par le prix Nobel, A. Fire (1998), et qui a ouvert de nouvelles perspectives sur les plans à la fois biologiques et technologiques, c'est un point chaud de la recherche aux multiples conséquences en thérapeutique (Gewirtz 2007). L'ARNi est présente dans la plupart des cellules et permet l'inhibition spécifique de l'expression des ARNm. Plusieurs mécanismes existent, comme par exemple celui qui passe par les siARN. De longs doubles brins d'ARN (dsRNA) sont transcrits, puis découpés par un Dicer en microARN (ou « *short interfering RNA* », siRNA) lesquels sont incorporés dans des « *RNA-induced silencing complex* », RISC, qui

vont contrôler spécifiquement la dégradation de l'ARNm. L'ARNi est très efficace et son efficacité est amplifiée par le fait que l'activité des siRNAs peut être amplifiée par l'effet d'une « *RNA-dependent polymérase* » qui fonctionne sans l'aide d'amorce.

Les « *microRNA* », miRNA, ont des propriétés analogues à celles des siRNA, mais ils bloquent la traduction par un mécanisme différent, en se liant aux ARNms, et non pas en les détruisant. Les miRNA proviennent de précurseurs qui sont des ARN très longs (70 nucléotides). Il y a des miRNA spécifiques dans le muscle, dans le myocarde etc., miR-1, miR133… Le rôle de l'ARNi dans la régulation du fonctionnement cellulaire est complexe et encore imparfaitement connu.

b) Après la traduction

Les produits de la traduction peuvent subir encore plusieurs types de modifications qui en altèrent profondément la nature. Certaines protéines, les protéoglycanes et les glycoprotéines, se lient aux chaînes latérales de différents glucides par glycolysation, en particulier celles qui sont secrétées ou exportées hors de la membrane. Les oligosaccharides sont préformés et ajoutés tels quels à l'extrémité de ces protéines. D'autres modifications peuvent survenir : phosphorylation par des kinases spécifiques, déphosphorylation par des phosphatases spécifiques, méthylation (d'une lysine) par des méthylases, hydroxylation (de la proline en hydroxyproline, acide aminé spécifique du collagène), carboxylation.

3.3.3 Réplication et réparation de l'ADN Cycle cellulaire

a) Réplication

Au cours du cycle cellulaire, pendant la phase S, avant la mitose, l'ADN se dédouble par un processus particulier et complexe qui met

en jeu un grand nombre d'éléments régulateurs[1]. Ce système dédouble l'information génétique, mais d'une part il n'est pas parfait, d'autre part il peut être altéré par des agressions chimiques ou physiques externes. Qui dit réplication sous-entend donc également mécanismes de réparation.

La réplication de l'ADN comprend d'abord la formation d'une fourche de réplication : une ADN hélicase va en effet dédoubler les deux brins de l'ADN, et former une fourche. Les brins seront maintenus séparé grâce à la fixation de protéines spéciales (« *single strand binding protein* », SSB)[2] (fig. 19). En même temps, il y aura synthèse de deux nouveaux brins complémentaires, synthèse qui se fera sur le modèle fourni par l'ADN parental, ce qui a deux conséquences : (i) lorsque le brin parental est 3'-5' la synthèse se fera de 5' en 3', et vis et versa; (ii) la synthèse est asymétrique, en effet la synthèse de l'un des deux nouveaux brins sera continue parce que se faisant dans le même sens que la fourchette de réplication, au contraire la synthèse du second nouveau brin ira en quelque sorte à contre-courant et sera discontinue, aboutissant à la formation de fragments de 1 000 à 2 000 paires de bases (dits fragments d'Okasaki). Une ligase devra dans un deuxième temps souder ces fragments entre eux. La réplication est donc asymétrique.

La réplication ne peut se faire qu'à partir d'amorces (ou « *primer* ») qui sont des petits ARN synthétisés par un complexe protéique appelé primosome et qui comprend entre autre une ARN polymérase (ou primase). La synthèse du brin principal ne demande qu'une seule amorce, celle des divers fragments qui composeront le second brin demande plusieurs amorces.

La réplication nécessite ensuite l'intervention de plusieurs ADN polymérases, l'une sert à la réplication (l'ADN polymérase III, Pol III), il y en a d'autres ou une autre dont le rôle est de réparer l'ADN les

1. Ce système est d'ailleurs à la base d'une technique importante d'analyse, la PCR (« *Polymerase Chain Reaction* »), voir plus loin 5.1.
2. Il y a onze protéines qui sont impliquées dans le processus de replication, certaines, en particulier la polymerase III existent sous forme de plusieurs isoformes.

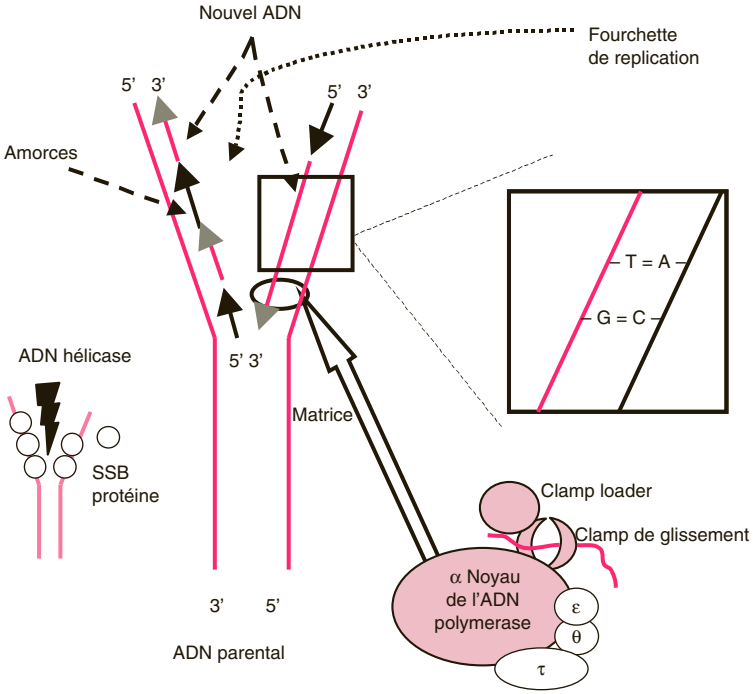


Figure 19 Réplication de l'ADN.

La réplication de l'ADN permet de diviser les deux brins de l'ADN en même temps et d'obtenir un nouvel ADN identique, ou presque, aux brins parentaux. La réplication comprend deux phases :
(1) les deux brins d'ADN parental doivent être séparés avec formation d'une fourchette de réplication;
(2) les deux nouveaux brins d'ADN sont ensuite synthétisés en se servant des brins originaux comme modèles. Finalement les deux nouveaux ADN seront composés d'un brin original et d'un brin nouvellement synthétisé. SSB protéine : « *single strand binding* protéine ».

erreurs commises lors de la réplication, en particulier pendant la soudure des fragments d'Okasaki, une ADN polymérase spécifique devra exciser les bases indûment incorporées et resynthétiser la base correcte en utilisant le brin complémentaire comme matrice. Cette réparation mettra en jeu successivement une ADN polymérase I (Pol I) et une ligase. La correction n'est néanmoins pas absolue et il subsiste toujours après réplication 10^{-10} à 10^{-11} d'erreurs.

L'ADN polymérase est un double complexe de plusieurs protéines (fig. 19) : (1) le clamp de glissement (« *sliding clamp* ») qui a une forme de beignet et qui est fait de deux sous-unités semi-circulaires qui entourent les deux nouveaux brins d'ADN comme des anneaux, il est associé à un complexe de charge (« *clamp loader* »); (2) le clamp de glissement est associé au noyau de l'ADN polymérase lui-même constitué de trois sous-unités (DnaE, la sous-unité alpha de polymerization; DnaQ, la sous-unité epsilon de lecture en charge de la réparation; HolE, la sous-unité theta. Chacun des deux complexes est en charge de l'un des nouveaux brins d'ADN, ces deux complexes sont réunis par une paire de sous-unités tau, ce qui suppose à la fois que le nouveau brin continu d'ADN et le nouveau brin discontinu (formé par les fragments d'Okasaki) sont synthétisés en même temps et que l'ADN s'est enroulé pour permettre cette synthèse simultanée.

La réparation de l'ADN est un mécanisme de sauvegarde absolument essentiel au cours de l'évolution. Les mutations de l'ADN sont des modifications de l'ADN transmissible, les mutations basales reflètent les erreurs spontanées de la ségrégation chromosomique lors de la mitose ou lors de la réplication et de la réparation qui lui fait suite. D'autres altérations sont dues à des réactions chimiques spontanées de dépurination, de désamination de la cytosine ou de dimérisation de pyrimidines. La détection de ces changements et leur réparation immédiate est la règle et met en jeu plusieurs mécanismes enzymatiques qui enlèvent la zone lésée et la répare à partir du modèle.

La réplication de l'ADN est la même chez les procaryotes et les eucaryotes, bien qu'elle soit mieux connue chez les premiers. Sur un

chromosome, la réplication se doit d'être synchronisée avec le cycle cellulaire ce qui est fait grâce à un site d'initiation (oriC) et à un complexe d'initiation fait de cinq protéines. L'initiation est contrôlée par au moins deux mécanismes la méthylation de l'ADN et l'attachement à la membrane. Elle se termine lorsque tout le chromosome a été répliqué, au niveau du site de terminaison (il y en a plusieurs). Les chromosomes des eucaryotes étant linéaires, leur réplication engendre des chromosomes filles plus courts, c'est à la télomérase qu'il appartiendra de synthétiser le télomère qui fera reprendre une taille normale au chromosome (voir plus haut « Télomère »). Par ailleurs, chez les eucaryotes, la réplication a plusieurs points de départ et il y a simultanément 10 000 à 100 000 renflements signalant la fourche de séparation sur une cellule somatique humaine en cours de division, ce qui pose un problème majeur de synchronisation, cette synchronisation est assurée par le « *replication licensing factor* ».

b) Cycle cellulaire

Les différentes étapes du cycle cellulaire sont régulées par des interactions complexes entre des signaux positifs et négatifs (fig. 20). Les régulateurs clés de la division cellulaire sont des kinases cycline-dépendantes. La progression normale du cycle de division cellulaire dépend de l'activation séquentielle de ces kinases et des phosphorylations de plusieurs substrats qui s'ensuivent. Les kinases forment des complexes quaternaires dont l'activité est contrôlée par l'activation des cyclines, la phosphorylation des sous-unités et l'association avec différents types d'inhibiteurs, dont les familles INK4 et CIP/KIP qui possèdent de multiples sites de liaison spécifiques. Ainsi, durant le cycle cellulaire, la progression d'une étape à l'autre est très finement contrôlée par de nombreuses protéines régulatrices. Le blocage de ce cycle est observé dans de nombreuses cellules adultes, comme les cellules cardiaques, il est corrélé avec la perte de l'activité de plusieurs cyclines et de plusieurs kinases, dont les cyclines D1 et D3, et avec l'activation de plusieurs inhibiteurs.

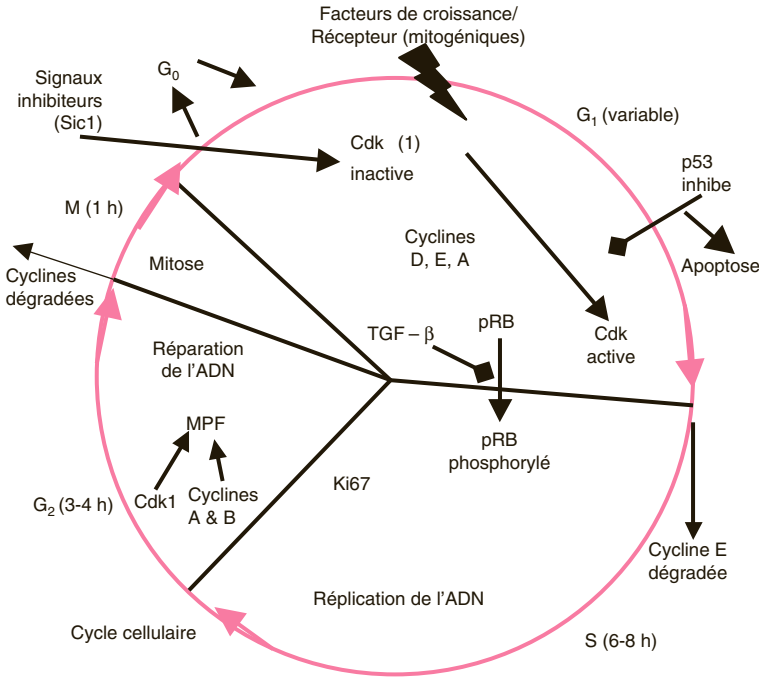


Figure 20 Cycle cellulaire.

G0, G1, S, G2 et M indiquent les différentes phases du cycle. Cdk : kinases dépendantes des cyclines. pRB : « *retinoblastoma protein* ». Cyclines A, B : cyclines mitotiques. MPF : facteur de promotion de la mitose. Ki 67 : marqueur de division cellulaire.

3.3.4 Mitose, méiose, fécondation

a) Mitose

Au moment de la division cellulaire, les chromosomes des cellules somatiques non germinales et la chromatine se dédoublent de sorte que chacune des deux cellules résultantes reçoive exactement le même capital génétique (fig. 21). Une cellule qui ne cesse de se diviser passe par

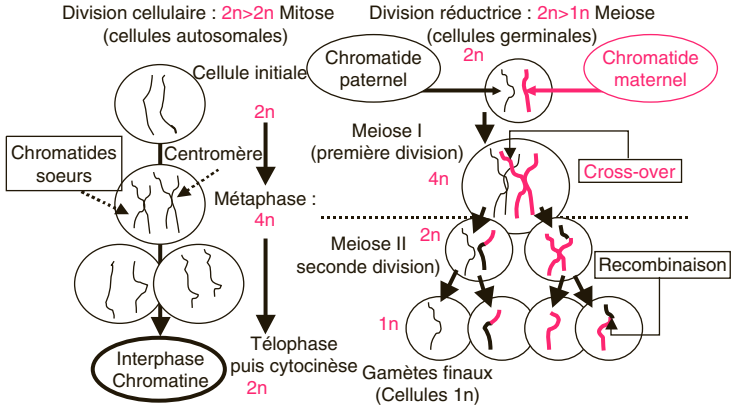


Figure 21 Mitose et méiose.

La mitose est le processus par lequel le matériel génétique des cellules somatiques est divisé en deux parties rigoureusement identiques, elle est généralement associée à la division cellulaire, bien que les deux phénomènes puissent être dissociés, il peut y avoir mitose sans division cellulaire (polyploïdie). Au cours de la mitose, la cellule est temporairement tétraploïde (4 copies du génome), en fin de mitose, au cours de l'interphase, l'individualité des chromosomes disparaît et laisse la place à une masse de chromatine. L'essentiel ici est que les cellules qui sont 2n au début restent 2n en fin de mitose. Par souci de simplification on n'a pas fait figurer les étapes intermédiaires, formation du fuseau horaire et de la plaque équatoriale. Les chromatides sont les deux éléments longitudinaux résultants de la réplication et réunis par un centromère. Ils sont visibles au début de la prophase et en métaphase. Après le clivage des centromères pendant l'anaphase, on ne parle plus de chromatides mais de chromosomes fils.

La méiose est le mode de division des cellules germinales, elle se distingue de la mitose sur deux points fondamentaux : (i) il y a non pas une division du génome, mais deux (méiose I et II dans la figure); (ii) alors que dans la mitose le matériel génétique terminal est rigoureusement identique au matériel initial, dans la méiose il y a recombinaison (souvent en des points particuliers dits points chauds), c'est-à-dire échange de matériel génétique entre chromosomes d'une même paire lors du « *cross-over* », les cellules haploïdes produites en fin de méiose sont génétiquement différentes les unes des autres. Le « *cross-over* » est à la base du polymorphisme de l'ADN. Le point d'attachement des fibres du fuseau mitotique au centromère est le cinétochore. Au moment de la métaphase le chromosome est constitué par deux chromatides et un centromère qui les relie (en bas et à droite).

diverses phases successives qui forment le cycle cellulaire. La phase préparatoire dite G1 (G pour Gap, c'est-à-dire intervalle) est la seule dont la durée est très variable. Vient ensuite la phase S, de Synthèse, au cours de laquelle se produit une division de l'ADN, appelée duplication, de 2n (2n signifiant que chacun des 22 chromosomes existe sous forme d'une paire) il passe à 4n. Au cours de cette phase il y a également synthèse de certaines protéines constituant la chromatine. Une deuxième phase préparatoire s'ensuit, appelée G2. Elle précède la mitose proprement dite. Au cours de la mitose (divisée elle-même en prophase, métaphase, anaphase, télophase), les chromosomes s'individualisent en utilisant le capital ADN obtenu précédemment par duplication, s'attachent sur des microfilaments tubulaires qui vont former des fuseaux, et se condensent sur la plaque métaphasique. Les deux chromosomes sont attirés vers les pôles opposés. Lors de la télophase, les chromosomes se condensent et les deux noyaux-fils (chacun avec 2n) du noyau initial vont se former. La cytokinése finale va donner deux cellules filles. Certaines cellules, comme les cellules cardiaques ou nerveuses, ont perdu cette capacité de se diviser et entre dans une phase de repos appelée Go.

b) Méiose

Le cycle cellulaire est différent pour les cellules germinales, c'est-à-dire pour les cellules qui produiront les ovules et les spermatozoïdes (fig. 21). Il y a en effet un double brassage de l'information génétique qui est un élément essentiel à la diversité génique. Les cellules germinales vont suivre le même cycle que les autres jusqu'au stade de la prophase qui est dans ce cas à la fois plus longue et plus complexe. La prophase comprend cinq stades : leptotène, zygotène, pachytène, diplotène, et diacinèse. Au cours de cette 1re prophase les chromosomes homologues vont s'apparier sur toute leur longueur en formant une sorte de double ruban appelé synapsis (au stade zygotène). Puis se produiront des coupures entre deux chromosomes homologues, suivis de religatures, qui aboutiront à l'échange, par recombinaisons, de fragments de chromosomes, donc d'informations génétiques, entre chromosomes homologues. C'est le « *crossing-over* ». Il y a en fait deux mitoses

successives, la première aboutit à la formation de deux cellules 2n possédant chacune le capital génétique initial, même si celui-ci a été brassé par le « *crossing-over* ». C'est une phase de brassage des gènes. Par contre, la seconde mitose - qui se produit juste après la télophase de la première division méiotique sans réplication de l'ADN, et sans phase S - est dite réductionnelle parce qu'elle aboutit à des cellules-filles 1n possédant chacune la moitié du capital génétique originel (la situation étant bien entendu compliquée par l'intervention du « *crossing-over* »). Ceci est obtenu au cours de l'anaphase II par synthèse de nouveaux chromosomes à partir d'un des chromosomes de chacune des paires originelles. Cette réduction définit la Méiose (Tab. 4).

c) Fécondation

La fécondation de l'ovocyte par un spermatozoïde aboutit à la formation d'un zygote (fig. 22) qui sera la première cellule d'un organisme.

Le zygote est 2n et donnera soit un homme si le chromosome sexuel est XY, soit une femme s'il est XX.

Au final le génome d'un organisme complet comprendra le génome nucléaire, celui des mitochondries, mais aussi le génome des bactéries dites commensales, c'est-à-dire essentiellement la flore intestinale laquelle joue un rôle physiologique métabolique qui complète le métabolisme contrôlé par le génome normal (fig. 23, tab. 4).

3.3.5 Régulations épigénétiques

D'une façon un peu restrictive, on entend par épigénétique, l'étude des modifications covalentes de l'ADN et des histones qui régulent l'activité des gènes dans le complexe chromatinien après la transcription sans qu'il y ait altération de la séquence. Cette discipline est en plein développement actuellement. Ces modifications sont réversibles et transmissibles sur au moins une génération. Les plus connues sont la méthylation des résidus cytosine dans les dinucléotides CpG sur la molécule ADN et les modifications plus diverses des histones. Ces modifications agissent sur la conformation spatiale de la chromatine.

Tableau 4 VUE D'ENSEMBLE DES MODES DE TRANSMISSION DES DIFFÉRENTES PARTIES DU GÉNOME HUMAIN

Femmes	Hommes
Cellules somatiques diploïdes (2n) Ces cellules se divisent par mitose en cellules filles identiques en termes d'ADN	
– 22 chromosomes, 2 paires par cellule autosomale – une paire de chromosomes X – ADN mitochondrial 1 000 copies par cellule	– 22 chromosomes, 2 paires par cellule autosomale – un chromosome X et un chromosome Y – ADN mitochondrial 1 000 copies par cellule
Mitoses et meioses qui dédoublent les chromosomes et entraînent des recombinaisons	
Ovogenèse	**Spermatogenèse**
Cellules germinales haploïdes (1n)	
Gamètes femelles = œufs – 22 chromosomes autosomaux – un X – ADN mitochondrial 100 000 copies par œuf (dans le cytoplasme)	Gamètes mâles = sperme – 22 chromosomes autosomaux – un X ou un Y – ADN mitochondrial 50-75 copies par spermatozoïde (dans la pièce intermédiaire à la base du flagelle, perdu lors de la fécondation)
Fertilisation	
Zygotes diploïdes (2n) Ce sont des cellules fertilisées résultant de l'union d'un gamète mâle et d'un gamète femelle	
– 22 chromosomes, 2 paires par cellule autosomale, une copie de chacun des parents – une paire de chromosome X – ADN mitochondrial 1 000 copies par cellule	– 22 paires de chromosomes autosomaux – un X de la mère, un Y du père – ADN mitochondrial de la mère

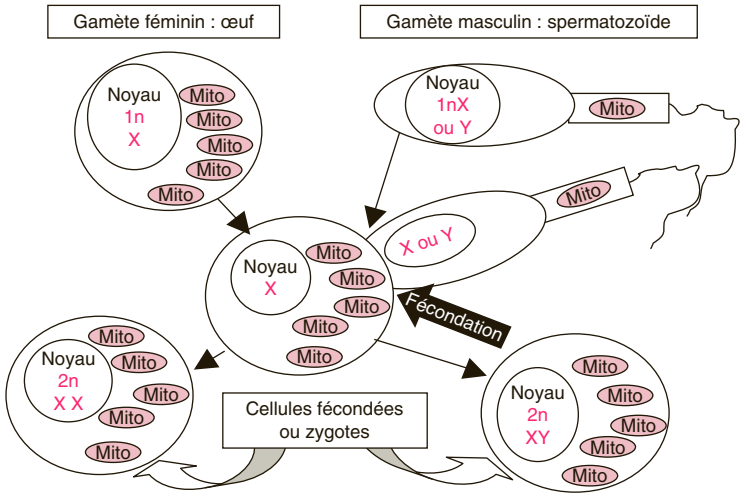


Figure 22 Fécondation et zygotes.

Les gamètes sont obtenus en fin de méiose et sont des cellules 1n. La fécondation d'un ovocyte par un spermatozoïde aboutit à la formation d'un zygote ou cellule fécondée qui sera 2n avec une copie chromosomique provenant de chacun des deux parents. Néanmoins cette copie est généralement modifiée, de façon aléatoire, par les recombinaisons survenues au cours du « cross-over » (voir la fig. 21). Le zygote est la première, chronologiquement parlant, des cellules de l'organisme, son génome sera recopié dans toutes les cellules de l'organisme. Le sexe de l'embryon qui va se développer ensuite sera déterminé par la présence du chromosome sexuel, XX pour les femmes, XY pour les hommes, Y ne pouvant être transmis que par le père.

L'ADN mitochondrial est également transmis lors de la fécondation. Néanmoins le spermatozoïde contient beaucoup moins de mitochondries, présentes dans la queue et donc éliminés lors de la fécondation, que les ovocytes, ce qui fait que, en pratique, chaque individu, mâle ou femelle, hérite des mitochondries de sa mère. En génétique des populations l'ADN mitochondrial est pour ces raisons un marqueur de l'hérédité féminine. L'ADN du chromosome Y est à l'inverse un marqueur de l'hérédité masculine.

L'ADN de la flore intestinale physiologique enfin est en majorité d'origine maternelle

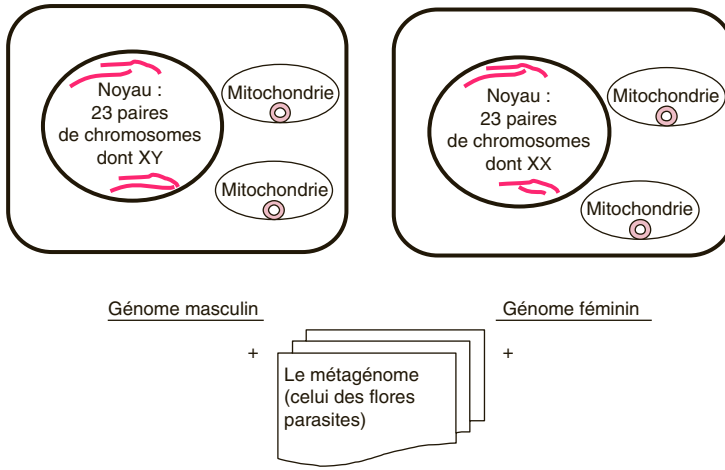


Figure 23 Génome cellulaire masculin chez l'homme et la femme.

Le génome haploide du noyau contient 3 200 Mpb, il y a deux copies de ce génome dans les cellules diploïdes. Ce génome est organisé en 46 molécules linéaires par cellule diploide, les chromosomes. Le génome mitochondrial représente 16 569 pb, avec 2 à 10 copies par mitochondrie et 1 000 copies par cellules, le génome mitochondrial est circulaire. Le génome d'un être humain comporte également celui des bactéries dites commensales, le microbiome qui comprend les flores intestinales, orales et cutanées, ce génome bactérien commensal est appelé métagénome. Métaboliquement parlant, le métagénome complète le génome humain (Gill 2006).

Tous les gènes ne peuvent être soumis à empreinte et ne sont pas méthylables. Le nombre de gènes dont l'expression peut être contrôlée par un mécanisme épigénétique de ce type est limité à 60-80 et comprend, entre autre, des gènes liés à l'insuline et *p53* qui est une cible en cancérogenèse. L'embryon ne peut se développer normalement que s'il possède et reconnaît deux génomes, l'un maternel, l'autre paternel. La reconnaissance de ces deux génomes implique qu'ils soient identifiables bien que

leurs séquences soient semblables, c'est l'empreinte parentale. L'impression se produit au cours de la gamétogenèse, c'est une méthylation des promoteurs de certains gènes qui peut persister toute la vie. Certains amas de gènes seront éteints, et le produit en sera l'expression monoallélique paternelle ou maternelle de gènes qui devraient avoir une expression biallélique.

Il y a beaucoup d'arguments en faveur d'une régulation épigénétique en physiopathologie (comme dans le syndrome métabolique). Chez le rat par exemple, un régime riche en donneurs de méthyl (folates, méthionine) pendant la gestation augmente significativement le nombre de porteurs de séquences méthylées et la survenue d'hypertension artérielle à l'âge adulte. La surnutrition glucidique de l'enfance est cause d'hyperinsulinisme, cette anomalie est transmise à la seconde génération chez l'homme et chez l'animal. L'épigénétique enfin joue vraisemblablement un rôle dans la cancérogenèse (leucémie et tératocarcinome).

3.3.6 Reproduction sexuée

Chez les mammifères les gonades embryonnaires sont potentiellement bisexuées et contiennent les cellules germinales. Dans l'espèce humaine, 7 semaines après la fécondation, elles possèdent deux paires d'ébauches des conduits génitaux, une paire femelle et une paire mâle. Le sexe définitif sera déterminé par le sexe génétique, c'est-à-dire par le couple de chromosomes sexuels, XX pour les femmes, XY pour les hommes. Les chromosomes X et Y ont un très grand degré d'homologie due à leur origine ancestrale commune. Le chromosome Y est plus petit que le chromosome X, il est porteur d'un gène SRY, pour *sex determining region Y*, qui est un gène de régulation codant pour un facteur de transcription responsable de la différentiation testiculaire. En son absence la gonade fœtale se différencie en ovaires. Dans un premier temps les hormones androgènes fœtales vont déterminer l'évolution des conduits mâles. En leur absence la différenciation se produit spontanément dans la voie femelle, il n'y a pas d'hormones gynogènes embryonnaires. L'activité androgène des testicules diminue jusqu'à la puberté. À la puberté, le relais est en quelque sorte pris par l'hypothalamus, l'hypophyse et les hormones gona-

dotropes qui vont déterminer les caractères sexuels secondaires, la spermatogenèse et l'ovulation. Le tableau 4 résume les modes de transmission des différentes parties du génome humain selon le sexe.

La reconnaissance du sexe d'un individu peut nécessiter un test génétique pour des raisons médico-légales, archéologiques ou médicales (voir § 6.3). Elle fait appel à un gène, celui de l'amélogénine, qui existe sous deux formes, un variant X et un variant Y qui n'ont pas le même poids moléculaire, les deux variants sont présents chez l'homme, seul le variant présent dans X se retrouve chez la femme (fig. 24 & 25).

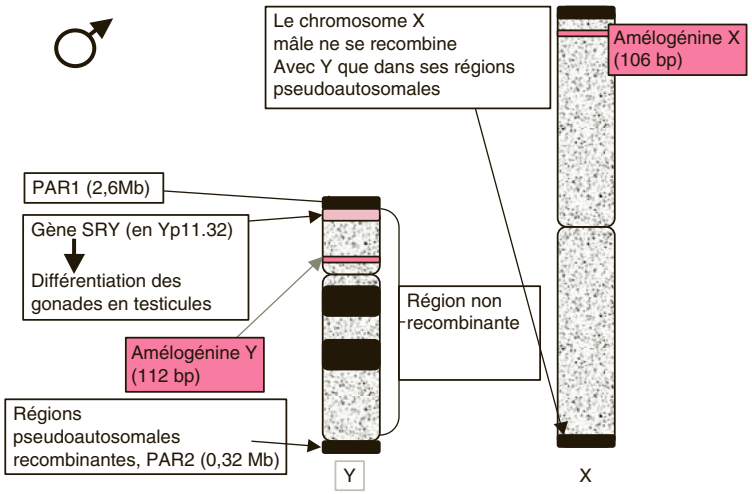


Figure 24 Les chromosomes masculins.

Chez l'homme il n'y a qu'un seul chromosome X et un second chromosome tronqué, le chromosome Y. La majeure partie des deux chromosomes est non recombinante, les deux seules portions recombinantes sont situées aux extrémités des chromosomes (PAR 1 et PAR 2). Le gène SRY situé sur le chromosome Y code pour les protéines responsables de la différentiation des gonades en testicules. Les allèles X et Y de l'amélogénine (*AMELY,* 112 paires de bases et *AMELX,* 106 paires de bases) permettent le diagnostic du sexe en médecine légale.

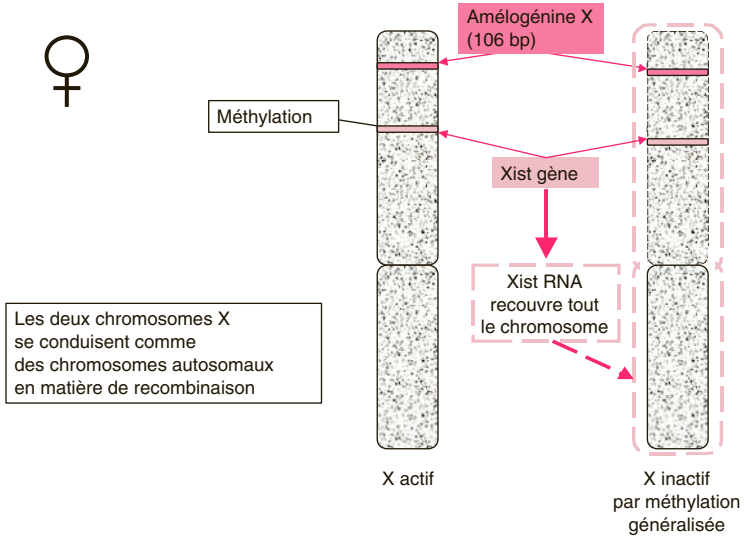


Figure 25 Les chromosomes féminins.

Les deux chromosomes X sont recombinants. Ils contiennent l'allèle X de l'amélogénine.

Chapitre **4**

Variabilité génotypique et variabilité phénotypique

4.1 VARIABILITÉ DU GÉNOME

4.1.1 Généralités

Les êtres vivants sont phénotypiquement et génotypiquement polymorphes (fig. 8). Le fait qu'un chat ne ressemble pas à un homme témoigne d'un certain polymorphisme dans les ADN de ces deux espèces. Il en va de même, mais à un moindre degré, au sein d'une même espèce, les ADNs de Monsieur Durand et Monsieur Dupont, sont à la fois homologues et dissemblables. Le polymorphisme de l'ADN peut certes être dû à des mutations[1] chimiques ou physiques, type Tchernobyl, mais il est sur-

1. Tous les changements dans la séquence de l'ADN sont des mutations, mais, par convention, on dit polymorphisme ou variant lorsque l'allèle minoritaire est présent à une fréquence supérieure à 2 % alors qu'on parle de mutation lorsque la fréquence du variant est inférieure à 2 %, mais ceci est pure convention et encore n'est-elle pas acceptée par tout le monde (discuté in Jobling 2004 Box 3.2.).

tout le reflet du brassage génique normal, physiologique, et entre autre de celui qui se produit lors de chaque mitose et lors de la méiose (fig. 21).

La diversité, et son opposé l'unicité, génétique sont toutes deux indispensables à la perpétuation de la vie en général et à celle de l'espèce humaine en particulier. Chez l'homme, la diversité est assurée par plusieurs mécanismes : (1) le caractère sexué de la reproduction qui assure un premier mixage entre les héritages parentaux, (2) l'existence de recombinaisons au cours de la division méiotique qui assure des échanges entre les deux héritages, (3) les variations ou mutations survenant lors de la mitose ou sous l'influence de facteurs environnementaux.

Sur un plan phénotypique, il faut se rappeler plusieurs données de base. (1) Sur le plan génomique, il y a certes entre deux hommes des milliers de différences, mais il y a aussi et surtout des milliards de ressemblance. (2) La très grande majorité de l'ADN est anonyme, non codante, et donc la très grande majorité des polymorphismes passera inaperçue en termes de phénotype. (3) Le fait pour un polymorphisme d'être situé sur un gène ne signifie pas obligatoirement qu'il y aura modification phénotypique : de nombreux acides aminés sont codés par des triplets différents dits synonymes et il faut, pour être sensible au niveau phénotypique, qu'une mutation change l'expression d'au moins un acide aminé et porte donc sur des triplets non synonymes. On peut aussi souvent modifier plusieurs acides aminés dans une protéine sans en altérer la fonction. (4) Un gène c'est l'ensemble de nucléotides qui contient toute l'information nécessaire pour transcrire un ARNm susceptible de fabriquer une protéine (voir § 3.2) Cette définition inclut donc la portion codante du gène, mais aussi tous les éléments régulateurs c'est-à-dire à la fois la portion régulatrice située en amont (en 5', voir figure 9, éléments dont la dimension et la situation sont souvent inconnues) de la portion codante, dans l'ADN anonyme et les éléments du génome codant pour les dsARN responsables de l'ARN interférence (voir § 3.3).

L'ADN comporte en moyenne une variation tous les 2-300 nucléotides, ce qui veut dire qu'il y a, en moyenne, au même endroit une fois sur 2-300 une base différente. Ce polymorphisme peut siéger soit dans la portion quantitativement la plus importante du génome, c'est-à-dire la portion

anonyme, soit au milieu d'un gène. Dans ce dernier cas le polymorphisme peut être situé dans la portion codante, ou dans la portion régulatrice du gène. Ce polymorphisme, même lorsqu'il est intra génique n'a généralement que peu de conséquences. Dans un nombre limité de cas il va soit modifier l'expression du gène, la protéine aura un acide aminé anormal, ce qui n'est à nouveau pas toujours pathologique si cet acide aminé n'a pas un rôle fonctionnel important, soit changer, bloquer le système de régulation, il pourra par exemple empêcher la transcription du gène (fig. 26).

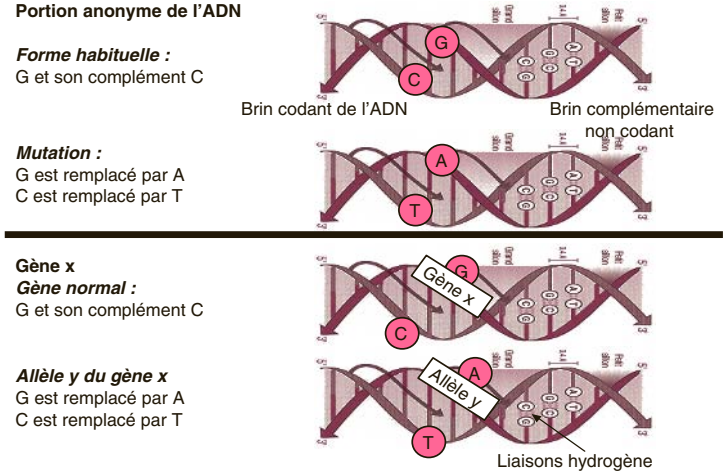


Figure 26 Polymorphisme de l'ADN.

Il y a polymorphisme d'un double brin à l'autre. À l'intérieur du double brin il n'y a pas polymorphisme mais complémentarité. Lorsque, au cours de la conversion génique, à l'intérieur du double brin, il n'y a pas complémentarité entre les deux brins, ce polymorphisme est, généralement, immédiatement corrigé par des mécanismes réparateurs qui complémentent correctement le deuxième brin qui n'était pas appareillé correctement (ce sont, entre autre, ces mécanismes qui sont déficients dans le cancer). Le polymorphisme peut siéger dans la partie anonyme du génome, qui est la plus importante, ou au sein d'un gène et dans ce cas au niveau d'un exon, d'un intron, ou à celui de la portion régulatrice du gène.

Les mutations peuvent porter sur des segments de longueur variable (tab. 5). Ces variations dans la structure du génome peuvent déterminer l'apparition de nouveaux phénotypes qui vont d'un changement mineur à une maladie ou une prédisposition génétique voir au cours de l'évolution à l'apparition de nouvelles espèces. On dit qu'il y a épistasis lorsqu'une mutation a lieu au début d'une chaîne métabolique qu'elle interrompt, de sorte qu'elle masque les effets des autres gènes qui en dépendent.

Tableau 5 CLASSES DIFFÉRENTES DE MUTATIONS

Polymorphismes limités *Polymorphismes ponctuels* Simples substitutions de bases Insertion ou délétion (Indel) d'une seule base *Polymorphismes touchant plusieurs bases* Insertion ou délétion de plusieurs bases Changement dans le nombre de répétitions dans un microsatellite
Polymorphismes touchant quelques dizaines ou centaines (kb) de paires de bases Insertion de séquences répétitives anonymes soit des minisatellites, soit des séquences de la famille dite Alu (parceque spécifiquement coupée par l'enzyme de restriction *Alu I*) ou de la famille L1 (LINE1)
Polymorphismes touchant plusieurs kilobases ou mégabases Insertions ou délétions de très gros fragments. Il peut s'agir de duplications en tandem, ou de segments orientés en sens inverse
Polymorphismes touchant des multimégabases voir tout un chromosome Translocation de chromosomes entiers, par exemple dans laquelle il existe un chromosome supplémentaire en plus de la paire existante normalement comme dans la trisomie 21. Ce type de polymorphisme peut être léthal ou donner des malformations graves

4.1.2 Définitions

a) Allèle

Les allèles sont les versions alternatives d'un même gène ou d'un même fragment d'ADN anonyme. Ils diffèrent entre eux par un ou plusieurs nucléotides. Il peut y avoir deux ou plusieurs allèles d'un même gène ou d'une même séquence (version multi-allélique, surtout pour des

séquences répétitives type satellite). La figure 27 montre les deux allèles d'un même gène l'un sur le chromosome paternel, l'autre sur le chromosome maternel. Elle montre également l'anti-sens correspondant. La figure montre également les versions alléliques du gène x, mais cette fois-ci au cours d'un cycle de division cellulaire. Au moment de la mitose les doubles hélices d'ADN paternel et maternel se dédoublent. Les allèles du gène x se dédoublent également. Le double d'un

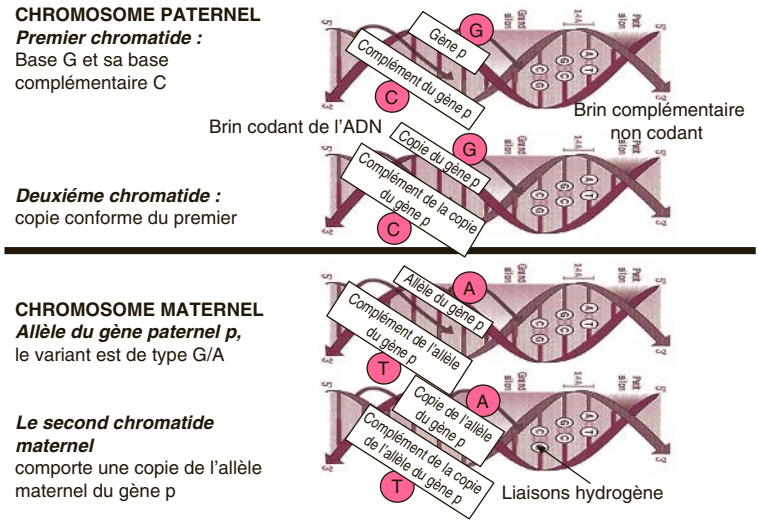


Figure 27 Allèles.

Le gène p du chromosome paternel a un allèle sur le chromosome maternel, cet allèle possède une adénine là où il y a sur le chromosome paternel une cytosine. Bien entendu le brin antisens reste toujours complémentaire, et il existe un allèle antisens. Il ne faut pas confondre allèle et séquence complémentaire antisens. Les mêmes séquences alléliques se retrouvent au cours de la mitose. Ils sont simplement dédoublés et se retrouvent copiés sur chacune des paires de chromatides. Il ne faut pas ici confondre allèle et simple copie sur la chromatide du même chromosome.

allèle est la copie identique du gène, ce n'est pas une nouvelle version allélique. Si le hasard fait que le polymorphisme supprime un site de restriction il est alors possible de le mettre facilement en évidence par PCR (fig. 28).

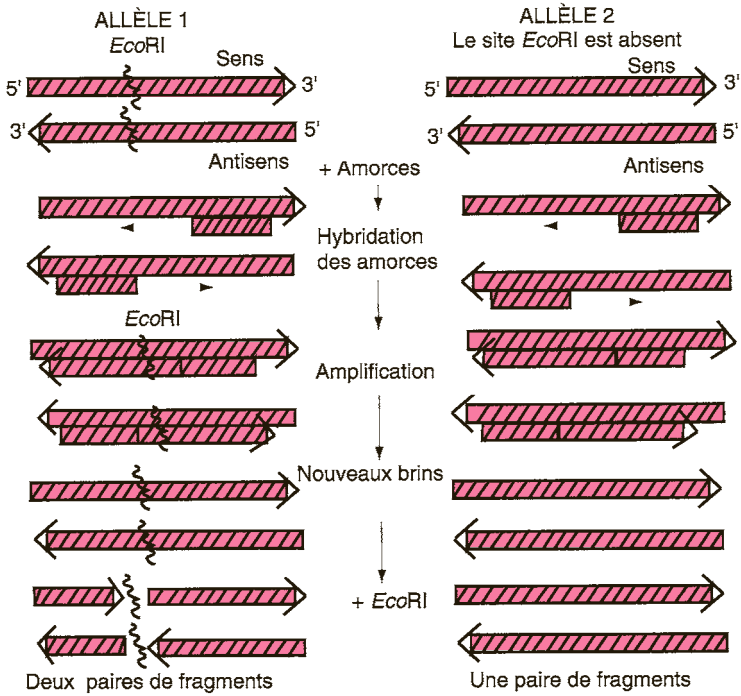


Figure 28 Mise en évidence d'un polymorphisme de restriction au moyen de la PCR.

Supposons le polymorphisme supprimant un site de restriction EcoR1 (en haut et à droite). On va amplifier par PCR (« *polymerase chain reaction* ») l'ADN normal (à gauche) et l'ADN anormal (à droite) en utilisant les mêmes amorces. La PCR va en quelque sorte extraire la séquence intéressante de l'ensemble de l'ADN, il n'y aura plus qu'à la traiter par l'enzyme de restriction pour mettre en évidence le polymorphisme sous forme de différence dans le poids moléculaire.

b) Haplotype et déséquilibre de liaison

Lorsque plusieurs allèles sont proches les uns des autres, sur la même molécule d'ADN (c'est-à-dire le même chromosome ou le même ADN mitochondrial) ils forment un haplotype (fig. 29)[1], et il existe dans ce cas un déséquilibre de liaison, DL. Il y a DL lorsque plusieurs allèles, du fait de leur proximité physique, sont associés entre eux plus souvent que ne le voudrait le hasard. Le DL c'est la tendance qu'ont des allèles d'être co-hérités du fait d'un nombre réduit de recombinaisons. Il y a par exemple DL entre des marqueurs génotypiques de proximité et la mutation caractéristique d'une maladie génétique monoallélique, c'est-à-dire d'une maladie résultant d'une mutation unique dans un seul gène et qui ne s'est pas répétée historiquement (voir plus loin 5.1. Les SNPs, et aussi 5.2. « *Genome-Wide Associations* »).

c) Recombinaison

On en a déjà parlé (§ 3.3). C'est l'échange de segments ADN entre deux chromosomes d'une même paire, cet échange se produit habituellement au cours de la méiose lors du « *cross-over* », il ne peut se faire que si les chromosomes sont strictement homologues et peuvent s'aligner de façon rigoureuse, le processus est réciproque, il n'y a, en effet, ni perte ni gain de matériel, mais simple échange.

Un autre schéma est constitué par les recombinaisons inégales qui expliquent la formation de Familles, voir de grandes familles multigéniques. Par exemple, la grande famille des récepteurs hormonaux nucléaires possédant un site de liaison avec l'ADN. Elle inclut les récepteurs de la thyroxine, de la progestérone, de l'aldostérone… Tous ces récepteurs ont plusieurs éléments structuraux extrêmement homologues (les segments permettant la fixation du ligand, ceux qui lient le récepteur à l'ADN…) qui leur ont été conférés au cours de l'évolution.

1. L'ADN mitochondrial et la très grande majorité de l'ADN du chromosome Y ne sont pas recombinants (fig. 23 & 24), la diversité des haplotypes ne peut, dans ces conditions, être obtenue que par une mutation. Dans tous les autres cas la diversité des haplotypes est obtenue soit par mutation soit par recombinaison.

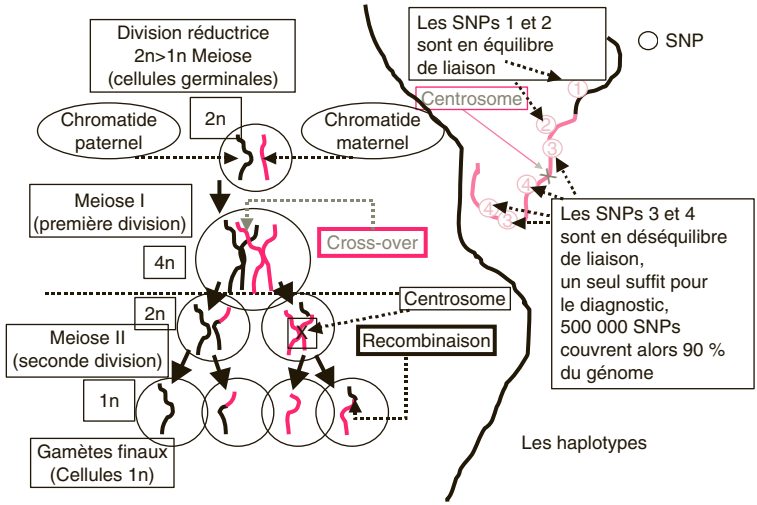


Figure 29 Déséquilibre de liaison et haplotypes.

On voit ici : – à gauche, une répétition de la partie droite de la figure 21, celle qui concerne la division dite réductrice se produisant au cours de la méiose dans les cellules germinales. Il y a lors de cette division un crossing-over consistant en un échange entre les chromosomes paternels et maternels avec recombinaison, ce qui signifie qu'à ce niveau, on mélange les chromosomes. – à droite on retrouve, grossi, le nouveau chromatide; les SNPs (single nucleotide polymorphisms) 1 et 2 proviennent l'un du père, l'autre de la mère et leur emplacement résulte de la recombinaison; ils ne sont pas liés l'un à l'autre et l'identification de 1 ne permet pas de localiser 2, ils sont dits en équilibre de liaison. Au contraire, les SNP 3 et 4 sont cohérités, ils ne proviennent pas de recombinaisons, surtout s'ils sont proches du centromère, et sont dits en déséquilibre de liaison. Identifier 1 ne permet pas, statistiquement parlant, d'identifier 2, par contre identifier 3 permet de savoir où est 4.3 et 4 forment des haplotypes (voir aussi plus loin la figure 42).

4.1.3 Les différentes formes de variabilités génétiques

L'ADN est polymorphe, et les variations de sa séquence sont une propriété essentielle de la molécule. Il est admis en génétique que lorsque la fréquence d'un allèle dans une population dépasse 0,01 % cela veut

dire que cet allèle a été l'objet d'une pression sélective. Il y a une définition officielle du polymorphisme : il y a polymorphisme lorsque la forme rare d'allèle est présente dans plus de 2 % de la population, en deçà il y a mutation (Brent 2000)[1]. On évalue à $1/10^6$-$1/10^8$ le taux d'apparition des mutations spontanées dans le génome. Le polymorphisme de l'ADN est à l'origine de l'évolution sélective, mais la très grande majorité des polymorphismes de l'ADN n'a aucune conséquence évolutive pour au moins deux raisons, d'une part une grande partie du génome n'est ni codante, ni régulatrice, d'autre part une grande majorité des mutations, bien que située sur un gène, n'ont aucune conséquence fonctionnelle notable. Par contre, il est important de se souvenir qu'un gène ne se limite pas à sa portion codante, et comporte une portion régulatrice, non codante, laquelle se subdivise en deux parties les séquences dites non spécifiques régulant l'expression, et surtout le niveau d'expression[2] (en cis, et les séquences spécifiques d'un ou de plusieurs facteurs de transcription (en trans). Le polymorphisme peut exister sur la portion codante où il est relativement facile de le détecter, ou sur les séquences régulatrices où cela est plus difficile.

La variabilité du génome peut apparaître à différents niveaux (Feuk 2006, Khaja 2006)[3] (Addendum 2). (1) Elle peut être détectable au sim-

1. Ce qui sous-entend que certains allèles, relativement fréquents aux USA par exemple, se définissent comme étant mutation dans une autre population où leur fréquence est inférieure au chiffre fatidique de 2 % (Brent 2000).
2. Chez l'homme, ces séquences sont situées à peu prés en nombre égal en position 5' ou 3' (c'est-à-dire en amont ou en aval) par rapport à la portion codante (Cheung 2005).
3. Les séquences du génome humain publiées en 2001 qui servent de référence sont en fait des séquences composites obtenues à partir de l'ADN de plusieurs personnes. Tout récemment, Craig Venter a publié la séquence d'un seul individu (lui-même) ce qui permet une estimation plus précise du polymorphisme de l'ADN. Il y a environ plus de 4 millions de variants ADN dont plus de 3 millions de SNPs, plus de 50 000 substitutions en blocs de 2 à 206 pb, et près de 300 000 indels (Levy 2007).

ple microscope et porter sur le chromosome dans son ensemble[1], sur le caryotype (> 3 kb) et avoir comme conséquence des changements dans le nombre ou la forme des chromosomes, il s'agit de variants structuraux microscopiques, un exemple bien connu est la présence d'un troisième chromosome comme dans la trisomie dont le phénotype est le mongolisme. (2) Le développement récent des techniques d'analyse de l'ensemble du génome a permis la mise en évidence de changements portant sur des segments du génome de taille élevée, entre 1 et 3 kb. Ce sont les nouveaux « variants structuraux sub-microscopiques » qui sont des variations sur grande échelle et incluent des copies en nombre d'un segment d'ADN qui peuvent être des insertions, des délétions, des duplications, et des translocations. Ce type de variants a beaucoup de conséquences médicales et probablement évolutionnistes. (3) Reste enfin les variations plus classiques portant sur un nombre limité de nucléotides (< 1 kb). Il peut s'agir de polymorphismes ponctuels, les polymorphismes de nucléotides uniques ou « *single nucleotide polymorphisms* », couramment appelés SNP ou snips, premiers identifiés dans le génome (1 tous les 300 nucléotides) (fig. 29 et 42), ou de polymorphismes de répétition, qui sont surtout des marqueurs et ont probablement peu de conséquences fonctionnelles (comme les micro- et les mini-satellites). Il peut, là aussi, s'agir d'insertions/délétions (les indels), d'inversions et de duplications (voir plus bas, § 5.1).

Nous ne détaillerons que les variations classiques non codantes, ponctuelles ou de répétition, ce sont en effet les seules qui, pour l'instant, sont utilisées dans les biotechnologies.

a) Polymorphismes ponctuels

L'exemple donné figure 28 est un exemple dans lequel il y a hétérogénéité (c'est-à-dire polymorphisme) ponctuelle. Une seule base est modifiée (quand une pyrimidine est échangée pour une autre pyrimi-

1. Les réarrangements chromosomiques sont un autre exemple. Par exemple, la comparaison entre la distribution des gènes du chromosome X de la souris et de l'homme. Les gènes sont les mêmes, mais leur orientation sur le chromosome est inversée. Les réarrangements se font par groupe de gènes, et touche par exemple six segments chromosomiques homologues chez la souris.

dine, il a transition, lorsqu'une pyrimidine est échangée pour une purine il y a transversion, les transitions sont deux fois plus fréquentes que les transversions). Ce sont les SNPs. La mise en évidence de ce type de polymorphismes a nécessité l'utilisation de couples formés par un enzyme de restriction et une sonde et l'établissement de cartes de restriction (fig. 29). L'insertion ou la délétion d'une seule ou de quelques bases s'appelle un Indel (INsertion DELétion), les substitutions de base sont dix fois plus fréquentes que les Indels. Ce type de modification se produit plus souvent à certains points privilégiés, dits point chauds, de la molécule d'ADN, le plus important étant la présence d'un CpG (p indiquant un phosphate).

b) Polymorphisme de répétition

Environ 25 % de l'ADN humain est fait de séquences dites modérément répétitives, ce qui inclut les copies multiples de certains gènes comme ceux codants pour les ARN ribosomaux et des séquences non fonctionnelles. Beaucoup de ces séquences forment des LINES (pour « *long interspersed elements* »), supposées dérivées de séquences virales, mesurant en moyenne 7 000 pb, il y en a environ 25-50 000 chez l'homme, certains contiennent des séquences codantes. Par ailleurs 10 % du génome humain contient des milliers de SINE (pour « *short interspersed element* »), la plus connue est Alu (300 pb), il y en a 300 à 500 000 copies.

Les microsatellites sont la répétition d'un motif de séquence simple d'une taille de 13 paires de base ou moins, ils sont constitués par la répétition, en tandem, de séquences nucléotidiques, très courtes, par exemple CA (avec sur le brin anti-sens GT) 5, 10, 20 fois (fig. 30). Il y en aurait environ 50 000 répartis dans tout le génome humain. Le polymorphisme c'est ici le nombre de fois où ce motif se répète. Les microsatellites ont un taux de mutation très élevé.

Les mini-satellites (ou VNTR pour « *variable number of tandem repeats* ») sont analogues aux microsatellites, mais le nombre de copies y est plus petit, typiquement 5 à 50. Leur nombre est modifié par « *cross-over* » inégal (voir aussi la fig. 21). Les VNTR existent sous diverses versions, dites alléliques, certains VNTR sont hypervariables

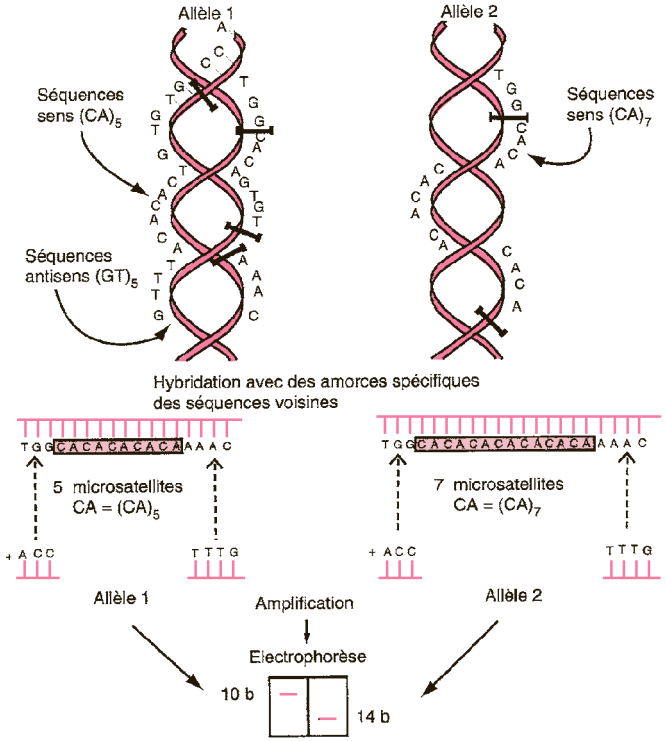


Figure 30 Polymorphisme de répétition : microsatellites.

Les microsatellites sont constitués par des séquences courtes répétitives, formant des sortes de blocs. Le polymorphisme réside dans le nombre de répétitions, mais pas dans la séquence elle-même. On voit ici de haut en bas :
– deux allèles situés sur une paire de chromosomes, l'un est dû à la répétition 5 fois du tandem CA [$(CA)_5$], dans l'autre allèle le motif est répété 7 fois [$(CA)_7$]. Ces microsatellites sont encadrés par des séquences identiques dans les deux allèles (ici arbitrairement GGT et AAAC. On voit également la séquence antisens appariée, elle est formée d'un microsatellite $(TG)_5$ sur l'allèle 1 et $(TG)_7$ sur l'allèle 2, flanquées des séquences CCA et TTTG,
–les mêmes séquences sont ici étalées, elles seront hybridées avec des amorces-PCR complémentaires des séquences flanquantes GGT et AAAC (en pratique ces amorces sont, bien entendu, plus longues), ➥

➥ –après amplification PCR on séparera les produits d'amplification sur électrophorèse selon leur poids moléculaire, et on mettra ainsi en évidence l'allèle 1 qui est plus petit que l'allèle 2.

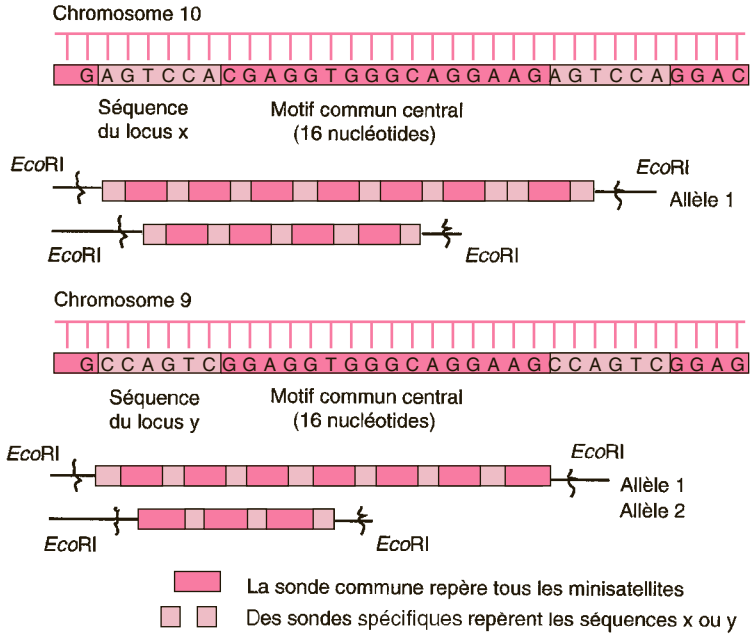


Figure 31 Polymorphisme de répétition : minisatellites.

Les minisatellites sont des séquences nucléotidiques répétitives qui comportent un motif commun central ici de 16 nucléotides (GGAGGTGGGCAGGAAG) encadré par des motifs plus spécifiques du locus considéré. On peut, en sous-clonant ces motifs obtenir des sondes spécifiques de très haute valeur informative. On voit ici deux exemples de minisatellites situés sur deux chromosomes différents, ils possèdent tous deux le même motif central, mais se distinguent par des motifs spécifiques différents. Une carte de restriction utilisant une sonde spécifique d'un de motifs et EcoR1 permettra leur mise en évidence grâce au fait que les fragments de restriction obtenus auront un poids moléculaire différent selon le type d'allèle.

et peuvent être spécifiques d'un même individu, c'est ce taux ce qui en fait un outil précieux en biométrie (voir § 6.3). Les mini-satellites sont autosomaux et n'existent pas sur les chromosomes X ou Y (fig. 31).

4.1.4 Causes du polymorphisme génétique

Les causes du polymorphisme de l'ADN sont nombreuses et ne sont pas toutes, et de loin, des accidents type Tchernobyl. La première d'entre elle est tout à fait naturelle, c'est une incorporation défectueuse lors de la réplication, la seconde ce sont les mutagenèses causées par différents agents chimiques ou physiques.

a) Causes naturelles

Lors de la réplication d'une cellule, les nouvelles bases sont incorporées si elles s'apparient correctement avec les bases correspondantes du simple brin originel (fig. 19). Le nombre de liaisons hydrogène appariées est cependant insuffisant pour que la nouvelle molécule d'ADN soit une copie exacte de l'ancienne. Par ailleurs l'ADN polymérase responsable de la synthèse a besoin d'un appariement parfait avant d'établir une liaison définitive entre les deux brins. Les défauts d'alignement sont corrigés par une exonucléase qui assure une fidélité presque parfaite avec seulement une erreur tous les 10^{-9}-10^{-11} nucléotides. Lorsque se produit un défaut d'incorporation si celle-ci a lieu sur une cellule germinale, elle est transmise et devient héréditaire, si au contraire il s'agit d'une cellule somatique, cela peut avoir peu de conséquences, sauf s'il y en a beaucoup comme dans le cancer. Les mutations qui affectent la fertilité ou qui sont léthales rapidement ne sont bien entendu pas transmises à l'autre génération.

Les recombinaisons qui se produisent lors de la méiose (fig. 21) sont de nature différente puisqu'elles ne font que mélanger des blocs entiers de chromosomes paternels et maternels. Le descendant peut ainsi avoir soit la totalité de l'héritage d'un de ses parents soit avoir un peu des deux (voir plus loin § 6.1 et figure 60).

b) *Causes environnementales*

L'intégrité du génome est soumise à de fortes agressions environnementales. Les principales agressions chimiques sont : la désamination chimique de la cytosine en uracile (qui s'apparie à l'adénine), d'autres réactions chimiques sont connues, les dommages causés par des agents mutagènes comme certains analogues des bases (le 5-bromouracile), certains agents modifiant les bases (l'hydroxylamine par exemple), d'autres qui se lient à l'hélice ADN (la mitomycine C) ou d'autres encore qui s'intercalent dans la séquence (l'acrydine orange). Les agressions physiques sont nombreuses : le rayonnement ultra-violet, les radiations ionisantes comme les rayons X ou celles émises par les isotopes. Ces agressions jouent un rôle déterminant dans la genèse du cancer.

Les lésions causées à la structure de l'ADN sont en général réparées spontanément par un certain nombre de mécanismes et d'enzymes : les séquences endommagées peuvent être excisées par des endonucléases et l'ADN est réparé en utilisant le brin intact par une ligase, les mauvais appariements demandent pour être identifiés une reconnaissance du nouveau brin laquelle se fait par une sous-méthylation des brins néoformés, enfin il existe des mécanismes de réparation des dommages double brin, elle fait appel à un certain nombre de protéines dont l'une, BRCA1, joue un rôle protecteur bien documenté dans le cancer du sein. Les mutations survenant dans les protéines des systèmes de réparation sont particulièrement graves et souvent cancérigènes.

c) *Quelques exemples*

À titre d'exemples voici quelques-uns des résultats obtenus après réparation des anomalies provoquées soit par des recombinaisons soit par des agressions externes.

La conversion génique est due à la recombinaison entre deux allèles ou deux séquences homologues qui forment des hétéroduplexes avec de mauvais appariements (« situation après cross-over »). La cellule va immédiatement réparer ces « *mismatchs* », mais elle le fera indifféremment sur le brin sens ou sur le brin antisens d'où la possibilité après réparation d'avoir des homozygotes et des hétérozygotes, à partir d'homozygotes.

Les recombinaisons inégales sont plus complexes. Ce mode de mixage de l'information génétique survient au moment du « cross-over » lorsqu'il y a beaucoup de séquences répétitives très analogues. Certaines de ces séquences s'apparient au moment du chiasma, mais en se décalant (ici vers le haut), ce qui aboutit à augmenter le nombre de copies sur un chromosome aux dépens de l'autre. Ce type de mixage explique la répétitivité de certains gènes comme ceux codant pour les histones, elle est aussi à l'origine des satellites qui sont des séquences répétitives non codantes.

Il y a d'autres sources naturelles de polymorphisme, au moment de la réplication de l'ADN au cours de la phase S du cycle cellulaire. Il peut en effet, à ce moment se produire un dérapage du brin néoformé sur le brin matriciel. Ceci aboutit à la re-synthèse de la séquence tandem, par exemple d'une séquence CA, ou à la néo-synthèse au cours de la réparation à partir du brin opposé.

4.2 VARIABILITÉ DU PHÉNOTYPE

4.2.1 Relations génotype/phénotype

Un grand physiologiste (Noble 2007) a, fort à propos, utilisé pour souligner la difficulté qu'il y a maintenant en biologie de passer du gène à la fonction, la métaphore suivante. Il compare les gènes aux notes de musique et la vie à une symphonie jouée par un grand orchestre. Sans notes pas de symphonie, ni d'interprètes, sans gènes pas d'être vivant, mais ni les notes, ni les gènes ne permettent d'expliquer le phénomène final. Le même type de métaphore est utilisé par Hans Noll (2003) qui, à propos du langage, compare les lettres aux gènes et, par exemple, les poèmes, à la vie, tout en soulignant à la fois l'importance de la multiplicité des combinaisons et le fait qu'il y a dans les deux cas utilisation d'un langage analogique.

C'est une véritable lapalissade de dire que la vie est complexe, mais il faut bien réaliser qu'il y a dix fois plus de protéines que de gènes, et que chaque protéine possède une structure secondaire particulière, puis une structure tertiaire c'est-à-dire que les sous-unités qui la composent peuvent établir plusieurs types de combinaisons et qu'enfin la structure spatiale de l'ensemble ne découle pas automatiquement (contrairement

à ce que l'on a cru longtemps) de sa structure primaire, mais aussi des véritables moules que sont les protéines-chaperons comme par exemple les « *heat-shock proteins* ». À ces niveaux de complexité proprement moléculaires se surajoutent les interactions entre molécules qui forment des réseaux hiérarchisés par des protéines comme certains facteurs de transcription, qui sont véritablement maîtres d'ensembles. La fonction physiologique enfin est elle-même le résultat d'interactions entre des réseaux qui peuvent être locaux ou systémiques.

L'unité de base de l'être vivant est la cellule, mais l'organisation cellulaire est elle-même l'objet d'une régulation multiple et complexe qui fait intervenir à partir d'un même moule génomique des facteurs de transcription dont l'expression est fonction du type de cellule, du moment de la vie, d'intervenants hormonaux, mécaniques, ioniques, neurologiques, psychologiques… mais aussi des mécanismes épigénétiques qui modifient la structure spatiale de la chromatine sans altérer la séquence des codons.

4.2.2 Plasticité phénotypique et normes de réaction

On peut quantifier la relation existant ente le génotype, le phénotype et l'environnement. Le génotype n'étant finalement que la manière dont un individu donné réagit à une donnée environnementale. La courbe environnement/phénotype pour un génotype donné s'appelle norme de réaction[1] (Griffith 2006, Woltereck 1909) (fig. 32), cette

1. On doit ce terme à Richard Woltereck (1909), zoologiste Allemand du début du siècle dernier qui, en étudiant, plusieurs lignées stables, morphologiquement distinctes, de Daphnées remarqua, le premier, que chacune de ces lignées réagissait spécifiquement à l'environnement. Par exemple, la dimension de la tête augmentait régulièrement de façon exponentielle lorsque le niveau d'alimentation croissait pour la lignée dite Moritzburg, alors que, pour la lignée Brisdorf, la variation exponentielle comportait un seuil, et qu'au contraire la lignée Kospuden était indifférente à cet effet environnemental. Les courbes phénotype (la taille de la tête) versus environnement (l'alimentation) sont clairement et de façon reproductible différentes selon la lignée (le génotype) considéré. Ces courbes ont été depuis retrouvées par de très nombreux auteurs, dans de très nombreuse autres espèces (le crabe, la drosophile…) et pour différentes variations environnementales.

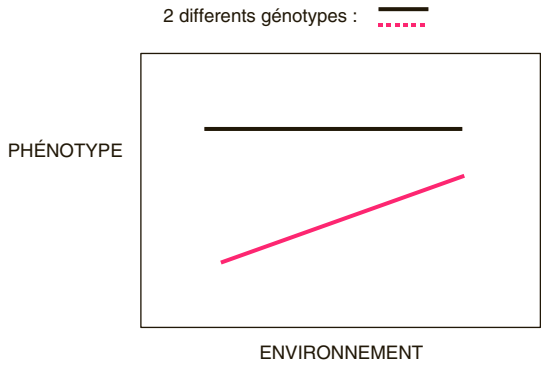


Figure 32 Normes de réaction.
La plasticité phénotypique est la propriété que possède tout être vivant de pouvoir adapter son phénotype à son environnement. Cette propriété varie selon le génotype, elle peut ne pas exister (courbe noire), elle peut exister (courbe rouge). La forme ou la couleur des ailes d'un papillon peuvent ainsi changer selon la température externe.

courbe est reproductible d'une souche animale ou végétale à l'autre et a fait l'objet d'un nombre considérable de publications. Elle définit la plasticité phénotypique d'un individu, ou plus exactement d'un génotype donné.

4.3 CONSÉQUENCES DE LA VARIABILITÉ

Les conséquences du polymorphisme sont nombreuses et pas toutes, et de loin, délétères. Le polymorphisme de l'ADN est d'abord le moteur de l'évolution des êtres vivants. C'est grâce au polymorphisme du génotype que les phénotypes sont différents, encore que les différences phénotypiques tiennent également à de très nombreux autres facteurs environnementaux. Enfin le polymorphisme génotypique est l'élément de base de la génétique médicale.

4.3.1 L'évolution des êtres vivants

« Rien n'est vrai en biologie excepté à la lumière de l'évolution »[1]. Le temps est ici un facteur majeur, puisqu'on estime grossièrement qu'il faut des dizaines de milliers d'années pour qu'une mutation fonctionnelle apparaisse (Ridley 2004, Stearns 2005).

a) Loi de l'évolution

L'évolution biologique est une loi, la seule véritable loi en biologie. Cette loi a reçu d'innombrables confirmations, et les controverses soulevées par les créationnistes sur sa nature ou son existence ne sont pas d'ordre scientifique (Ridley 2004). L'évolution est un processus discontinu, toujours en cours, caractérisé par des périodes de stase au cours desquelles rien ne se passe comme si le processus s'était gélifié, suivies généralement par des ponctuations, c'est-à-dire des périodes au cours de laquelle se produisent des changements d'espèce très rapides. Par définition, l'évolution est calquée sur l'environnement, c'est-à-dire sur l'histoire de la terre laquelle comporte trois grands volets : les changements climatiques avec alternances de périodes de glaciation et de réchauffement, l'apparition de grandes catastrophes d'origine volcanique ou météoritique qui aboutissent à l'extinction massive de nombreuses espèces[2], et les mouvements tectoniques des continents[3].

b) Évolution darwinienne

Tous les êtres vivants ont le même ancêtre, cette affirmation est probablement ce qu'il y a de plus essentiel dans l'apport de Charles Darwin (Darwin 1859), mais il s'est écoulé plus de 3 milliards d'années pour

1. Cet aphorisme célèbre est de Dobzhansky T, il justifie la place donnée à l'évolution dans cet Abrégé.
2. La plus célèbre étant celle survenue il y a 65 millions d'années et qui a supprimé les dinosaures et les ammonites (voir Addendum 3).
3. Comme celui qui a créé l'isthme de Suez et a permis les deux migrations *Out-of-Africa* qui ont permis aux premiers hommes d'origine Africaine de coloniser le reste du monde.

que la vie se diversifie comme elle l'est actuellement. Qui dit diversification ne signifie pas progrès, sur un plan strictement évolutionniste les bactéries par exemple qui ont 3 milliards d'années d'âge sont toujours parfaitement adaptées à leur environnement. Le mécanisme général de l'évolution darwinienne est adaptatif, il est basé sur la pression sélective : des mutations surviennent naturellement à chaque mitose, elles engendrent des changements d'expression phénotypique qui confèrent à celui qui la porte soit un avantage lui permettant à la fois de mieux survivre dans l'environnement qui est le sien et de mieux se reproduire, soit au contraire un désavantage[1]. Le terme consacré pour qualifier l'avantage sélectif positif est celui de « *fitness* », terme anglais intraduisible dans ce contexte, car il ne signifie pas seulement « forme physique », mais aussi « capacité à se reproduire ».

Il est possible de mesurer la pression sélective exercée dans une espèce donnée par rapport à une autre en mesurant le rapport dN/dS, (ou rapport oméga), c'est-à-dire le rapport entre les substitutions observées sur les codons non synonymes, N, et les substitutions synonymes, S (voir légende du tableau 2). Les premières vont en effet changer l'expression des acides aminés, mais pas les secondes, et il est nécessaire qu'un acide aminé soit changé pour que la protéine soit modifiée et donc pour que la pression sélective puisse s'exercer par l'intermédiaire de la fonction physiologique déterminée par la protéine.

c) *Évolution neutre*

L'évolution neutre porte sur les gènes et les protéines, elle est due au hasard. Les changements dus au hasard dans la fréquence d'allèles qui ne causent pas de modifications dans la fitness, s'appellent dérive géné-

1. Il y a un moyen très simple de reproduire l'Evolution en laboratoire, c'est d'incorporer un gène de résistance à un antibiotique dans une bactérie. On met cette bactérie dans une culture où se trouvent d'autres bactéries, puis on ajoute l'antibiotique. Toutes les bactéries seront alors détruites sauf celles qui contiennent ce gène, lesquelles survivront et pourront se développer même s'il n'y en avait que quelques unes au départ.

tique (« *genetic drift* »). Dans quelle proportion la pression sélective et la dérive jouent-elles un rôle, est un débat qui est loin d'être clos (voir Ridley 2004) (fig. 33).

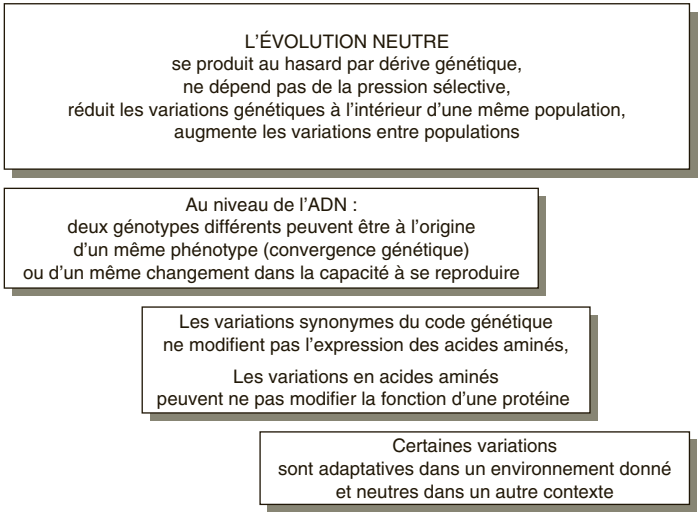


Figure 33 Évolution neutre ou par simple dérive.
L'évolution biologique peut en effet n'être pas due à la pression sélective, mais survenir par simple dérive.

d) Histoire de la vie

➤ Les débuts

Le système solaire résulte probablement de l'explosion d'une ou plusieurs supernova il y a 4,6 milliards d'années, MA. Les plus anciennes traces de vie terrestre ont été retrouvées il y a 3,8 MA. La paléontologie moléculaire a apporté plusieurs données nouvelles éclairantes concernant la nature de l'ancêtre commun prédit par Darwin. La première molécule capable de se reproduire est probablement de l'ARN (acide ribonucléique). Ce monde ARN (4,2-3,6 MA) aurait lui-même été

précédé par un monde pre-ARN dominé par la chimie prébiotique et caractérisé par des polymères capables de s'autorépliquer, mais de façon moins efficaces que l'ARN. L'ADN aurait finalement été sélectionné pour assurer l'hérédité parce que c'est une molécule beaucoup plus stable que l'ARN. À partir de cet ancêtre commun, la première des radiations est celle qui a abouti à la séparation des eucaryotes (avec noyau) des procaryotes (sans noyau) lesquels comprennent les bactéries ou eubactéries et les archées (fig. 34).

Les premiers êtres vivants ont été des bactéries. (1) Leur évolution est lente car les bactéries sont une des formes de la vie les mieux adaptées qui soit. (2) Les procaryotes ont souvent un génome très petit comparé aux eucaryotes, ils ne possèdent souvent qu'un simple chromosome circulaire, ce qui leur permet de se multiplier très rapidement.

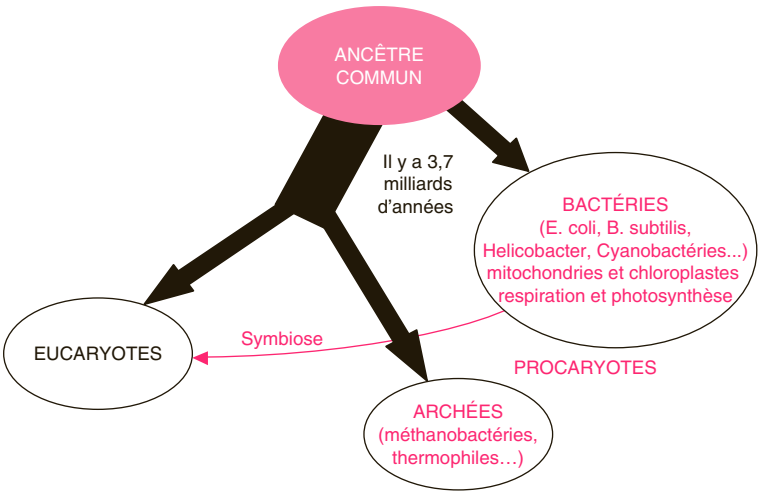


Figure 34 Les trois premières radiations du vivant à partir d'un ancêtre commun, hypothétique pour l'instant.
Il y a maintenant forte probabilité que les Archées ont eu un tronc commun avec les eucaryotes.

(3) L'évolution s'y fait le plus souvent de façon horizontale par transfert de gènes d'une espèce à l'autre et non verticalement comme chez les eucaryotes par changement dans la structure du génome d'une génération à l'autre. Ce sont les bactéries qui ont changé l'atmosphère par la photosynthèse oxygénique (la production de l'oxygène atmosphérique a commencé autour de 3,5 Gyr) et créé les dépôts minéraux.

Les premiers êtres multicellulaires (les métazoaires) datent d'environ 1 MA. Ils ont été rendus possibles par l'apparition de chromosomes linéaires qui est à la base de la reproduction sexuée, l'apparition des exons et son corollaire la complexification des protéines, l'acquisition des facteurs de transcription et leur hiérarchisation. La différenciation cellulaire et l'embryogenèse en particulier ne sont possibles que si l'activation de quelques gènes dits maîtres, suffit à déclencher un processus qui touchera tout un ensemble cellulaire.

➤ Les grandes radiations

L'histoire permet d'établir un certain nombre de repères grâce à qui on peut dater, grossièrement, une cascade métabolique en la retrouvant dans tel règne et pas dans tel autre. La présence d'une cascade métabolique importante à la fois chez l'homme et le poisson, permet par exemple de dire que cette cascade a 345 millions d'années, peut-être plus, mais pas moins.

On sait maintenant que l'évolution morphologique n'est pas rigoureusement calquée sur l'évolution moléculaire. Les dates les plus marquantes en sont l'explosion cambrienne (0,54 MA) au cours de laquelle sont apparues la plupart des grandes structures corporelles actuellement existantes, la naissance d'Urbilateralia (0,60 MA, un peu avant le Cambrien), c'est-à-dire des premiers animaux symétriques, et des gènes qui déterminent cette symétrie, la famille *Hox*[1], l'apparition simultanée,

1. Le complexe *Hox* est responsable de l'activation spatiale et temporelle de plusieurs gènes qui jouent un rôle déterminant dans le développement. Avant l'explosion du Cambrien, *Hox* était isolé. Il s'est ensuite dupliqué plusieurs fois dans chacun des Bilateralia.

au cours du Silurien (0,500 Ma), du premier végétal terrestre et des premiers véritables poissons; les premiers oiseaux ont été détectés au Jurassique alors qu'il y a déjà à cette période plusieurs espèces de mammifères (addendum 3 et fig. 35 & 36), les premiers hominides marchant sur leurs deux pieds sont apparus en Afrique il y a 4 millions d'années et le premier *Homo sapiens* il y a environ 200 000 ans.

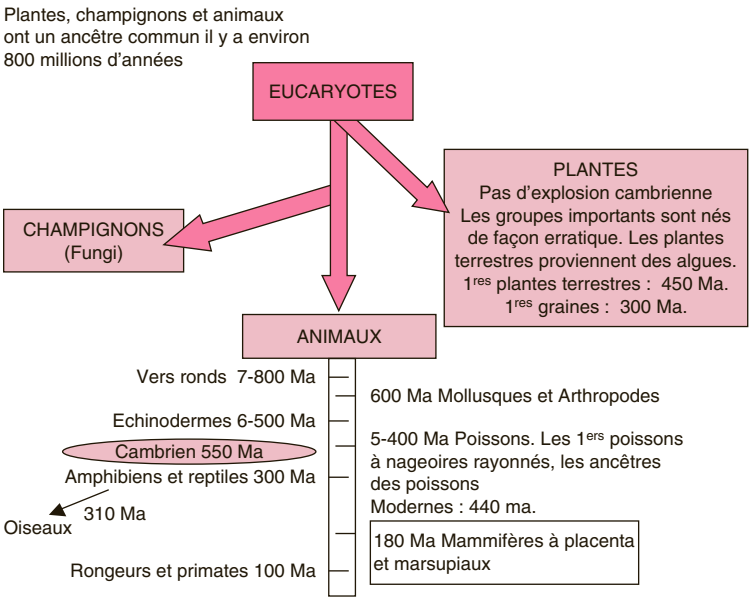


Figure 35 L'évolution des eucaryotes depuis leur début.
Les eucaryotes ne sont pas tous pluricellulaires, les levures par exemple sont des Champignons et donc des eucaryotes. Il n'y a pas eu d'explosion cambrienne chez les plantes qui ont divergé par étapes successives. Le grand événement est l'explosion cambrienne (550 Ma) au cours de laquelle sont apparues la plupart des espèces animales actuelles.

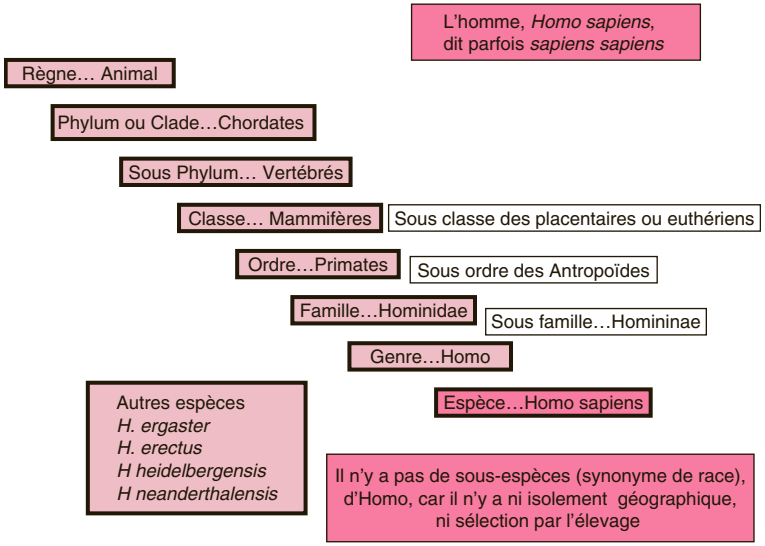


Figure 36 Cladistique.
L'homme appartient à l'ordre des primates, au règne animal…

4.3.2 Variabilité du génome humain

a) La notion de « race humaine »

Les différences physiques inter-ethniques ne portent en fait que sur des aspects mineurs, non significatifs sur le plan de la reproduction, et les « races humaines » ne sont en fait que des entités de nature sociologique, éducationnelle et économique. La couleur de peau, si importante quand on parle de race, est un trait multigénique. Les génomes des noirs, des blancs, des Asiatiques sont remarquablement identiques et les variations rapportées sont inférieures à 0,1 % des 35 000 gènes identifiés dans le Programme Génome Humain. On peut quantifier la diversité génétique et démontrer que la variabilité à l'intérieur des populations rend compte de 85 % de la variabilité totale de l'espèce humaine. Les 15 % restant représentent la variabilité entre les popula-

tions et la variabilité entre les « races ». Le programme informatique STRUCTURE est basé sur un modèle mathématique qui permet de classer les génotypes en fonction du type de microsatellite dont on connaît l'emplacement exact sur le génome. STRUCTURE classe alors ces polymorphismes en quatre groupes sans tenir compte de la « race » apparente supposée de l'individu. Une telle classification démontre de notables différences entre le génotypage et le phénotype, c'est-à-dire la « race » apparente[1].

Les gènes de l'homme et du singe sont identiques à 2-4 % prés, et pourtant les phénotypes sont clairement distincts ne serait-ce que sur des points aussi essentiels que le langage. Une première hypothèse serait qu'il existe en fait des différences importantes dans la partie anonyme du génome portant sur les séquences régulatrices des gènes lesquelles ne sont en fait pas identifiables en tant que telles. Une seconde explication est que les différences entre homme et singe seraient en fait localisées en des endroits déterminants pour l'activité fonctionnelle, il suffit de peu de chose. Enfin, un trait génétique comme le langage, la taille n'est pas le produit de l'activité d'un seul gène, mais celui de l'activité de plusieurs gènes. C'est le caractère combinatoire et hiérarchisé des gènes et, probablement, certaines données épigénétiques, qui créent la diversité et qui font l'homme si différent du singe alors que leurs génomes sont si semblables.

La race est un sous-type de l'espèce. Deux êtres sont d'une espèce différente lorsqu'ils ne peuvent se reproduire l'un avec l'autre (fig. 36). Ce n'est pas le cas pour les races. Les races animales sont le fruit d'isolements, rarement géographiques, plus souvent voulus par l'homme lui-même pour assurer une fonction précise. Le phénotype y est toujours très précis. Génétiquement parlant, l'espèce *Homo sapiens* est à la fois trop homogène et d'apparition trop récente et les « races » humaines sont des créations socio-économiques et historiques qui reflètent

1. 63 % des Ethiopiens par exemple rentrent dans le même groupe que celui des privilégiés, les caucasiens…, avec les Norvégiens, mais aussi avec 21 % des Afro-Caribbéens, soulignant, s'il le fallait, la réalité des échanges ethniques. Classer Chinois et Papous dans le groupe Asiatique, comme cela se fait, est, génétiquement parlant, un non sens.

dans un pays donné, à une époque donnée, l'état de la société parfois jusqu'à l'absurde. Il paraît infiniment plus productif de bannir ce terme de la littérature biologique comme le font tous les grands traités et comme le recommande par exemple le journal « *Nature Genetics* ».

b) Les marqueurs d'origine géographique

Il existe de nombreux marqueurs de l'origine géographique des hommes, les plus utilisés sont des marqueurs anonymes. De plus on peut retrouver l'origine géographique de nombreuses affections génétiques comme le diabète non insulino-dépendant des Indiens Pima, la susceptibilité héréditaire à la tuberculose des Gambiens, sans compter certaines maladies de l'hémoglobine pour lesquelles on a pu démontrer que la mutation première avait eu lieu en Centre Afrique.

L'espèce humaine est caractérisée par sa capacité à se mélanger, et les grandes migrations spontanées ou forcées, comme les traites négrières[1], ont fortement contribué à ce mélange. Sur un plan génétique les Africains sont, de toutes les ethnies possibles, l'entité génétiquement la plus diverse. En effet il est maintenant certain que les premiers hommes sont originaires de l'Afrique de l'Est, il y a à peu près 200 000 ans et qu'ils ont eu 100 000 ans environ pour se disperser et se diversifier sur le plan génétique (et aussi linguistique) à l'intérieur de l'Afrique. Deux migrations « Out-of-Africa » ont eu lieu par l'isthme de Suez il y a seulement 100 000 ans, elles ont colonisé tout le reste du monde, génétiquement parlant l'Eurasie et les Amériques sont plus homogènes que l'Afrique.

4.3.3 Genèse des maladies : génétique versus environnement

Une des conséquences du polymorphisme génique, ce sont les maladies génétiques. Les maladies monogéniques dans lesquelles le phénotype morbide est dû à une mutation touchant les parties codantes ou régulatrices d'un gène fonctionellement lié à l'affection, ce sont les plus rares.

1. 17 millions de captifs semblent l'avoir été par les différentes traites orientales entre le XVII[e] siècle et 1920.

Les maladies multigéniques sont de nature différente et la génétique peut n'y jouer qu'un rôle mineur. Ces affections sont habituellement multifactorielles. Rentrent dans leur déterminisme non seulement plusieurs gènes dits de susceptibilité généralement assez rares (voir le chapitre 6).

La figure 37 schématise un peu la situation. La maladie n'est jamais qu'un seuil dans une norme de réaction, résultante phénotypique de l'interaction génotype/environnement. À l'une des extrémités la génétique joue un rôle majeur, comme dans les maladies monogéniques type mucoviscidose par exemple, à l'autre bout se trouvent les affections dans lesquelles la génétique ne joue aucun rôle, comme la fracture de jambe. Entre ces deux extrêmes se trouvent la grande majorité des affections courantes, cancer, diabète, maladies coronariennes, c'est-à-dire des affections dans la genèse desquelles l'hérédité joue un rôle qui varie de 30 % (dans l'hypertension artérielle) à 60 % (dans l'obésité par exemple).

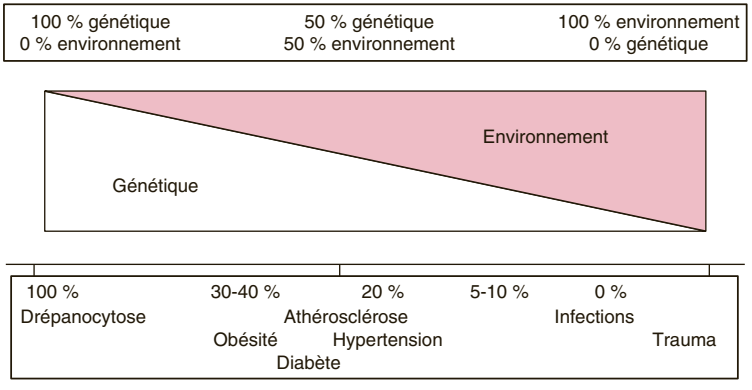


Figure 37 Génétique et environnement (nature versus nurture), les deux partenaires incontournables du fait médical.

La pathologie va du tout génétique (à gauche) au tout environnement (à droite), la majorité des affections contemporaines se situent entre ces deux extrêmes.

Chapitre **5**

Techniques et biotechnologies

5.1 QUELQUES OUTILS DE BASE

On se propose ici d'expliquer les principes généraux des techniques les plus courantes dans ce qu'il est convenu d'appeler le génie génétique et, dans l'ordre, de développer certains principes techniques comme la notion de sonde, d'hybride, d'enzyme de restriction, de polymorphisme, et de donner ensuite les principes des principales techniques d'analyse de l'expression du gène (dosages de transcrits), puis celles d'analyse de l'ADN, de clonage, de séquençage.

5.1.1 Outils basés sur la structure

a) Sondes

Les sondes sont des fragments d'ADN ou d'ARN dont la séquence est connue et utilisés pour la détection spécifique de séquences cibles. Ce sont des outils qui servent à identifier des fragments d'ADN ou d'ARN

présents parmi des milliers d'autres en confectionnant des hybrides spécifiques et quantifiables. Le concept de sonde repose sur le principe selon lequel une séquence de nucléotides donnée s'apparie spécifiquement à la séquence complémentaire. Cet appariement se fait au moyen de liaisons hydrogène qui peuvent être rompues par des moyens physiques.

Les sondes (ou « *probes* ») ADN sont de trois types (fig. 38) :

– Les sondes génomiques qui sont soit la copie d'un gène et peuvent comprendre des séquences introniques ou des portions de la partie régulatrice du gène, soit des copies de portions anonymes non codantes du génome qui sont utilisées pour les études de polymorphisme. Rappelons que dans un organisme donné l'ADN génomique est le même dans tous les noyaux et dans tous les tissus ayant le même équipement diploïde ou haploïde, quel que soit le tissu.
– Les sondes ADNc, c pour complémentaire qui sont des copies monobrins des ARN messagers d'un tissu donné. On les obtient par transcription réverse. Ces sondes sont spécifiques d'un tissu donné (myocarde, foie, mais aussi myocarde adulte ou embryonnaire, foie normal ou cancéreux etc.). Il n'y a généralement pas de sondes ARN, car cette molécule est trop fragile, plus que l'ADN. Par ailleurs ces sondes ne représentent que les exons du gène, et pas ses introns.
– Les oligonucléotides sont des sondes de petite dimension et ne représentent qu'une portion du gène ou de la partie anonyme du génome, elles sont fabriquées artificiellement par synthèse et disponibles dans le commerce. Ce sont des petites séquences d'une vingtaine de nucléotides qui servent souvent d'amorces par exemple au cours de la PCR par exemple. On peut confectionner de cette manière des séquences à la demande, complémentaires de n'importe quelle portion de gène. Un insert est une sonde recombinée à l'ADN d'un plasmide, et donc insérée dans ce plasmide.

Les sondes pour être utilisées doivent être marquées au moyen de marqueurs radioactifs ou fluorescents.

Les hybrides formés entre la séquence d'intérêt et la sonde sont obtenus en faisant appel au principe de l'appariement des bases homo-

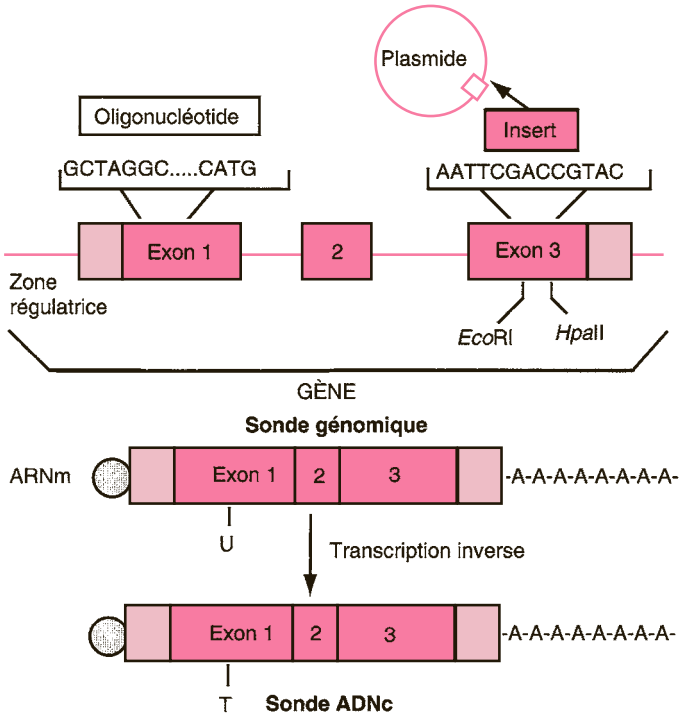


Figure 38 Les sondes.

Les sondes génomiques sont des sondes ADN de même structure que le gène. Par opposition, les sondes ADNc reflètent le contenu spécifique d'un tissu en ARNm, c'est-à-dire l'expression tissulaire du gène. Elles sont obtenues par transcription inverse. On peut aussi (partie haute du schéma) fabriquer artificiellement (il y a des industriels qui offrent ce service) des oligonucléotides qui sont des séquences nucléotidiques connues fabriquées in vitro. Il est assez rare que l'on utilise des sondes couvrant la totalité d'un gène. Il est plus habituel d'utiliser des inserts, c'est-à-dire des fragments de plus petite dimension obtenus après coupure au moyen d'enzymes de restriction (ici *EcoRI* et *HpaII*). Les sondes ADNc (« *cDNA* ») sont obtenues par transcription inverse à partir de l'ARN messager d'un tissu donné. Elles comprennent la partie codante des messagers. U est une base spécifique des ARN, alors que T ne se retrouve que dans l'ADN.

logues. La liaison hydrogène à l'origine de cette hybridation n'est stable que dans certaines conditions particulières de température et de force ionique que l'on appelle conditions de stringence, elles sont spécifiques pour un hybride donné. À l'inverse, à forte température les deux brins d'une molécule d'ADN se séparent, là aussi dans des conditions de température et de force ionique précises et spécifiques.

b) Enzymes de restriction

Les enzymes de restriction sont des endonucléases capables de couper des séquences nucléotidiques à des emplacements reproductibles et définis de façon spécifique en termes de bases. Cette définition n'est pas rigoureusement vraie puisqu'il existe des enzymes de restriction capables de couper des séquences partiellement définies (la Nsp II figure 39). Les plus utilisés des enzymes de restriction font des coupures sur des séquences dites palindromiques[1], ce qui veut dire que la séquence est identique sur les deux brins lorsqu'elle est lue de 5' en 3' dans les deux cas (EcoR I). Après ce type de coupure les extrémités de la séquence se chevauchent (« *sticky ends* »). Il y a des enzymes qui font des coupures franches, ils sont particulièrement utiles lorsque l'on veut par exemple re-souder les bords d'un plasmide (Alu I). Ces coupures sont si spécifiques et si reproductibles qu'elles permettent de définir une séquence, il est courant et habituel de définir une séquence nucléotidique, un gène par exemple, au moyen d'une carte de restriction (schématisée figure 39) sur laquelle on figure les sites dits de restriction, qui indiquent des éléments partiels de la séquence.

Le principe des cartes de restriction est double : (i) il faut d'abord fragmenter l'ADN, qui est une énorme molécule, et il faut que cette fragmentation se fasse à des endroits reproductibles, ce sera le fait des enzymes de restriction. (ii) Il faut ensuite savoir où l'on est, on ne peut pas mesurer la distance séparant le locus d'intérêt du début de la

1. Un exemple de palindrome : *Esope reste ici et se repose,* qui peut se lire dans les deux sens.

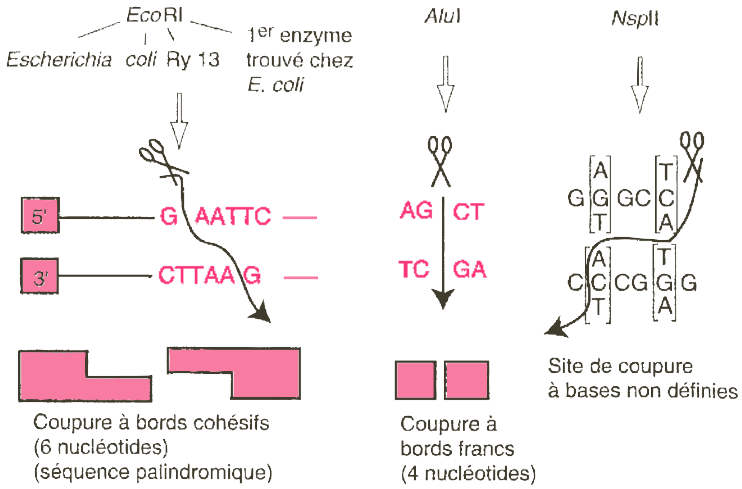


Figure 39 Enzymes de restriction.

La nomenclature retient pour définir ces enzymes le nom de la souche bactérienne dont ils ont été extraits. Des exemples des trois familles d'enzymes de restriction sont donnés : ceux qui coupent des séquences palindromiques, ceux qui effectuent des coupures à bords francs, ceux enfin qui ne sont que partiellement spécifiques.

séquence d'ADN, il faut donc prendre des points de repère, c'est-à-dire utiliser des sondes marquées (avec de la radioactivité ou sans radioactivité) connues, ce qui nous permettra de nous localiser. La partie haute de la figure 40 montre l'aspect diffus, inutilisable, d'une électrophorèse obtenue après coupure avec des enzymes de restriction, mais sans utilisation d'une sonde. Le nombre de bandes obtenues pour une si grande molécule est beaucoup trop important pour pouvoir identifier chaque bande.

La carte de restriction se fait au moyen d'une technique d'électrophorèse, appelée du nom de son inventeur P. Southern. Le « *Southern*

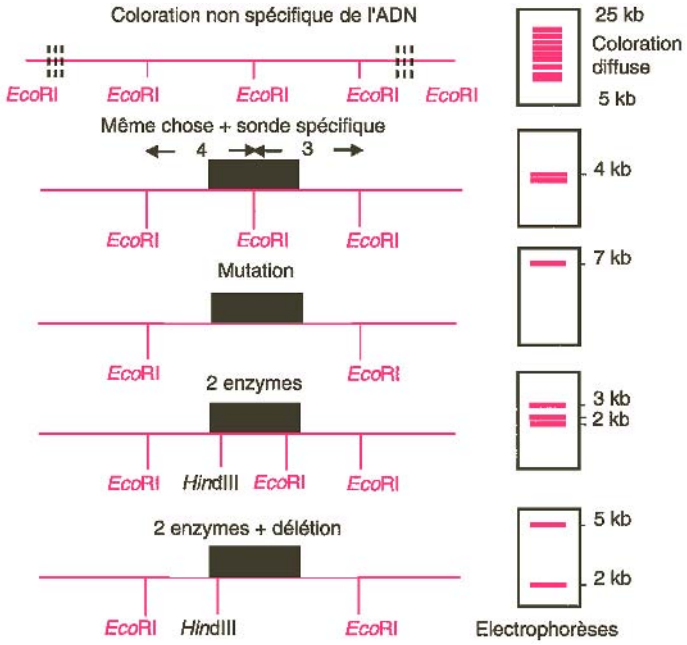


Figure 40 Carte de restriction.

Cette carte est obtenue après digestion d'un fragment d'ADN, voir de tout l'ADN génomique, au moyen d'un ou plusieurs enzymes de restriction (voir figure 39). On ne peut localiser la délétion, le gène, ou le fragment d'intérêt, quel qu'il soit, qu'après hybridation au moyen d'une sonde spécifique (génomique ou cADN) rendue radioactive et qui va rendre l'hybride radioactif. On pourra alors faire migrer le tout en électrophorèse, identifier le segment hybridé en faisant une autoradiographie (ce qui va en quelle que sorte colorer la bande d'intérêt), et connaître le poids moléculaire des fragments coupés par l'enzyme. Comme les sites de restriction (voir figure 39) sont très précis et spécifiques d'une structure donnée, on peut par cette technique identifier rapidement un gène connu par exemple, avec un risque d'erreur qui n'est pas nul (le hasard peut placer deux sites de restriction au même endroit sur deux gènes différents possédant beaucoup d'homologies, mais la probabilité d'une telle coïncidence est très faible), mais qui est très faible. Cette technique est une des techniques de base utilisée en génétique pour analyser le polymorphisme de l'ADN.

blot » consiste à faire migrer en électrophorèse sur un gel les fragments de l'ADN obtenus par l'action d'enzymes de restriction. À la fin de la migration les séquences nucléotidiques sont transférées sur une membrane et hybridées à une sonde spécifique.

L'existence d'une mutation change de façon imprévisible la carte de restriction, comme on peut le voir figure 40. Le nombre de bandes observées, et donc la finesse de l'outil, est proportionnel au nombre de sites de restriction et au nombre d'enzymes utilisés en même temps, une mutation peut aussi bien créer une bande nouvelle qu'en enlever une autre. L'existence d'un allèle, même s'il ne s'agit que d'un seul allèle donne des aspects très différents selon qu'il s'agit d'un homo- ou d'un hétérozygote. Les cartes de restriction ne sont plus guère utilisées en génétique.

c) *Séquençage*

Le séquençage d'un fragment d'ADN ou d'ARN est actuellement facile, rapide et se fait en pratique automatiquement. Il est beaucoup plus facile que le séquençage d'une protéine, ce qui fait que l'on a souvent eu la séquence d'un gène, ou tout au moins de la partie codante d'un gène, bien avant celle de la protéine. La méthode de Sanger est actuellement la méthode de référence. Elle consiste, dans son principe à synthétiser un fragment de l'ADN que l'on souhaite séquencer à partir d'une sonde-amorce dont une extrémité est marquée (rond vert figure 41) et à interrompre de façon aléatoire cette synthèse au moyen de ddTTP, ddATP, ddCTP et ddGTP marqués. De simples électrophorèses permettront de lire la séquence comme indiqué figure 41. La méthode est automatisée, elle permet des recoupements et des alignements de séquence. Les résultats en sont généralement rendus sous forme de courbes de couleur différentes, chaque courbe correspondant à une base. D'autres méthodes existent (voir Gibson 2004 et Griffiths 2004).

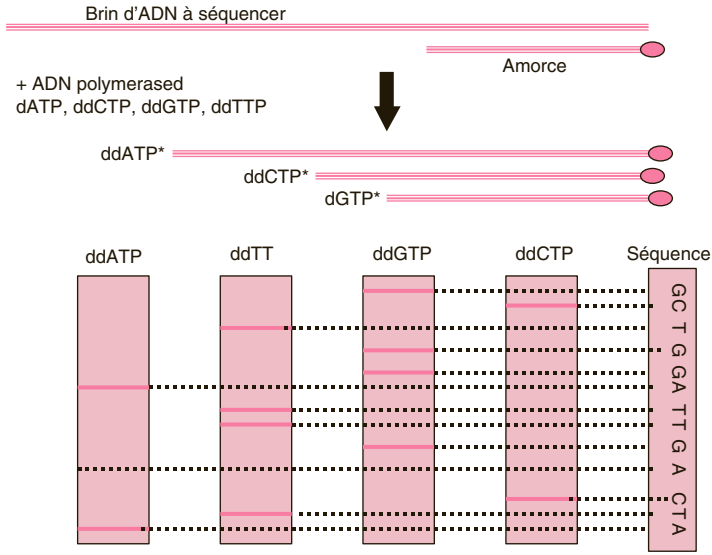


Figure 41 Séquençage de l'ADN selon Sanger.
Principe du séquençage utilisant des didésoxynucléotides (ddA, ddT, ddC, ddG). Voir texte.

d) Les SNP ou snip

Ces témoins du polymorphisme ponctuel (voir plus haut § 4.1) sont des outils actuellement très utilisés en génétique (voir plus bas § 5.2). Ces variations dans la séquence du génome peuvent se retrouver dans les régions codantes ou dans les régions régulatrices des gènes, et peuvent parfois, lorsqu'ils sont situés dans les codons non synonymes, affecter la fonction. Ils forment ainsi des variants voir même des mutants du gène considéré (voir les définitions § 4.1). Les SNP situés dans la portion anonyme du génome sont, bien évidemment plus nombreux puisque la distribution des SNP est en grande partie aléatoire et que la portion anonyme du génome est beaucoup plus importante que la portion occupée par les gènes. Ils se retrouvent tous les 11 à 300 paires de

base. Il existe maintenant dans le commerce des kits de SNPs couvrant le génome avec des densités plus ou moins grandes selon le nombre de SNPs, entre 100 et 500 000 SNPs (Affymetrix 100K ou 500K). Leur utilisation nécessite la connaissance de la carte des haplotypes (fig. 29 et surtout fig. 42).

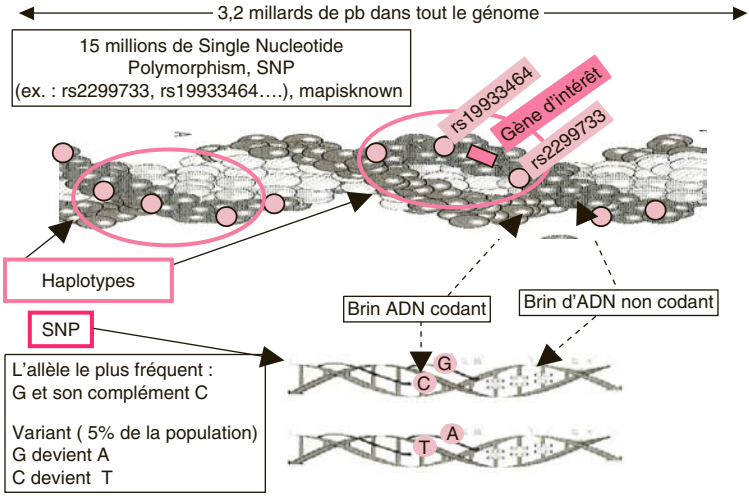


Figure 42 SNPs et haplotypes.

Connaître les seuls SNPs ne permet d'identifier que quelques % du génome. La connaissance des haplotypes permet une économie considérable, il suffit d'un SNP pour identifier tout un groupe de SNPs et de gènes.

5.1.2 Outils basés sur des propriétés biologiques

a) Amplification par « Polymerase chain reaction », PCR

L'un des progrès techniques les plus cruciaux en biologie moléculaire a consisté à utiliser des sondes spécifiques ou amorces (« *primer* » en anglais) et à mettre en pratique le principe de la réplication de l'ADN

(voir fig. 19) pour amplifier des séquences et doser de tout petits fragments d'ADN.

Connaissant la séquence nucléotidique d'un gène on peut faire fabriquer artificiellement des oligonucléotides spécifiques de la séquence. Ces amorces pourront servir de point de départ à la synthèse d'une copie du gène grâce à un enzyme l'ADN polymérase aussi appelée Taq (de *Thermus aquaticus*, le microorganisme dont elle est extraite) polymérase. Cette ADN polymérase ressemble à la polymérase physiologique qui est responsable de la duplication de l'ADN, mais elle est de plus thermostable, ce qui lui confère un avantage pratique.

La PCR est une technique qui permet d'amplifier des séquences d'ADN presque à l'infini et qui est basée sur le principe de la réplication. Comme pour la réplication, il y a action d'une ADN polymérase qui duplique la séquence à partir d'amorces, il faut ensuite artificiellement, par la chaleur, séparer les deux brins d'ADN (alors qu'au cours de la réplication cette opération se fait sous l'action d'une hélicase). La réaction (fig. 43 et 44) se produit dans un petit appareil (peu coûteux) qui est capable de changer vite de température et de passer de 95°, température qui dissocie l'ADN, à 55 °C, température qui permet aux amorces de s'hybrider et enfin à 72 °C, température qui permet à la polymérase d'avoir son action optima. Les amorces sont des copies homologues des extrémités de la séquence à copier, l'une est 5'-3', s'hybride à l'extrémité 3' du brin non-codant et permettra la synthèse de novo, avec des substrats radioactifs, du brin codant, l'autre est 3'-5'.

La PCR est utilisée pour amplifier des séquences ADN peu abondantes, ou, après réverse transcription, des séquences ARN (RT-PCR). La séquence nucléotidique située entre les amorces sera entièrement amplifiée, les deux amorces indiquant de chaque côté (fig. 43) les limites du processus. On peut ainsi amplifier une séquence dont on ne connaît que les extrémités, alors que l'on ignore la structure de la portion médiane, cette possibilité est à la base du développement de l'analyse des microsatellites en génétique, on en parlera plus loin. La puissance de cet outil est telle qu'elle permet d'amplifier sur coupe his-

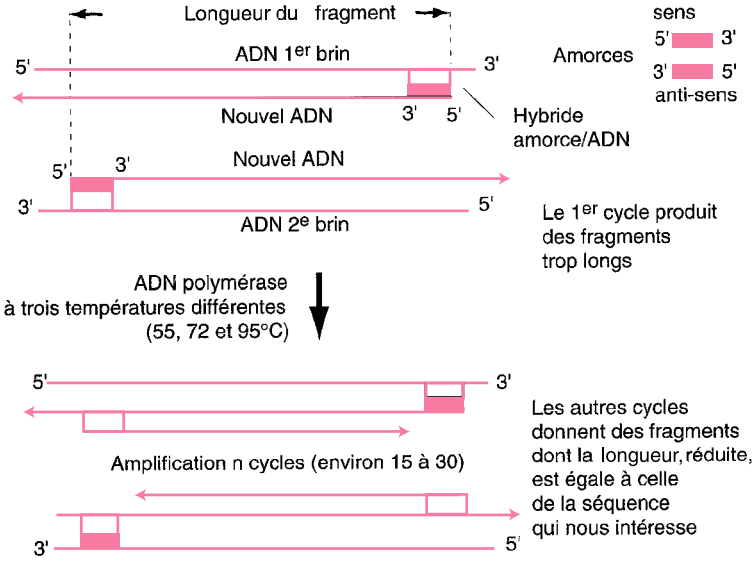


Figure 43 Principe d'une PCR.

La « *Polymerase Chain Reaction* » est une amplification artificielle in vitro à partir d'amorces. La structure des amorces délimite la longueur de la séquence qui sera amplifiée. La multiplication des cycles permet d'amplifier des quantités infimes d'ADN jusqu'à ce qu'elles deviennent visibles en UV en électrophorèse.

tologique des séquences extraites de cellules. L'un des problèmes majeurs de cette technique est l'étalonnage, c'est-à-dire de connaître le nombre de cycles d'amplification. On utilise pour cela des témoins, soit ADN endogène connu, soit séquence analogue à celle qui est étudiée, mais de poids moléculaire légèrement différent, ou sur laquelle on a pu créer par mutagenèse un nouveau site de restriction (fig. 28). De nouvelles techniques permettent de s'affranchir de cette contrainte (« *Real-Time* PCR »).

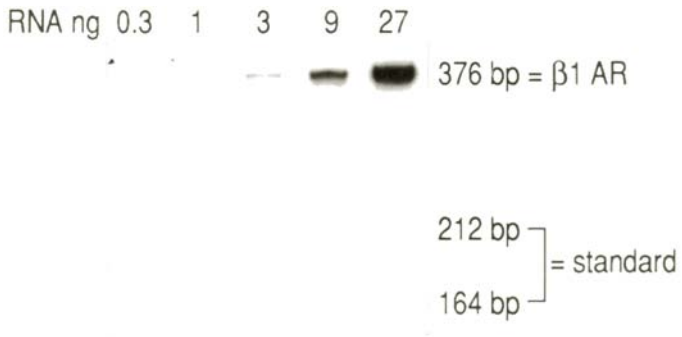


Figure 44 Résultats d'une PCR quantitative.

Cette photo est une autoradiographie d'une électrophorèse sur laquelle on a fait migrer l'ADN obtenu après amplification par PCR. La technique consiste à ajouter au milieu d'amplification une copie mutée de la sonde que l'on souhaite amplifier. la mutation a généré dans la sonde un site de restriction absent dans la séquence sauvage. après amplification on traite le mélange ADN à tester/sonde mutée par l'enzyme de restriction correspondant. La sonde mutée est facilement repérée puisqu'elle est plus petite, ici a peu prés la moitié puisque le site de restriction a été ajoute au milieu de la sonde. Le dosage PCR mis au point ici concerne le messager codant pour le récepteur de l'adrénaline, le récepteur beta1-adrénergique. Le matériel de départ dans lequel on veut faire ce dosage est l'ARN total, qui contient l'ARN messager cardiaque. Ces ARN messagers doivent d'abord être rétrotranscrits, il s'agit donc d'une RT-PCR. Deux amorces complémentaires des extrémités du gène ont été synthétisées, elles ont permis d'amplifier des quantités croissantes d'ADN, ici en NG, de l'extrait de cœur, et une quantité fixe de la sonde mutée. On obtient une bande de poids moléculaire élevé, le gène du récepteur que l'on veut doser qui ne forme qu'une bande puisqu'il ne contient pas le site de restriction, et deux bandes plus petites correspondant aux deux fragments de la sonde mutée. La réponse est dose-dépendante pour l'ADN étudié. elle ne l'est pas pour la sonde mutée dont la quantité est toujours la même (document CEA/INSERM, dû à l'obligeance de M. Elalouf et de J.-M. Moalic).

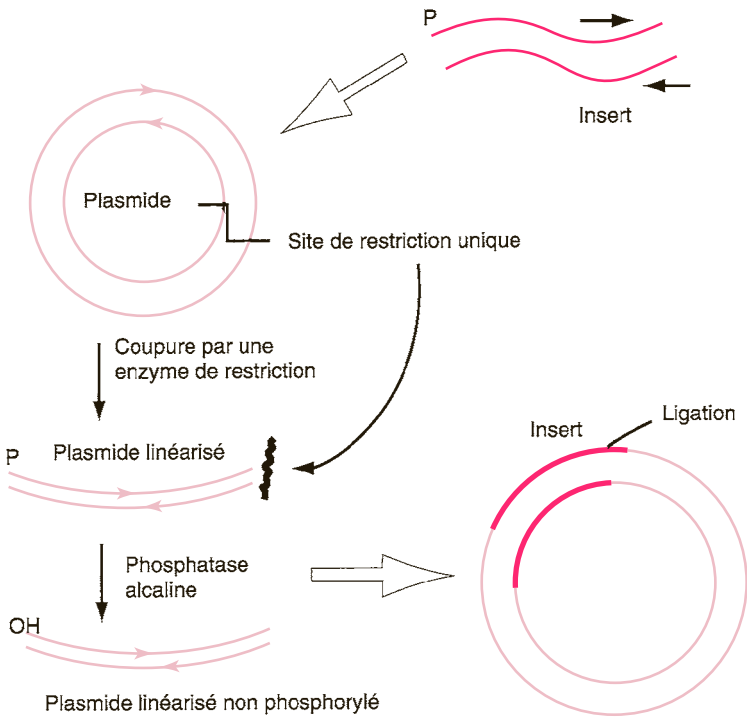


Figure 45 Insertion d'un fragment d'ADN dans un plasmide.

Le fragment double brin, ou insert, sera inséré dans un site de restriction unique, ou « *polylinker* ». Il est nécessaire après l'action de l'enzyme de restriction et avant celle de la ligase de déphosphorylé les extrémités 5′ du plasmide linéarisé afin d'empêcher qu'il redevienne à nouveau circulaire. Le phosphore nécessaire à sa nouvelle ligation (ce mot est un des nombreux anglicismes passé dans le langage courant et signifie ligature) sera fourni par l'insert ou par le gène qui confère au plasmide sa résistance aux antibiotiques dans le cas où l'on veut introduire ce gène.

b) Dosage des transcrits

Les physiologistes sont plus intéressés par l'expression du gène que par sa structure, ils seront donc plus intéressés par l'ARN messager, ARNm, que par les séquences du génome. La quantification des ARNm est basée sur les techniques d'hybridation. On peut détecter un ARNm parmi 10 000 différents ARNm au milieu des 10^6 molécules d'ARN que contient une cellule grâce à des sondes spécifiques, marquées. L'un des problèmes lorsqu'on travaille avec l'ARN est la présence d'ARNase contaminante qui peut altérer les préparations. Manipuler l'ARN demande donc des précautions particulières, des gants entre autre.

Le « *Northern blot* » est la plus ancienne des techniques. Ce terme est utilisé par opposition à « *Southern blot* », le Northern blot est la même technique que le Southern blot simplement ici on sépare des ARNm et on les hybride à des sondes ADN (généralement ADNc, fig. 38). Lorsque les conditions d'hybridation et de déshybridation sont parfaitement bien mises au point on peut hybrider directement le mélange d'ARN dans de petits puits (« *dot* » ou « *slot blot* »).

Le même principe technique peut s'appliquer à l'histologie, on peut, dans certaines conditions, identifier de cette manière la localisation cellulaire de certains transcrits.

c) Vecteurs – Clonage et Amplification

Les applications de routine de la biologie moléculaire nécessitent souvent l'utilisation de techniques dérivées de la microbiologie afin d'amplifier des sondes ou d'établir des banques ADN. Le principe général en est de transférer la séquence d'intérêt dans un vecteur qui lui permettra de pénétrer dans des bactéries pour s'y multiplier et aussi pour bénéficier des nombreux avantages de la génétique bactérienne.

Un vecteur est un ADN qui permet de transporter une séquence nucléotidique d'intérêt dans une bactérie, où il sera amplifié, ou dans des cellules, voir dans des organismes. Cette séquence est souvent artificielle, sa composition en nucléotides et sa carte de restric-

Principaux vecteurs

1. Plasmides : ADN circulaire extrachromosomique d'origine bactérienne, qui se réplique dans les bactéries indépendamment du fonctionnement bactérien. Taille : 2-5 kb. Capacité : 8 kb. Le plus utilisé; pBR 322.
2. Bactériophages ou phages : ce sont des virus de bactéries qu iont la capacité de pénétrer dans les bactéries, de s'y multiplier et de les détruire en se multipliant. Leur capacité de stockage est plus grande que celle des plasmides. Le plus utilisé est la phage λ.
3. Les vecteurs artificiels, comme les vecteurs-navettes et les cosmides. Les chromosomes artificiels de levure (yeast artificial chromosome, YAC) permettent de transporter de très gros fragments d'ADN de 10 à 1000 kb.
4. Les vecteurs naturels : rétrovirus, adénovirus, cellules hépatiques, lymphocytes, et les liposomes.

tion sont en tous les cas connues, ce qui permet d'y introduire la séquence d'intérêt, voir des séquences comme celles qui confèrent la résistance à un antibiotique. Le vecteur peut se répliquer dans le système récepteur, et il doit le faire sans perturber le fonctionnement de l'hôte.

Les figures 45 et 46 schématisent la technique utilisée pour amplifier un fragment ADN d'intérêt. Il faut préalablement l'insérer dans un plasmide, qui est un vecteur disponible sur le marché et qui contient au moins un site de restriction unique, pour permettre l'insertion. La plupart des plasmides sont vendus avec également un gène. Le plasmide contenant deux séquences inhabituelles, non contiguës à l'état naturel, l'insert et le gène de résistance aux antibiotiques, est dit recombinant. Le plasmide est ensuite incorporé dans des bactéries comme Eschirichia Coli, ne contenant elles-mêmes pas de sites de restriction.

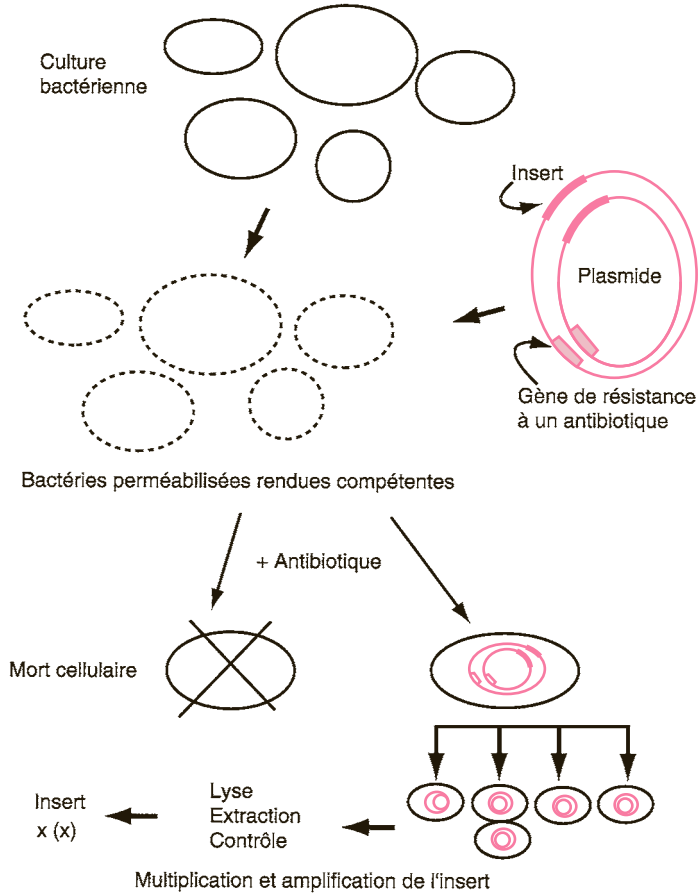


Figure 46 Amplification.

Le plasmide est incorporé par des bactéries rendues compétentes. Comme il contient un gène de résistance aux antibiotiques, le traitement de la préparation bactérienne par cet antibiotique ne laissera survivre que les bactéries contenant le plasmide et donc permettra d'amplifier spécifiquement les bactéries contenant l'insert.

Les figures 47 et 48 montrent un exemple de banque ADN effectué avec des phages. Les phages, ou bactériophages, sont des virus qui peuvent infecter les bactéries s'y multiplier et se faisant les lyser. L'utilisation

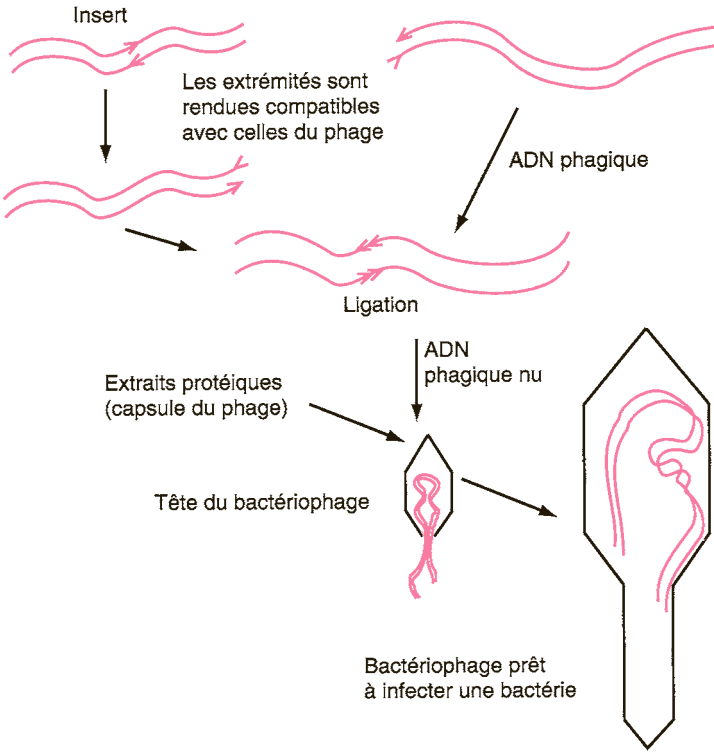


Figure 47 Utilisation des phages comme vecteurs.
La ligation se fait selon le même principe que pour un plasmide (voir figure 45), mais l'ADN phagique n'est pas circulaire et il faut rendre compatible les extrémités de l'ADN de l'insert avec celles du phage. On reconstitue ensuite la capsule phagique au moyen de ses protéines.

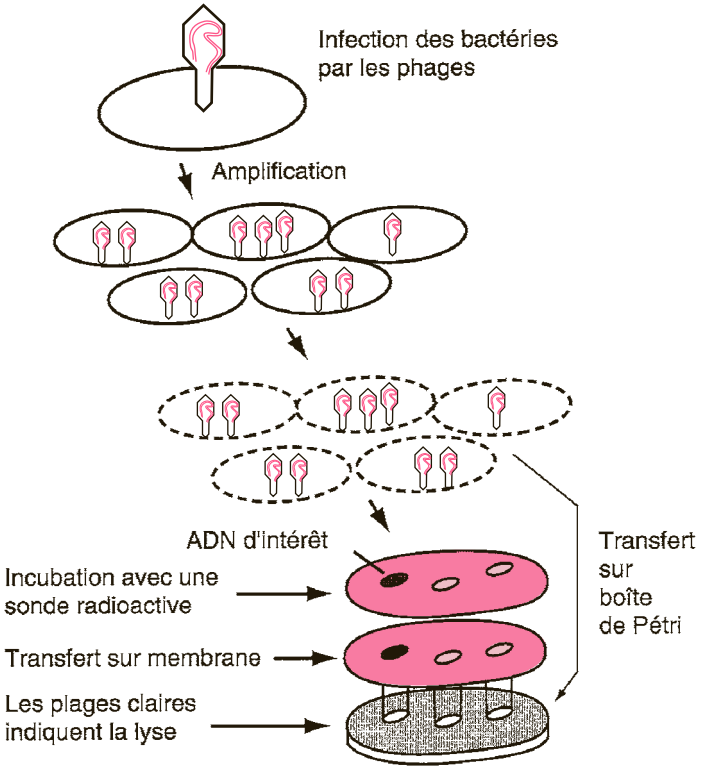


Figure 48 Confection d'une banque d'ADN en utilisant des phages comme vecteurs.

Le phage contient l'ADN d'intérêt et va infecter les bactéries, puis s'y multiplier et en se multipliant les lyser. Les bactéries sont étalées sur des boîtes de gélose, les plages claires indiquent la lyse bactérienne. On effectue ensuite un transfert des ADN sur des membranes. La banque ainsi établie devra ensuite être criblée, ce qui sera fait ici au moyen d'une sonde homologue chaude. Le spot radioactif sera prélevé avec une spatule puis réintroduit dans de bactéries pour y être amplifié.

des Phages comme vecteur permet d'augmenter les rendements et de se servir de fragments d'ADN beaucoup plus grands. Il y a lieu d'abord de libérer l'ADN phagique qui est normalement encapsulé. On effectue ensuite une ligation, comme avec un plasmide, mais en rendant les extrémités de l'insert compatibles avec celles de l'ADN phagique, et on reconstitue la capsule phagique à partir de ses protéines. Le phage est alors prêt à infecter les bactéries. Les cultures bactériennes doivent être très abondantes, et sont étalées sur boîtes de Pétri, les plages claires indiqueront la lyse bactérienne. Il faudra ensuite transférer l'ADN sur une membrane et procéder à la phase dite de criblage. Les banques ADN peuvent être génomiques, et dans ce cas elles contiennent la totalité du génome et peuvent donc permettre l'isolement de fragments non codants, voir anonymes, elles peuvent être faites d'ADNc et dans ce cas elles ne représentent que l'ADN exprimé dans un tissu donné.

d) Criblage d'une banque

Le criblage d'une banque consiste à rechercher le gène d'intérêt dans un des spots sus-décrits. On veut par exemple isoler et connaître les protéines présentes dans le tissu fœtal et absentes du tissu adulte, sous-entendu les protéines responsables ou caractéristiques de la différenciation cellulaire. La technique consiste à préparer les ARNm totaux des deux types de tissu. Par transcription reverse, on transforme les ARNm du tissu fœtal en ADNc et l'on hybride ces ADNc avec les ARNm du tissu adulte. Les ADNc spécifiques du tissu fœtal resteront monobrins et pourront être isolés et incorporés dans un vecteur pour former une banque. Les clones isolés à partir de cette banque pourront s'hybrider avec les ADNc fœtaux, pas avec l'ADN adulte.

Dans la plupart des cas la protéine dont on veut isoler le gène est connue, et, au moins partiellement isolée et séquencée. On peut alors tirer avantage de ces données pour isoler la partie codante du gène. Une technique très semblable consiste à cribler une banque au moyen d'une sonde radioactive homologue, en utilisant la séquence d'une isoforme, lorsqu'elle existe. C'est la manière habituelle d'isoler les isoenzymes ou les isoformes.

Principales techniques utilisées pour le criblage d'une banque ADN

1. On possède un fragment de la protéine d'intérêt :
   a) fabrication d'un oligonucléotide de synthèse radioactif à partir de la séquence en acides aminés ;
   b) fabrication d'un anticorps : on utilise des vecteurs d'expression comme le phage λgt11 qu ivont fabriquer les proétines correspondant aux divers fragments ADN, l'anticorps sera rendu radioactif.
2. On ne possède pas la protéine d'intérêt :
   a) hybridation des spots ADN obtenus au moyen d'une purée d'ARNm contenant au moins le messager d'intérêt, après lavage chaque spot va retenir son messager spécifique qu'il conviendra ensuite de traduire *in vitro ;*
   b) banque différentielle : l'ADNc préparé à partir des ARNm d'un tissu A est hybridé avec l'ARNm d'un tissu B, ce qui reste représentera le ou les gènes exprimés dans A et pas dans B.

L'utilisation d'anticorps présuppose que l'on a fait s'exprimer les clones ADN qui ont été séparés sur boîte de Pétri. On utilise pour ce faire des vecteurs d'expression, c'est-à-dire des Phages capables de traduire en protéines les fragments d'ADN qui leur ont été incorporés. On identifie les complexes antigène/anticorps en liant à l'anticorps une protéine A (qui se lie aux immunoglobulines) sur laquelle on a fixé un isotope radioactif.

e) Blocages des ARNm, utilisation de l'ARN interférence

Depuis longtemps, on sait comment utiliser des séquences antisenses et les introduire dans les cellules pour bloquer spécifiquement (on dit « *knock-down* », *KO*) l'expression d'un ARNm. On a utilisé pour cela

des oligonucléotides antisenses (Scherrer 2003). Cette technique est utilisable in vivo et permet, on le verra plus loin, d'obtenir des animaux transgéniques.

L'utilisation de l'ARN interférence, ARNi, est nouvelle. L'ARNi est présente dans la plupart des cellules et permet l'inhibition spécifique de l'expression des ARNm soit en le dégradant (siRNA), soit en se liant à eux (miRNA). Les siRNA et les « *short hairpin RNAs* », shRNA, sont des produits naturels, mais ils sont, depuis peu, utilisés pour l'inhibition spécifique de l'expression d'un ARNm. Cette technique est en train de devenir la méthode de choix. Elle a aussi la capacité de pouvoir être utilisé en thérapeutique humaine. Ces microRNA (~ 22 nucléotides) peuvent être générés à partir des longs dsRNA et servir comme outils spécifiques pour inhiber un ARNm.

5.2 ANALYSES GLOBALISÉES DU GÉNOME ET DE SON EXPRESSION

La publication de la séquence complète du génome humain a été le point de départ d'une série de techniques permettant d'explorer dans son ensemble la structure du génome et l'expression de l'ensemble des gènes. D'une manière générale, la génomique inclut l'étude de la structure, du contenu et de l'évolution du génome (Gibson 2004), ce qui recouvre de fait la génétique moléculaire, et tout ce qui concerne l'expression génique aussi bien au niveau des ARN messagers (le transcriptome) qu'à celui des protéines (la protéomique), et de leur fonction (le physiome). Le point commun à toutes ces techniques est qu'elles analysent la totalité des gènes ou la totalité des ARNs messagers ou des protéines, et pas seulement quelques éléments sélectionnés *a priori*, comme c'était le cas autrefois. La métabolomique met potentiellement en évidence les effets nets des réactions enzymatiques en en mesurant les produits.

5.2.1 Les analyses globalisées de la structure du génome

La recherche génétique utilise actuellement une approche technologique radicalement différente de celle utilisée pendant longtemps. Traditionnellement, les généticiens étaient focalisés sur des gènes simples aux effets potentiels importants et recherchaient de grandes familles porteuses de maladies rares et graves à transmission mendélienne, comme la mucovicidose par exemple. Le nombre de gènes en cause y était relativement limité et les études de liaison (« *linkage* ») ont permis des avancées importantes sur un nombre limité d'affections familiales. C'est de cette manière que furent identifiées les mutations à l'origine de la très grande majorité des maladies monogéniques.

Ce fut ensuite le temps des gènes candidats, c'est-à-dire de gènes potentiellement en cause, choisis à partir d'hypothèses physiopathologiques, le gène de l'insuline par exemple dans le diabète, ou celui de l'angiotensinogène dans l'hypertension artérielle. La technique consistait à essayer d'établir une association entre une mutation présente sur ce type de gène et l'affection. Des milliers d'études de ce type ont été, et sont encore, publiés. Cette technique a elle aussi permis la mise en évidence de certains gènes pathogènes, mais peu ont fourni des résultats reproductibles, comme ceux qui concernent le système HLA dans le diabète de type 1. La nouvelle approche est différente.

« Genome-Wide Association Studies, GWAS »

Il y a parmi les 3,2 milliards de paire de base du génome humain environ 7 millions de variants existant à une fréquence > 5 %, ce sont en majorité des polymorphismes ponctuels, c'est-à-dire des SNPs. La détection de ces SNPs se fait au moyen d'un signal qui est proportionnel au nombre de copies de l'allèle dans l'échantillon considéré. On utilise pour ce faire des kits tout faits du commerce qui couvrent actuellement une grande majorité des variants dont les allèles ont une fréquence > 5 % (fig. 42). Les allèles plus rares ainsi que les copies supplémentaires de gènes peuvent échapper à cette recherche, et ne peuvent être « capturés » qu'en croisant plusieurs collections de SNP. Le problème c'est qu'avec même 500 000 SNPs, on ne couvre guère que quelques pourcents du génome.

La technique n'a pu être rendue possible qu'après la publication de la carte complète des haplotypes humains (HapMap). Les combinaisons d'allèles (ou de SNPs) proches sont des haplotypes, l'existence de ces haplotypes réduit de façon considérable la diversité qu'il y aurait dû avoir s'il n'y avait pas de corrélation ente les SNPs. Le principe de la GWA est le suivant. Chaque SNP de référence « capture » des SNPs de proximité situé dans le même haplotype, parmi lesquels peut se trouver un variant génique lié au phénotype, c'est-à-dire à l'affection que l'on étudie. Cette capture signifie que le ou les SNPs de cette région sont en déséquilibre de liaison (Kruglyak 2005) (fig. 42).

Cette méthode a permis, pour la première fois, l'identification de variants de gènes ou de simples zones d'ADN spécifiquement liés à certaines maladies courantes dont le caractère familial n'est pas évident sur le plan individuel, comme l'infarctus du myocarde, les diabètes type 1 et 2, l'asthme, la sclérose en plaques, l'obésité… Il s'agit de maladies dans lesquelles le facteur génétique joue un rôle, mais un rôle qui n'a rien à voir avec le rôle unique et déterminant joué par les mutations en cause dans les maladies monogéniques à transmission mendélienne. Les facteurs environnementaux, allergie, auto-immunité, alimentation, infections… y sont des déterminants à part pratiquement égale avec les déterminants génétiques.

Les points critiques dans ce genre d'approche sont la rigueur de la caractérisation du phénotype (la maladie), la robustesse de l'évidence statistique, la dimension de la population étudiée (les cohortes étudiées regroupent plusieurs milliers de patients et de contrôles) et surtout la reproductibilité des résultats dans plusieurs types de population, véritable étalon-or de ce type d'études.

5.2.2 Analyses globalisées de l'expression génique, les « omiques »

L'expression des gènes peut s'analyser à plusieurs niveaux de complexité croissante : les ARN messagers, les protéines, les ensembles fonctionnels de protéines comme les cascades métaboliques et les fonctions physiologiques. Les techniques issues de la connaissance de la séquence

globale du génome ont permis le développement parallèle des techniques d'analyse de l'ensemble de l'expression des gènes, ensemble qui est bien entendu spécifique d'un tissu donné, et qui varie selon l'âge, la thérapeutique… Toutes ces techniques ont en commun une partie informatique qui fait de la bio-informatique un outil incontournable en la matière (Tab. 6).

Tableau 6 Quelques sites Web de bioinformatique appliquée à la génétique

Genbank et sites associés http://www.ncbi.nlm.nih.gov/Entrez/ http://www. ncbi.nlm.nih.gov//genome/guide/human/
Institute for genomic research (TIGR) http://www.tigr.org/tdb
human Genome Database http://www.gdbwww.gdb.org
European Bioinformatics Institute (inclus le site de l'EMBO et Swissprof) http://www.ebi.ac.uk
RCSB Protein Data Bank http://www.www.rcsb.org/pdb/
Protein Information Resource (annotation basée sur des séquences d'acide nucléique qui n'ont pas encore été vérifiées chez l'homme) http://www-nbrf.georgetown.edu/pir/searchdb.html/
Swiss-Prot (banque de données des séquences des protéines) http://www.expasy.org/sitemap.html.ch
Représentations tridimensionnelles http://ca.expasy.org/cgi-bin/get-sw3d-entry ?PO2836 http://www.rcsb.org/pdb
Biologie modulaire http://www.ingenuity.com pour Ingenuity Pathways Analysis, Ingenuity® Systems
Ontology des gènes http://www.geneontology.org pour Gene Ontology, GO http://www.kegg.com pour Kyoto Encyclopedia of Genes and Genomes, KEGG http://www.genecards.org nomenclature des gènes

a) Transcriptomique (les « microarrays » ou micro-alignements)

Le premier niveau d'expression est celui des ARN messagers, appelés également transcrits, d'où le nom de transcriptomique donné à cette approche. Il y a environ 2-3 fois plus d'ARN messagers que de gènes. Les puces à ADN contiennent des échantillons d'ADN correspondants à des gènes dûment identifiés et déposés sur un support par robotique (fig. 49). On utilise plus couramment pour les alignements à haute

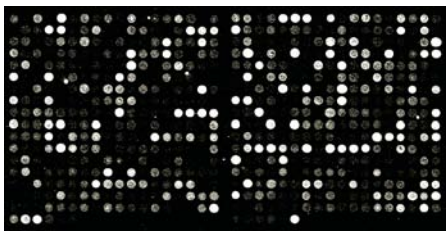


Figure 49 Micro-alignements (« microarrays »). L'intensité des taches est proportionnelle à la quantité d'ARN messager (document dû à l'obligeance de J.-J. Léger).

densité des oligonucléotides, chaque gène étant représenté par 10-20 oligonucléotides spécifiques synthétiques. L'ensemble des ADNc est déposé sur ces plaques. Ces ADNc sont des copies des ARNm d'intérêt provenant d'un tissu donné et obtenu par transcription réverse. Ces ADNc ont été préalablement marqués au moyen de sondes fluorescentes. Des liaisons homologues s'établissent entre la sonde et ceux des ADNc marqués correspondants. La liaison établie produit un signal fluorescent au Laser. Ce signal est normalisé, il est, dans certaines conditions, est quantitatif, c'est-à-dire qu'il correspond à la quantité du gène ou de la séquence d'intérêt présente dans le mélange étudié. La figure 49 est un exemple, chaque tache représente un gène, son intensité est proportionnelle à la quantité du transcrit présent. Il faut bien

souligner que cette technique mesure l'expression d'un gène dans un tissu donné, au cours du développement ou dans des circonstances physiologiques ou pathologiques déterminées. L'expression de ces micro-alignements, couramment appelés « *microarrays* » se présente souvent sous forme de cartes au nom évocateur, les « *heat-maps* » dans lesquelles l'intensité du signal est proportionnelle à la coloration rouge de la tache. Ils sont en train de révolutionner la recherche biologique en permettant de mesurer d'un seul coup la totalité de l'expression génique, la conséquence en est que les publications utilisant cette technique fournissent maintenant des millions de données nouvelles.

b) Protéomique

Les ARN messagers n'ont qu'une seule fonction, ils servent à synthétiser les protéines, mais il y a environ 10 fois plus de protéines que d'ARN messagers, et un peu plus d'ARNm que de gènes. Le moyen le plus traditionnel d'étudier un mélange de protéines est de le faire migrer en deux dimensions en fonction de leur masse moléculaire et de leur point isoélectrique (fig. 50). Cette technique peut être automati-

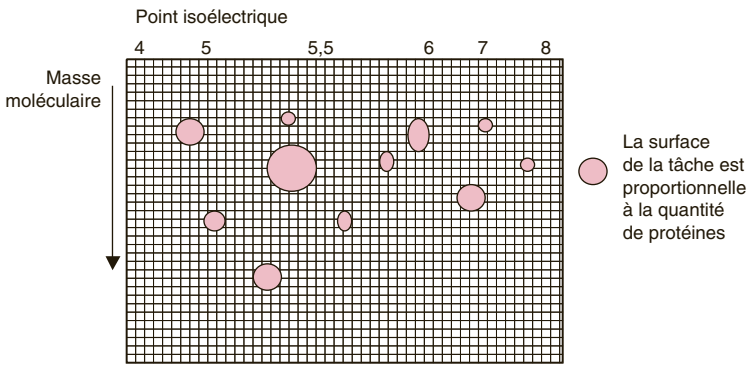


Figure 50 Électrophorèse bidimensionnelle en gel de polyacrylamide d'un mélange de protéines.
Technique de protéomique.

sée et permet d'identifier un nombre raisonnable de protéines. Elle peut être plus précise en la couplant avec un séquençage des protéines par spectrométrie de masse. Plus récemment, on a développé des techniques de micro-alignements des protéines basées sur le clonage de fragments aléatoires de protéines.

c) Métabolomique

Elle consiste à établir le profil de la structure et de la répartition des métabolites, notamment des composés organiques. Pour séparer et identifier ces métabolites, on fait appel à tout un ensemble de techniques qui vont de la spectrométrie de masse après pyrolyse à la chromatographie liquide ou gazeuse (Kell 2004). La métabolomique, couplée à d'autres techniques de génomique, devrait, à terme, permettre la caractérisation complète des voies métaboliques. Elle est, à ce titre, la dernière étape de la génomique fonctionnelle.

5.2.3 Biologie modulaire

Le catalogue des gènes existe, il reste à l'organiser, à en hiérarchiser les éléments et à établir les connections existant entre tous ces gènes. La biologie des réseaux (« *network* » ou « *modular biology* ») est à la fois une nouvelle branche de la biologie, et un nouveau mode de raisonnement en physiopathologie (Barabasi 2004) et en biologie de l'évolution. L'analyse finale de ces bioréseaux est calquée sur la technologie du Web (« *web* » signifie toile) et repose sur une loi simple, plus un nœud a de connexions, plus il a de chances d'être connecté (ce qui est l'évidence sur le réseau Web, comme chacun peut l'expérimenter tous les jours) (fig. 51). Elle permet d'identifier des voies métaboliques-clés, lesquelles deviennent ainsi, sans hypothèse préalable, des cibles physiopathologiques ou pharmacologiques privilégiées. Les premiers essais sont déjà là, et les premières applications ouvrent déjà des perspectives pour la compréhension de maladies comme l'athérosclérose, le cancer de la prostate et le diabète.

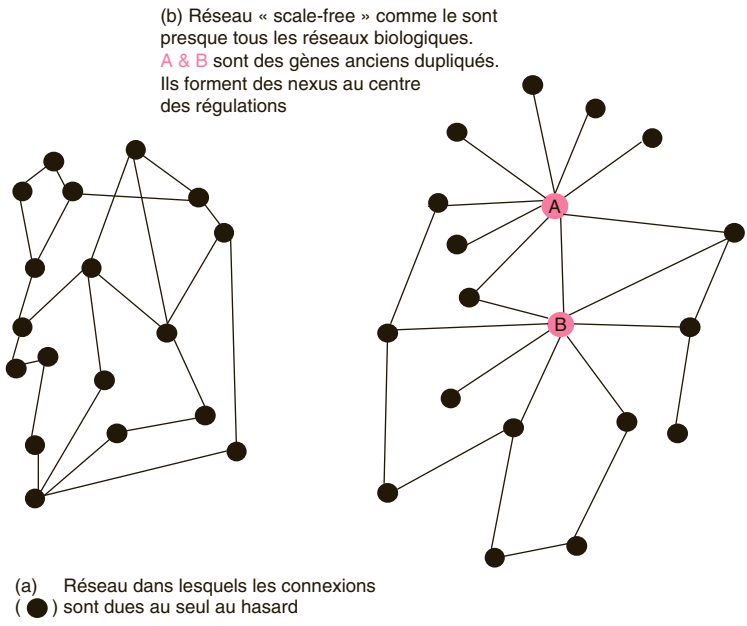


Figure 51 Modèles de réseaux.

(a) Les noeuds sont connectés au hasard par des liens. (b) Dans les réseaux sans échelle, les nœuds sont aussi connectés par des liens. Néanmoins il existe des attachements préférentiels qui génère des « hubs », ou moyeux (en A et B) grâce au mécanisme « les riches-deviennent-plus-riches ».

Les réseaux biologiques hiérarchisés ont très probablement une origine évolutionniste ancienne. Les gènes-maîtres, ceux qui forment des nexus ou ceux qui sont au sommet d'une cascade métabolique comme dans les réseaux hiérarchisés sont, selon toute vraisemblance des gènes dupliqués au cours de l'évolution. Une duplication de gènes va en effet produire deux protéines identiques lesquelles ont forcément les mêmes connexions ce qui va automatiquement amplifier le réseau. L'identification de tels gènes est cruciale en médecine, car elle permet

de mettre le doigt sur les éléments métaboliques déterminants soit pour la compréhension d'une maladie, soit pour son traitement. Il a été proposé que ces réseaux puissent constituer de véritables modules fonctionnels. Ces modules seraient de véritables unités physiologiques qui sont transmises d'un seul bloc au cours de toute l'évolution. La pression sélective s'exercerait non pas sur tel ou tel gène particulier mais sur le module lui-même, et les éléments qui en déterminent l'unité.

5.3 TRANSFERTS GÉNIQUES

5.3.1 L'infection virale

a) Généralités

La manière la plus naturelle dont se fait un transfert de gènes est l'infection virale. Les virus sont de simples capsules membranaires (on dit des capsides) dont le génome est fait selon les cas d'ADN ou d'ARN mono- ou double brin et qui ont la particularité de ne pouvoir se multiplier qu'à l'intérieur de cellules hôtes spécifiques, dites permissives. Les rétrovirus sont des virus qui ne peuvent se multiplier qu'après avoir intégré le génome d'une cellule hôte (figure. 52). Pour s'intégrer dans le génome ADN-hôte, ils doivent d'abord être rétro-transcrits, il s'agit d'une auto-rétrotranscription, car la transcriptase reverse nécessaire est fabriquée par le virus lui-même qui contient le gène *POL* correspondant. La réaction est par ailleurs plus complexe car la transcriptase réverse possède d'autres propriétés et permet, entre autre la synthèse d'une séquence nucléotidique, la séquence *LTR*, pour « *Long Terminal Repeat* » qui joue le rôle d'un promoteur fort non spécifique capable d'activer n'importe quel gène. À cette séquence peuvent s'ajouter d'autres gènes : un oncogène qui jouera un rôle transformant comme le virus du sarcome de poulet *Rous*, ou des gènes régulateurs comme dans le cas du virus HIV, le virus du SIDA.

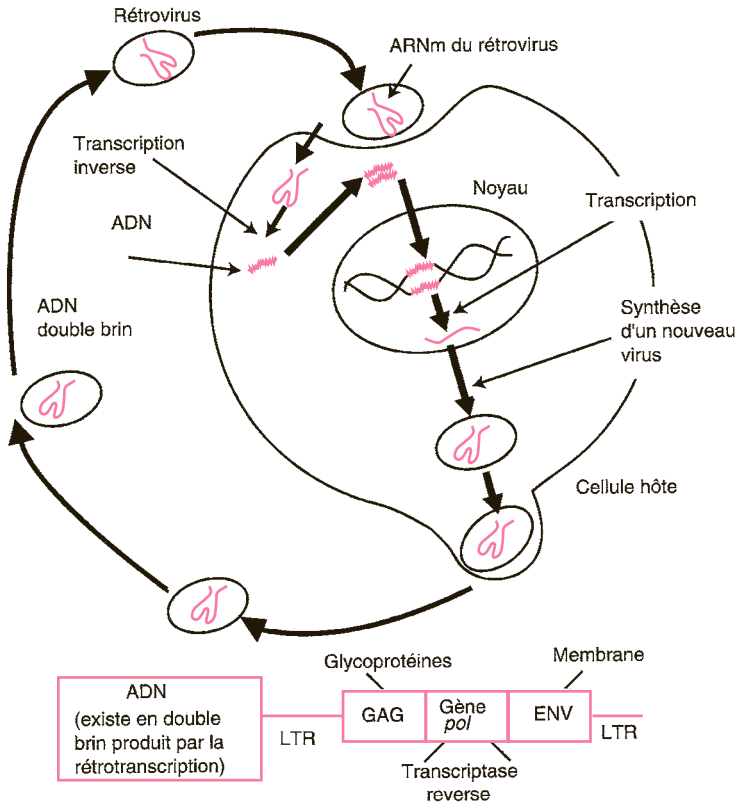


Figure 52 Rétrovirus.

Les rétrovirus ont un génome fait d'ARN qui doit donc être transcrit en ADN pour pouvoir être intégré dans le génome hôte. Cette rétro-transcription est catalysée par une transcriptase réverse codée par un des gènes propres du virus, le gène POL. Cet enzyme a de multiples fonctions et synthétise un ADN double brin copie de l'ARN, et lui adjoint deux séquences qui l'encadrent et sont appelées « *Long Terminal Repeat* », LTR, selon le schéma indiqué en bas de la figure. Ce nouvel ADN peut s'intégrer dans l'ADN-hôte, y former un provirus et y transcrire les ARNs correspondants qui vont à leur tour fabriquer de nouveaux virus.

b) Infection virale, cancérogenèse et SIDA

Les cellules cancéreuses sont des cellules qui ont à la fois acquis l'immortalité, et l'indépendance à l'égard des facteurs de croissance, elles ont aussi perdu l'inhibition de contact. La pathogénie du cancer ne peut bien entendu pas se résumer en quelques lignes. On se bornera ici à en dire quelques mots à propos de la théorie oncogénique et du rôle des rétrovirus dans la cancérogenèse d'origine virale. Les oncogènes sont une famille extrêmement hétérogène de gènes et les oncoprotéines sont des protéines participant toutes, à des degrés divers, au processus de croissance et de développement. Selon cette théorie, le cancer serait dû à une perversion des oncogènes. Cette perversion peut porter sur le promoteur ou la portion codante, elle est créée par le virus lui-même (mais pourrait aussi être créée par une agression chimique ou physique comme celles provoquées par le tabac ou la radioactivité). Cette théorie tire son origine de la découverte des effets cancérigènes d'un oncogène, *v-src*, présent dans le Virus du Sarcome de Rous (VRS) et l'identification de cet oncogène, *c-src*, dans le génome de nombreuses espèces animales. On a depuis identifié chez l'homme quelques exemples de rétrovirus humains cancérigènes, comme HTLV responsable de leucémies (qui sont des cancers des globules blancs). Le virus de l'hépatite B, n'est pas à proprement parlé un rétrovirus, mais il peut dans certains cas (5 % des hépatites B), s'intégrer dans le génome des cellules hépatiques et les transformer.

Les virus HIV, « *Human Immunodeficiency Virus* », sont responsables du SIDA. Leur description nécessiterait tout un livre. Les virus HIV sont des rétrovirus qui s'intègrent donc dans le génome de la cellule-hôte et infecte les lymphocytes T, l'infection ici détruit les cellules immunitaires, à la différence de l'infection due au virus HTLV, et supprime de fait la capacité de se défendre contre les infections microbiennes normalement inoffensives chez des individus normaux. Ces malades meurent d'infections parfois bénignes. Ce rétrovirus possède la séquence commune à tous les rétrovirus, mais plusieurs gènes supplémentaires lui sont rattachés qui jouent un rôle particulier dans la reproduction du virus.

5.3.2 Transfections

Le terme de transfection est généralement réservé aux transferts géniques faits *in vitro* dans des cultures cellulaires. Il y a plusieurs manières de transfecter une cellule, la plus élémentaire consistant à micro-injecter avec une seringue la construction intéressante. Habituellement on a recours à des techniques plus productives : précipitation de l'ADN par du phosphate de calcium pour pénétrer la barrière membranaire, le DEAE-Dextran agit de la même façon, l'électroporation consiste à perméabiliser la membrane cellulaire au moyen d'impulsions électriques à haute tension[1], on peut enfin projeter par « bombardement » des microparticules revêtues. La liste est importante.

D'autres modes de transfection utilisent les rétrovirus comme vecteurs. Ils permettent des constructions stables, encore faut-il inactiver les virus et leur donner la capacité d'exprimer la séquence d'intérêt. Ces virus ne peuvent s'incorporer dans l'ADN-hôte que s'ils restent d'une longueur relativement constante, il faut donc avant d'incorporer la séquence nucléotidique que l'on veut étudier supprimer par délétion certaines des séquences déjà existantes, mais comme ces séquences servent à la multiplication des virus il faudra les réintroduire à part sous forme d'« *helper* ». Le virus vecteur pourra se développer et s'amplifier dans la cellule-hôte, et ensuite transfecter la préparation qui nous intéresse. On peut aussi utiliser le virus à ADN SV40 qui est un virus du singe capable d'induire des cancers chez le rongeur mais pas chez l'homme ou des adénovirus. Il y a aussi possibilité de transférer des gènes en utilisant des vecteurs plus « naturels » comme les lymphocytes eux-mêmes.

Qui dit transfection dit préparation et manipulation du matériel génétique qui sera transfecté (le discours sera plus loin le même pour les transferts sur animal entier). On peut créer des délétions (fig. 53), c'est-à-dire supprimer une séquence en utilisant la carte de restriction et en supprimant un segment compris entre deux sites de restriction.

1. En utilisant par exemple 3 trains de 100 impulsions de 100 µs à 50 mA chacun fournies par un électropulser type GET42 (www.elecinfopilat.com), ce qui donne un ordre de grandeur.

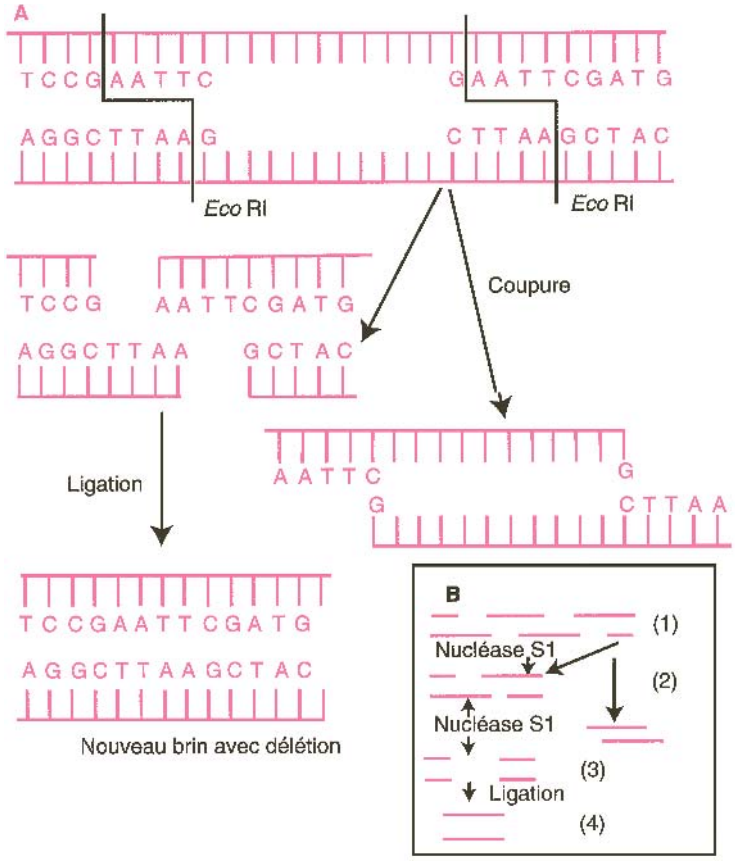


Figure 53 Création de délétions.

En A on se borne à exciser un fragment ADN situé entre deux sites de restriction. Les deux extrémités restantes seront ensuite liées au moyen d'une ligase. On récupère en général le fragment néo-synthétisé (qui est forcément plus petit que le fragment initial, en éluant les bandes obtenues après électrophorèse). En B, il s'agit de la même expérience qu'en A, simplement la coupure par enzyme de restriction a été complétée par une digestion des brins libres non-hybridés par une nucléase S1.

La technique peut être affinée si l'on ajoute une nucléase qui supprimera tous les fragments non appariés.

Il est souvent nécessaire d'obtenir des mutations plus ponctuelles et l'une des méthodes utilisées est celle de la mutagenèse dirigée qui doit utiliser un vecteur particulier comme le phage M13. Ce phage a la particularité d'être un phage monobrin qui peut se répliquer et former deux brins sous l'action d'une ADN polymérase après avoir transfecté une bactérie. La technique permet d'obtenir par exemple des sondes légèrement plus petites que leur original et qui pourront servir de témoin lors d'une amplification PCR.

5.3.3 Animaux transgéniques

La transfection dans une culture cellulaire fournit des renseignements utiles sur la régulation de l'expression d'un gène donné, mais l'interprétation des résultats est toujours limitée par le fait qu'il s'agit de cellules en culture, isolées de leur contexte. La technologie transgénique pallie à cet inconvénient en permettant l'incorporation de constructions dans des cellules germinales et permettant ainsi à la fois une diffusion du transfert génique à tout l'organisme et sa transmission à sa descendance.

Cette technique consiste à injecter une construction génique dans les cellules germinales d'un animal (ce qu'il est formellement interdit de faire chez l'homme). La pénétration de la construction dans le génome est, avec la technique habituelle, aléatoire, c'est-à-dire qu'elle peut se faire n'importe où dans le génome, y compris dans des zones dites silencieuses, et le rendement de la technique est faible de l'ordre de 4-5 %.

L'injection est une micro-injection à la seringue, elle se fait dans le pronucleus mâle d'un oocyte fertilisé de souris (fig. 54). La construction pénètre ainsi directement dans le noyau et s'intégrera, généralement en cascade, c'est-à-dire que plusieurs constructions s'intégreront à la queue leu leu dans le génome en même temps. Il faudra ensuite examiner l'ADN des portées de souriceaux et rechercher la construction initiale par PCR. Les animaux positifs, possédant la construction, seront sélectionnés. Comme il s'agit par définition d'hétérozygotes (puisque

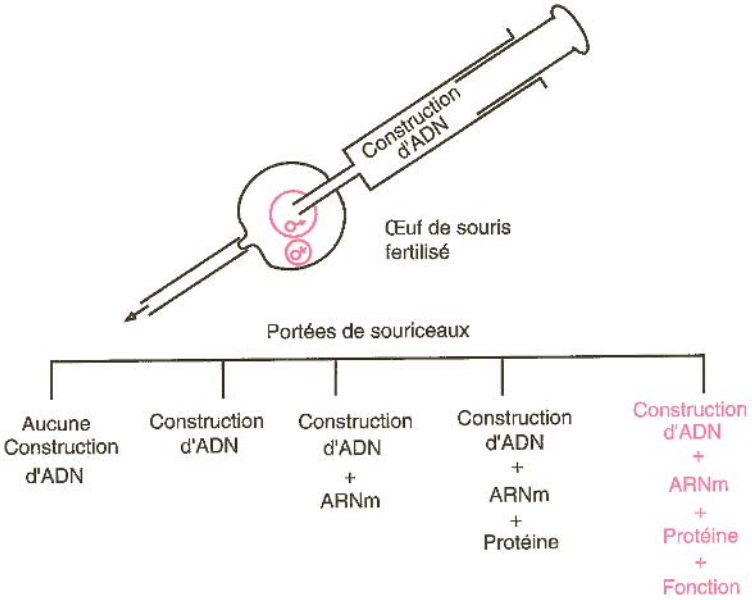


Figure 54 Principe de la technologie transgénique.
On voit sur cette figure l'œuf de souris fertilisé, les deux pronucleus, mâle et femelle, une pipette de contention de l'œuf et la micro-seringue qui sert à injecter la construction dans le pronucleus de l'œuf. On analysera les portées de souriceaux en termes d'ADN, d'ARN, de protéine et éventuellement de fonction.

l'injection est faite dans le noyau mâle), ils seront croisés entre eux afin d'obtenir des homozygotes. On recherchera ensuite l'ARN messager correspondant, puis, soit le marqueur caractéristique du génie reporter (souvent un colorant ou un marqueur fluorescent), soit, s'il y a lieu, la modification physiologique correspondante. La technique est coûteuse, mais son inconvénient majeur vient de son coté aléatoire, on ne sait jamais ni dans quelle portion du génome va s'incorporer la cons-

truction (ce peut être dans une zone silencieuse), ni combien de copies (généralement plusieurs) seront incorporées. Elle est pour des raisons éthiques évidentes formellement interdite chez l'homme. La figure 55 en donne un exemple.

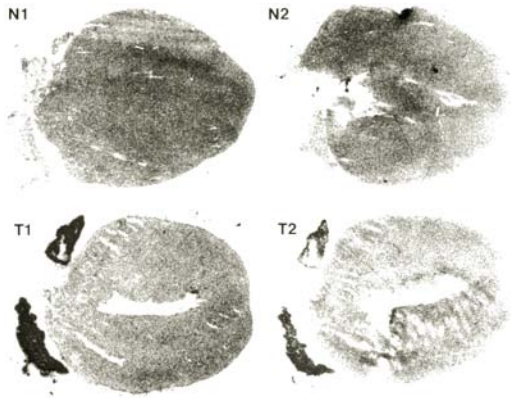


Figure 55 Autoradiographie sur des coupes de cœur de souris transgéniques.

Visualisation des récepteurs beta-adrénergiques au moyen de pindolol (un bêtabloquant) radioactif. On distingue sur les quatre coupes la masse ventriculaire et les deux oreillettes. N1 et N2 : cœurs normaux. T1 et T2 : cœurs transgéniques dans lesquels les récepteurs adrénergiques sont surexprimés en utilisant comme vecteur le promoteur du facteur atrial natriurétique qui est spécifique des oreillettes. La surexpression est clairement plus marquée dans les deux oreillettes.

La technique a plusieurs usages. On peut les regrouper en deux catégories. Dans un premier groupe, on souhaite essentiellement étudier un promoteur, ou plus habituellement on fait des constructions avec divers fragments du promoteur afin de localiser la séquence réellement active. Cette manière de faire est aussi celle de ceux qui souhaitent

démontrer le niveau transcriptionnel d'une régulation. Dans ces cas de figure on utilise une partie codante appelée « *reporter* » et qui est faite de la portion codante d'un gène dont le produit, la protéine, est facilement identifiable. On peut ainsi utiliser l'hormone de croissance qui se dose facilement dans la circulation avec une technique radio-immunologique, la beta-galactosidase qui donne une protéine facile à colorer, un enzyme facile à doser comme la chloramphenicol-acyl transferase, CAT, ou donnant un signal fluorescent comme la luciférase.

Dans d'autres cas le but est d'étudier les effets de la sur-expression ou de la sous-expression d'un gène donné. On utilisera alors pour la construction un promoteur fort non spécifique, comme le promoteur des métallothionéines, ou on essaiera de cibler l'expression du gène dans un tissu donné. Dans ce cas le promoteur sera celui d'une protéine qui n'existe que dans ce tissu. La partie codante pourra coder pour une protéine dont on espère qu'elle modifiera une fonction physiologique. C'est la technique utilisée pour étudier des maladies génétiques et savoir, par exemple, comment fonctionne le canal Chlore muté des mucoviscidoses, la dystrophine et ses formes mutées dans la myopathie de Duchenne ou encore une forme anormale de la myosine dans certaines cardiopathies héréditaires. On s'en sert également pour fabriquer des modèles animaux de maladies cardiaques ou artérielles en faisant se surexprimer une protéine candidate. On peut enfin bloquer des synthèses par plusieurs procédés, comme celui qui consiste à fabriquer en excès un ARN messager anti-sens qui va s'hybrider avec le messager sens présent dans la cellule et l'empêcher de synthétiser la protéine.

La technique dite du *knock-out* est basée sur un phénomène naturel, les recombinaisons homologues que les cellules utilisent pour réparer leur ADN lorsque celui-ci est endommagé. Les souris *knock-out*, ou *KO*, peuvent, en utilisant les cellules souches embryonnaires, transmettre le gène qui a été éteint par ce procédé à leurs descendants. On peut aussi contrôler le moment où, après la naissance le gène est éteint en utilisant le système *Cre-lox*. Cet ensemble technologique est d'actualité puisqu'il a valu à ses auteurs (Oliver Smithies, Mario Capecchi et

Martin Evans, tous Américains) le dernier prix Nobel (2007). Il leur a permis de développer plusieurs modèles de pathologie expérimentale dont un modèle de mucoviscidose et un modèle d'hypertension artérielle en vue de promouvoir une thérapie par réparation des gènes défectueux.

5.3.4 Transferts de gènes dans des cellules non germinales

Il est possible d'injecter dans certains tissus une construction génétique et de faire exprimer le gène ainsi injecté, *in situ*. Cette technique a initialement été rapportée dans le muscle squelettique, puis dans le cœur. Elle permet des études de fond sur la régulation de l'expression, la valeur d'un promoteur, sa sensibilité à des agents hormonaux par exemple. Elle est aussi potentiellement d'usage thérapeutique et a été envisagée dans le traitement des dystrophies musculaires.

L'un des tout premiers tissus qui a été la cible est le muscle d'origine squelettique, d'accès facile. L'injection, dans le muscle quadriceps de souris, d'une construction faite d'un gène « *reporter* » et d'un vecteur d'expression, le plasmide pSP qui possède un « *polylinker* » qui leur permet de se lier à n'importe quelle séquence et un promoteur (généralement pour la polymérase SP6, ici pour la polymérase T7) qui leur permet de transcrire en RNA la séquence insérée. En utilisant d'autres promoteurs et la beta-galactosidase comme « *reporter* », il a pu localiser cette expression, environ 10-30 % des cellules sont positives autour du point d'injection et la réponse est dose-dépendante. Le niveau d'expression reste encore élevé après 30 jours, plusieurs arguments laissent à penser que la construction reste circulaire et extra-chromosomique. Les muscles, d'origine squelettique ou cardiaque, sont apparemment les seuls tissus dans lesquels ce genre d'expression est possible à un niveau important, peut-être à cause de leur structure ou à cause de leur capacité à se régénérer rapidement. Les mêmes essais se sont avérés pratiquement négatifs dans d'autres tissus, comme le foie par exemple.

5.3.5 Transferts géniques à visée thérapeutique

Il y a deux manières d'envisager la thérapie génique. On peut chercher à remplacer un gène défectueux par un gène normal, ce type d'essai en est pour l'instant essentiellement au stade expérimental. La thérapie par les gènes est par contre plus avancée, elle en est déjà aux premiers essais chez l'homme et consiste à cibler dans un tissu malade l'expression d'un gène destructeur, ou celle d'un gène exprimant, par exemple, le peptide manquant, et utilise comme vecteur les lymphocytes, un rétrovirus, un adénovirus (si la cellule ne se divise pas voir une suspension de liposomes, sur des cellules autosomales uniquement bien entendu.

Le premier essai de thérapie génique entrepris chez l'homme est celui de l'équipe de Rosenberg, pour traiter le mélanome qui est une forme de cancer redoutable – ce qui justifie les essais thérapeutiques à risque. Rosenberg a eu l'idée d'utiliser un vecteur naturel déjà ciblé sur la tumeur : les « *Tumor Infiltrating Lymphocytes* », *TIL*, c'est-à-dire des lymphocytes particuliers, prélevés dans la tumeur elle-même. Ces lymphocytes sont cultivés et réinjectés avec de l'Interleukine 2 qui stimule l'immunité, cette immunothérapie d'un type nouveau a entraîné dans les premiers essais une réduction significative du nombre de métastases.

Le premier traitement d'une maladie génétique humaine a été entrepris avec succès par F Anderson et M Blaese. Il a porté sur une maladie immunitaire gravissime due à une déficience enzymatique en adenosine désaminase. Le traitement substitutif a utilisé comme vecteur des lymphocytes normaux dans lesquels on a introduit le gène codant pour la protéine déficitaire. Le résultat clinique a été spectaculaire, mais il faut bien entendu répéter ces injections tous les 4-5 mois car la durée de vie des lymphocytes est limitée.

Le prix Nobel de médecine a été attribué en 1985 à Brown et Goldstein qui ont découvert la cause d'une forme familiale grave d'hypercholestérolémie et en même temps les récepteurs des lipoprotéines (Brown 1986). Il existe une maladie génétique dans laquelle ces récepteurs sont anormaux et incapables d'assurer une épuration normale du cholestérol. Ces malades font des infarctus du myocarde graves à l'âge de 20 ans du fait de leur hypercholestérolémie. La thérapie a

été facilitée ici par l'existence d'un modèle animal, le lapin dit Watanabe, qui possède la même déficience en récepteur que les humains. Il a d'abord été montré que des lapins Watanabe chez qui on a ciblé par technologie transgénique l'expression hépatique de récepteurs normaux avaient un cholestérol plus bas. Plus récemment on est passé à l'homme. J Wilson a en effet été capable de faire baisser le taux de cholestérol d'une Américaine ayant une hypercholestérolémie sévère en lui ré-injectant le gène correspondant et en utilisant comme vecteur les propres cellules hépatiques de la patiente, après culture *in vitro*.

Des essais sont également à l'étude pour traiter la mucoviscidose, d'autres formes de cancer, d'autres enfin utilisent des séquences nucléotidiques antisens ou les microARNs. La thérapie génique est-elle la thérapie du futur, l'ADN deviendra-t-il un médicament comme les autres, personne ne peut réellement répondre à cette question.

Chapitre **6**

Quelques applications

6.1 GÉNÉTIQUE MÉDICALE

6.1.1 Hérédité

Les attributs essentiels du gène ont été définis il y a plus de 100 ans, avant la découverte du gène et de l'ADN par Gregor Mendel (fig. 56) dont on peut résumer l'œuvre sous forme de deux lois. La première loi de Mendel décrit la ségrégation des allèles « les allèles n'ont aucun effet permanent l'un sur l'autre lorsqu'ils sont présents sur la même plante (Mendel était botaniste et a décrit ses lois sur des petits pois), mais ils ségrèguent inchangés lorsqu'ils passent dans des gamètes différents » (Lewin 1987). Quand les deux allèles sont identiques, l'organisme est dit homozygote et le phénotype est l'exact reflet du génotype. Lorsque les allèles sont différents on a à faire à une hétérozygote, le phénotype sera le reflet de l'allèle dominant, l'autre allèle sera dit récessif (fig. 57).

Figure 56 Statut de Gregor Mendel dans l'abbaye de Brno. (Document de l'auteur).

La seconde loi de Mendel résume l'assortiment indépendant de gènes différents. Elle a trait au croisement d'un homozygote dominant pour deux caractères différents avec un homozygote récessif pour ces deux mêmes caractères (fig. 58).

Un caractère génétique est un phénotype, ce n'est pas un gène. Il est dominant s'il se manifeste chez l'hétérozygote, il est récessif s'il ne se

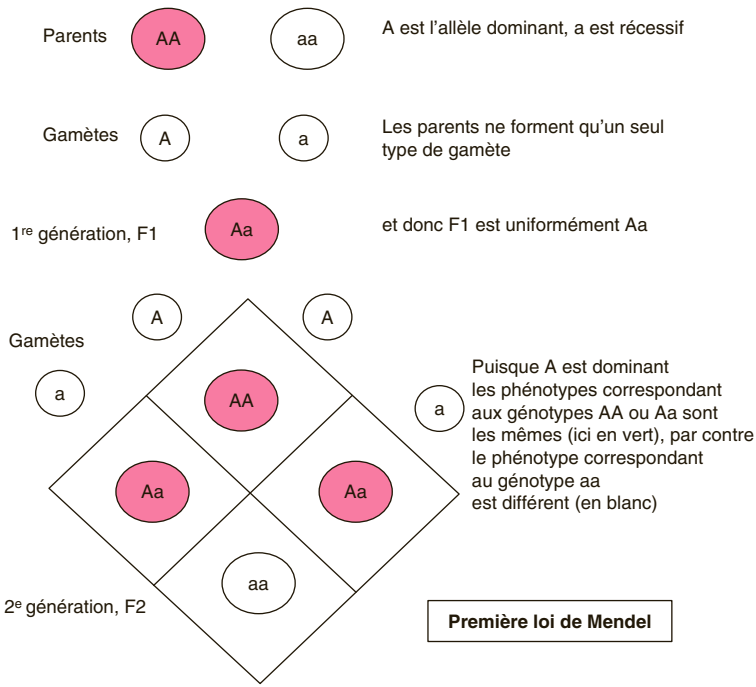


Figure 57 Première loi sur l'héritabilité de Mendel.

manifeste pas chez les hétérozygotes. On dit par ailleurs qu'un caractère peut être autosomal c'est-à-dire lié aux chromosomes non sexuels ou au contraire lié aux chromosomes sexuels, et dans ce dernier cas il faut distinguer les caractères liés au chromosome X qui peuvent être dominants ou récessifs, et les caractères liés au chromosome Y qui est unique (sauf rarissimes exceptions) (fig. 59). Dans ce dernier cas, les problèmes de dominance ou de récessivité ne se posent pas. Il y a donc cinq types d'hérédité mendélienne : hérédité autosomale dominante ou récessive, ou liée au chromosome X dominante (les deux sexes sont

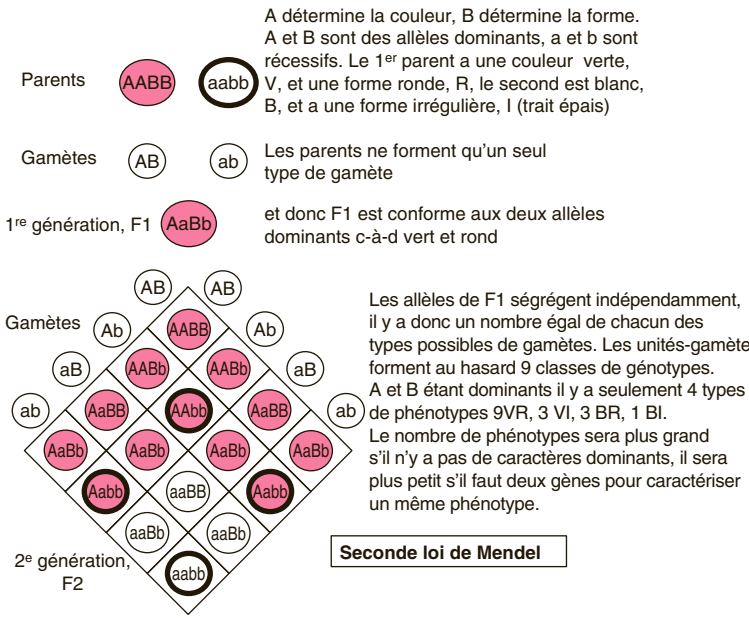


Figure 58 Seconde loi de Mendel.

atteints, toutes les filles d'un homme atteint sont atteintes, mais aucun des garçons, l'enfant d'une femme atteinte a une chance sur deux d'être atteint quelque soit le sexe) ou récessive (atteint les garçons, les mères sont porteuses asymptomatiques) et hérédité liée au chromosome Y (seuls les hommes sont atteints). L'hérédité des mutations mitochondriales est maternelle, elle n'est pas mendélienne.

La liste des 5 000 caractères mendéliens recensés chez l'homme est régulièrement mise à jour. Le catalogue peut être consulté sur Internet. Cette banque de données est un outil routinier pour la recherche en génétique, chaque caractère est identifiable sous forme d'un nombre MIM à six chiffres (par exemple, la chorée de Huntington est MIM 143100).

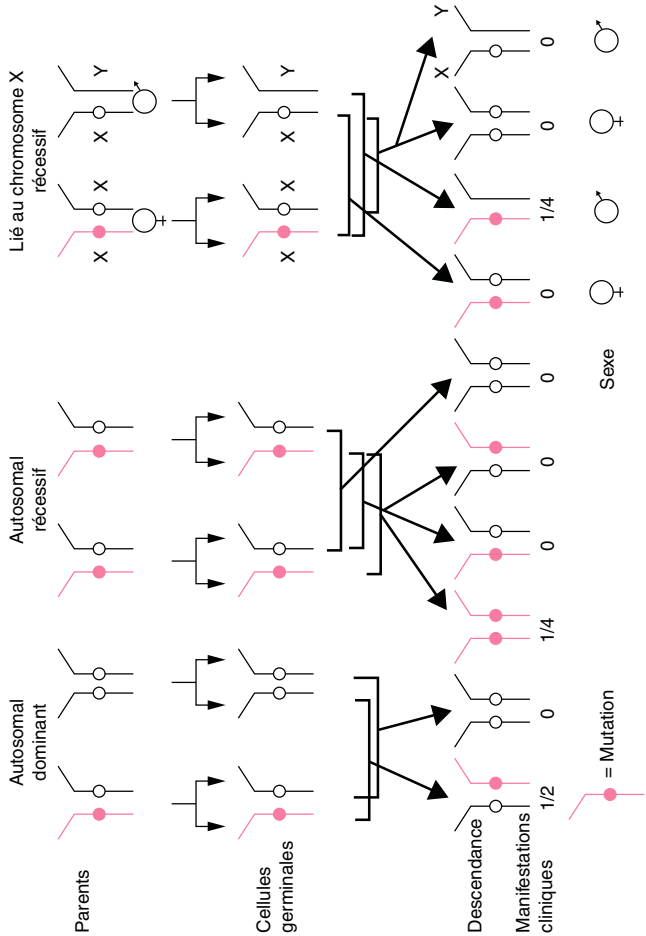


Figure 59 Modes d'héritabilité.
3 types principaux d'héritabilité mendélienne.

Les caractères récessifs se retrouvent chez les hétérozygotes (un point noir et un point blanc) qu'il s'agisse de caractères autosomaux ou de caractères liés au chromosome X, les caractères dominants se retrouvent chez les homozygotes.

6.1.2 Détecter et mesurer le polymorphisme (liaison génétique)

La génétique est une science dont l'un des buts essentiels est de trouver les anomalies génétiques à l'origine des maladies héréditaires. Les maladies génétiques au cours desquels on a mis en évidence une protéine anormale sont évidemment d'étude plus facile puisqu'on peut remonter facilement de la protéine au gène, mais elles sont relativement rares.

Liaison génétique. Dans la majorité des cas on ignore la protéine en cause, ce qui fait que l'essentiel de la recherche en génétique utilise ce que l'on appelle la génétique inverse et le principe de la liaison génétique. La figure 60 illustre ce principe. Il y a au cours de la méiose un mélange d'informations génétique qui est une source majeure de polymorphisme au sens large du terme (voir § 4.1). On admet sur cette figure que le gène en cause est M, et qu'il est la version mutée de m présent sur l'autre chromosome de la paire. M et m sont inconnus, donc non identifiables par une sonde. On va s'efforcer de trouver un gène, ou une séquence connue non codante, appelés ici P et p, qui va se transmettre (le généticien dira « ségrégé ») en même temps que M et m. Il est évident que du fait du « *cross-over* », les chances, au sens statistique du terme, de voir P et M co-ségrégé seront d'autant plus grande que la distance qui les sépare (distance génétique) est plus petite. On voit sur la figure que lorsque P et M sont loin l'un de l'autre, P peut rapidement perdre son caractère informatif. Dans cet exemple, les allèles du marqueur utilisé (P et p sur la figure 60) ne sont pas eux-mêmes en cause dans la genèse de la maladie.

Les unités de mesure de la liaison génétique sont des unités statistiques qui ont des équivalences, ou plutôt des approximations physiques. On voit bien sur la figure 60 que la distance qui sépare physiquement par exemple les loci p et m sur le second chromatide du chromosome paternel est plus grande à gauche qu'à droite. Cette distance s'exprime en pb (paires de base, « *bp* » en anglais). On voit également sur cette même figure qu'au cours du « cross-over » il y a beaucoup plus de chances que le recombinant final contienne à la fois

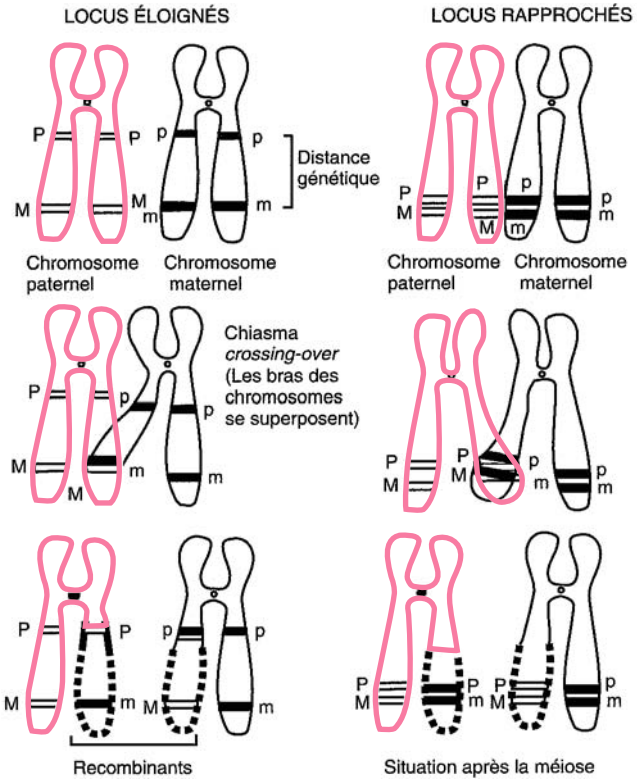


Figure 60 « Cross-over » au cours de la méiose.

Exemples de ce qui se passe pour deux locus éloignés ou pour deux locus rapprochés. De haut en bas : situation avant, pendant et après le « *cross-over* ». P et p : locus marqueur Polymorphe. M et m : locus Morbide. Il est évident que des locus très éloignés auront plus de chances de se retrouver séparés après le « *cross-over* » que des locus rapprochés, dont le caractère informatif sera donc meilleur. Cet éloignement est la distance génétique. Le locus morbide M est lié au marqueur P, mais la détection de P aura beaucoup plus de chances de renseigner sur l'emplacement du locus morbide M si les locus sont rapprochés que s'ils sont éloignés. Le recombinant final (en bas) contient à la fois P et M en bas et à droite, il ne les contient pas en bas à gauche.

p et m dans le couple de chromosomes situé à gauche, là où les loci sont rapprochés qu'à droite. L'unité génétique de mesure qui est le centiMorgan, cM, représente non pas le nombre de bp qui sépare p de m, mais le risque de survenue d'une recombinaison entre ces deux loci. 1 cM veut dire que ce risque est de 1 %. On trouvera par exemple que p et m sont séparés de 10 cM à gauche et de 1 cM à droite. L'équivalence entre bp et cM n'est pas absolue et peut varier selon le type de chromosome. Le génome humain mesure environ 3 milliards de bp et 3 300 cM, en moyenne un cM mesure donc 1 000 000 bp (1 000 kb).

Fraction ou fréquence (en %) de recombinaison q et lod score. La fréquence de recombinaison q est égale au nombre de recombinants divisé par le nombre total de méioses examinées. La fréquence avec laquelle deux loci différents peuvent se recombiner varie entre 0 et 50 %, 0 représente l'existence d'une liaison absolue, le marqueur est dans le gène morbide et la distance entre le marqueur et le locus morbide est de 0 cM; au contraire 50 % représente le cas où il n'y a pas de recombinaison possible, comme par exemple lorsque le marqueur est situé sur un chromosome différent. Il est évident que dans ce cas la liaison est impossible, la distance entre le marqueur et le locus est > 50 cM, la ségrégation des deux loci se fera selon les lois de Mendel à la fréquence 25 % PM; 25 % pm; 25 % Pm; 25 % pM, soit 50 % des cas où il y aura recombinaison. q s'exprime en cM.

L'analyse d'une liaison génétique doit s'effectuer sur un grand nombre de méioses, c'est-à-dire sur une famille la plus complète possible (trois générations est un minimum), en possédant pour le plus grand nombre possible d'individus d'une part les paramètres phénotypiques (les signes cliniques d'une maladie génétique ou… les yeux bleus; transposé à la figure 60 on dira, s'il y a maladie, phénotypiquement parlant, que l'on a à faire à M, et s'il n'y a pas maladie qu'il s'agit de m, on ne présuppose pas de cette manière de la nature du ou des gènes morbides, on se borne à étudier la liaison entre le marqueur et la maladie), d'autre part les données fournies par l'analyse génotypique, c'est-à-dire l'existence ou l'absence d'un

des deux éléments d'une paire d'allélique utilisés comme marqueur qu'il s'agisse d'un polymorphisme ponctuel ou d'un microsatellite (P ou p sur la figure 60).

On émet alors successivement un certain nombre d'hypothèses, à savoir que l'un des deux allèles-marqueurs a une chance sur 100 (theta = 0,01 ou 1 cM),…, 10 chances sur 100 (theta = 0,10 ou 10 cM),…, 50 chances sur 100 (le maximum, theta = 0,50) d'être lié au locus morbide, ce qui définit la distance génétique. On examine alors les données fournies par l'enquête génétique et l'analyse du génome et on calcule la vraisemblance de cette hypothèse. Ce calcul assez complexe est entièrement informatisé. Il fait appel au rapport de vraisemblance (ou de probabilité) qui traduit les chances qu'a une telle liaison d'exister. Ce rapport s'exprime en logarithmes décimaux et s'appelle « *lod score* », Z, il est égal au rapport :

$$Z = \frac{\text{vraisemblance de la liaison pour une valeur donnée de } \theta}{\text{vraisemblance d'} \text{ une absence de liaison}}$$

Il existe toujours une valeur de theta pour laquelle le lod score (Z) est maxima. Z est égal ou supérieur à = 3 indique qu'il y a 1 000 chances contre une pour que la liaison ne soit pas due au hasard. Si theta est égal à zéro, le « *lod score* » sera égal à l'infini puisque le log décimal de zéro est l'infini, ceci ne fait qu'exprimer le fait que lorsque θ est égal à zéro, l'hypothèse est que le marqueur est le locus morbide lui-même. On peut avec cette technique approcher le locus morbide, mais aussi exclure des gènes candidats en démontrant que le « *lod score* » entre locus morbide et un marqueur connu d'un gène candidat est inférieur à −2 (moins d'une chance sur 100 d'être liés).

Les résultats de cette analyse peuvent différer dans deux familles différentes, soit parce qu'il y a deux loci morbides différents à l'origine de la maladie, soit parce que les informations dans les deux familles sont moins complètes dans l'une et que l'informativité de l'analyse est moins bonne dans une des familles comparée à l'autre. Le « *lod score* » donne en effet des indications de probabilité.

6.1.3 Maladies génétiques

Les mutations susceptibles de modifier une fonction, et qui sont à l'origine des maladies génétiques, peuvent survenir au niveau de la séquence codante d'un gène, sur un exon[1], à celui des séquences régulatrices du gène, à celui des séquences non codantes intragéniques au niveau des introns (mutations d'épissage dans les maladies du collagène), au niveau de l'ADN mitochondrial[2], enfin à celui des gènes codant pour des ARNs régulateurs comme les microARNs, bien que cet aspect de la génétique en soit encore à ses débuts.

a) Maladies monogéniques

➤ Drépanocytose

C'est une maladie africaine dont l'incidence dans les populations noires est de 1 pour 500 naissances. Elle se transmet selon un mode autosomal récessif. Cette maladie est grave chez les homozygotes et l'issue en est souvent fatale, la mort survenant par anémie et thrombose. La mutation qui est à son origine est unique (c'est une maladie monogénique unicentrique monoallélique). Elle est située sur le codon 6 du premier des 3 exons du gène codant pour la beta-globine. Elle aboutit à substituer à l'acide glutamique normalement codé à cet endroit, une valine (fig. 61). Les conséquences de cette mutation extrêmement ponctuelle sont consi-

1. Il peut s'agir de substitutions non synonymes, généralement au niveau d'une des deux premières bases du codon. L'hypothèse du « *wobble* » admet en effet une certaine flexibilité dans l'appariement au niveau de la troisième base (à ce niveau les appariements G-U peuvent être admis). Il existe pour nombre de gènes des *points chauds*, c'est-à-dire des emplacements au sein de la séquence où la fréquence des mutations est plus élevée qu'ailleurs (l'acide aminé 403 sur la molécule de myosine cardiaque dans les cardiomyopathies hypertrophiques familiales, par exemple).
2. La génétique de l'ADN mitochondrial est différente du fait de la multiplicité de ces particules et de l'hérédité maternelle. Au cours de la division cellulaire les mitochondries se répartissent au hasard et de ce fait, dans une même cellule, peuvent coexister des mitochondries normales et des mitochondries possédant la mutation (hétéroplasmie).

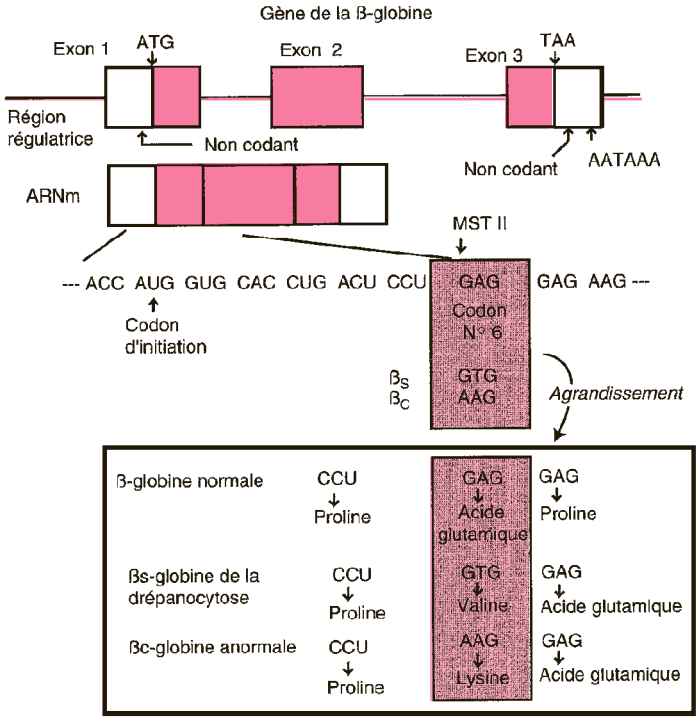


Figure 61 Drépanocytose.

Mutations faux-sens du gène de la beta-globine de l'hémoglobine au niveau du Codon 6, sur le premier exon. De haut en bas : le gène de la β-globine et son ARNm, la structure du début du premier exon montrant le lieu des mutations, et enfin les acides aminés correspondants. La mutation caractéristique de la drépanocytose beta^s remplace une charge négative, l'acide glutamique, par une charge neutre, la valine, la proteine va donc changer de charge électrique et pourra être distinguée de la protéine normale par électrophorése. Le diagnostic peut également se faire par une carte de restriction : les codons 5, 6 et 7 forment en effet un site de restriction CCTNAGG (N veut dire que le nucléotide peut être n'importe quel nucléotide) sensible à l'enzyme Mst II. La mutation supprime ce site. On a également figuré une seconde mutation pathogène, beta^c celle qui aboutit à l'hémoglobine H^c. Le codon 6 qui est le siège privilégié de plusieurs mutations est appelé point chaud, « *hot spot* ».

dérables puisque l'hémoglobine mutée précipite lorsqu'elle n'est pas oxygénée, ce qui donne aux hématies un aspect en forme de faux, d'où le nom d'anémie *falciparum*. Ce phénomène est réversible à la ré-oxygénation pendant un certain temps, puis il devient progressivement irréversible et l'hémoglobine précipite dans les petits vaisseaux et va causer thromboses et dégâts cellulaires par ischémie, ce qui entraînera l'issue fatale. Mais la mutation a la particularité d'avoir également des effets bénéfiques, elle protège ceux qui en sont porteurs contre le paludisme, et de fait la mutation confère un avantage sur le plan évolutionniste et sa carte de diffusion se superpose à celle du paludisme, lequel est essentiellement Africain. La mutation sur l'hémoglobine est très certainement née en Afrique, par hasard, mais elle s'est maintenue sur ce continent du fait de l'avantage sélectif qu'elle donne à ses porteurs. La durée de vie dans l'Afrique sub-Saharienne est de l'ordre de 40 ans, mais un Africain a plus de chances de mourir de paludisme dans sa jeunesse que de mourir d'anémie.

➤ Myopathie de Duchenne

C'est une myopathie dégénérative grave qui aboutit à la mort par dégénérescence des muscles respiratoires vers l'âge de 20 ans. Elle ne touche que les garçons. Une très longue marche d'approche a été nécessaire pour découvrir que la cause de cette maladie était située au sein d'une protéine normale, la dystrophine, jusque maintenant inconnue et qui fut découverte à cette occasion. Les mêmes signes cliniques sont observés dans des cas où les mutations sont différentes, mais toujours situées au sein de la dystrophine. Sur ce point la myopathie de Duchenne est différente de la drépanocytose, il n'y a pas effet fondateur, la maladie est certes monogénique mais elle est pluricentrique et multiallèlique. Fait important la longueur des délétions n'est pas proportionnelle à la gravité des symptômes.

➤ Mucoviscidose

C'est une affection fréquente dans les populations Européennes où elle touche un nouveau-né sur 2 500. Elle est caractérisée par l'épaississement des sécrétions pancréatiques et bronchiques, avec augmentation de la concentration en Chlore de la sueur. Ces enfants ont une espé-

rance de vie autour de 25 ans et meurent d'infection pulmonaire. La protéine en cause est le canal au Chlore (ou un facteur en contrôlant l'activité), et là aussi au même tableau clinique correspond de très nombreuses mutations (avec peut-être certaines nuances pronostiques). La mucoviscidose est une des maladies génétiques pour lesquelles on envisage dans un avenir pas trop lointain une thérapie génique.

➤ Cardiomyopathie Hypertrophique Familiale

La cardiomyopathie hypertrophique est familiale dans un cas sur deux. La forme familiale est habituellement autosomale et évolue sur un mode dominant. L'incidence de l'affection est grade (2,5 pour 100 000 habitants). La pénétrance est incomplète chez les sujets jeunes et augmente avec l'âge. La maladie est grave de pronostic sévère, c'est le grand responsable de la mort subite chez le sujet jeune. C'est une maladie monogénique, mais plusieurs gènes sur des chromosomes différents (on dit qu'il y a hétérogénéité non allélique) et plusieurs types de mutation sur le même gène (hétérogénéité allélique) sont en cause, on en a recensé prêt d'une cinquantaine. Ces mutations portent toutes sur les protéines contractiles (myosine, protéine C, troponine, tropomyosine), elles donnent toutes le même tableau clinique, l'hypertrophie myocardique compense probablement le déficit contractile.

La fonction de la myosine anormale a pu être étudiée de deux manières : (i) la chaîne lourde bêta de la myosine existe également dans le muscle d'origine squelettique où elle est accessible à la biopsie, on a pu de cette façon étudier les conséquences physiologiques de la mutation et montrer que les molécules anormales avaient *in vitro* une contractilité anormale; (ii) mais la manière la plus habituelle de démontrer qu'une mutation est bien à l'origine de l'anomalie fonctionnelle consiste à incorporer le gène anormal dans un organisme par transgenèse, ou dans une cellule et d'en étudier les conséquences. Les souris chez qui on a fait exprimer le gène muté ont une hypertrophie cardiaque comparable à celle observée chez l'homme.

Cette histoire récente exemplaire souligne plusieurs des difficultés de la démonstration génétique : la même maladie peut répondre à plu-

sieurs mutations situées soit sur le même gène, soit sur des gènes voire des chromosomes différents. Le lien entre la symptomatologie clinique et l'anomalie protéique est loin d'être évident et la manière dont une myosine anormale peut créer une hypertrophie cardiaque est encore du domaine de l'hypothèse. Les territoires à explorer sont gigantesques (15 cMorgan) ; la thérapie génique par substitution d'une maladie aussi hétérogène n'est pas pour demain.

b) Maladies polygéniques

Il s'agit de maladies dues à des anomalies géniques dispersées, chaque allèle étant nécessaire, mais pas suffisant (Wright 2007). Ces maladies comportent toutes un facteur environnemental, nutritionnel, toxique, qui, dans certains cas joue un rôle déterminant et révélateur dans l'apparition de la maladie (un fumeur fils d'hypertendu a beaucoup plus de chance de faire un accident cardiaque qu'un non-fumeur également fils d'hypertendu mais il n'est pas exclu que le tabagisme soit également en partie lié à un facteur génétique) (fig. 37).

L'approche de ces grands secteurs de la pathologie, qui couvrent les maladies cardiovasculaires, le diabète, le cancer et les maladies mentales (ce qui, à nouveau, ne veut pas dire et de loin que tous les cancers ou toutes les maladies mentales soient d'origine génétique par exemple) présuppose d'étroites collaborations entre généticiens épidémiologistes, cliniciens, et même économistes et… financiers. Elle peut être basée sur des hypothèses, celles de gènes candidats, par exemple ceux qui codent pour les récepteurs des lipoprotéines pour les hyperlipidémies., et sur la recherche d'une liaison génétique entre ce gène candidat et un ou plusieurs marqueurs.

6.2 PHARMACOGÉNÉTIQUE ET PHARMACOGÉNOMIQUE

La pharmacogénétique a trait aux variations génétiquement déterminées de la sensibilité des individus à un médicament, elle est issue d'un certain nombre d'observations faites dans les années 50. Ces observa-

tions ont démontré le caractère héréditaire d'un certain nombre d'accidents thérapeutiques : l'hémolyse sous antipaludéens et ses relations avec les anomalies héréditaires en G6PDH érythrocytaire par exemple. Historiquement, la discipline a pris consistance en 1980 à propos de la débrisoquine qui est un hypotenseur ayant un nombre inhabituel d'effets secondaires. Ce composé est métabolisé par une hydroxylase qui appartient à la famille des CYP2D6. Il existe des patients, les plus nombreux, qui métabolisent rapidement ce médicament, dits métaboliseurs rapides, des patients métaboliseurs ultrarapides et des patients, les plus rares, métaboliseurs lents. Ce trait est transmis sous une forme autosomale récessive. Il n'est pas nécessaire pour identifier ces différentes classes de patients d'analyser leur génotype, on peut assez facilement savoir à quel groupe ils appartiennent en dosant dans leurs urines un des métabolites de la débrisoquine, et selon la concentration savoir si le sujet est ou non un bon métaboliseur de ce médicament. On sait depuis qu'à chaque classe de patient correspond un allèle spécifique du gène CYP2D6 qui a été identifié.

Pharmacogénomique a une signification plus large et inclut à la fois la pharmacogénétique et le champ immense des cibles des médicaments en y incluant les aspects multigéniques des grands cadres morbides comme l'athérosclérose, l'écogénétique et la toxicogénétique, c'est-à-dire les aspects génétiques des accidents médicamenteux. La pharmacogénomique est fille du Programme Génome Humain. La connaissance de l'ensemble des gènes et l'arrivée de techniques qui permettent d'explorer l'ensemble de l'expression génétique, ont fourni au pharmacologue (comme au toxicologue) des outils permettant l'étude de l'ensemble des gènes et de leur expression contrôlant à la fois l'efficacité, la sensibilité et la recherche thérapeutique.

Il faut d'abord distinguer trois aspects différents à la fois sur un plan conceptuel et sur un plan stratégique.

(**1**) La sensibilité au traitement est un trait multigénique. Il peut s'agir de polymorphismes géniques modifiant le transport ou les mécanismes d'inactivation des médicaments, certains de ces allèles peuvent bloquer l'élimination des médicaments, lesquels s'accumulent et devien-

nent toxiques, d'autres au contraire sont plus actifs et accélèrent le catabolisme des médicaments en métabolites inactifs, ce qui va obliger à augmenter la posologie courante. L'exemple le plus connu est le CYP2D6 lorsqu'il active les opioïdes en morphine par O-déméthylation. Le variant non fonctionnel du *CYP2D6* explique que 2-10 % de la population soit très peu sensible au traitement. La codéine perd ses effets respiratoires, psychomoteurs et pupillaires chez les sujets porteurs de ce variant. On a pu démontrer que chez ces sujets on ne retrouvait pas les métabolites habituels de la codéine, métabolites (dont la morphine) qui sont en fait la forme pharmacologique active du médicament.

Mais il peut au contraire s'agir de modifications dans la structure même des gènes qui codent pour les cibles du traitement, que ces cibles soient celles dont les effets sont souhaités ou au contraire les cibles responsables des effets secondaires délétères. Il est, par exemple, facile

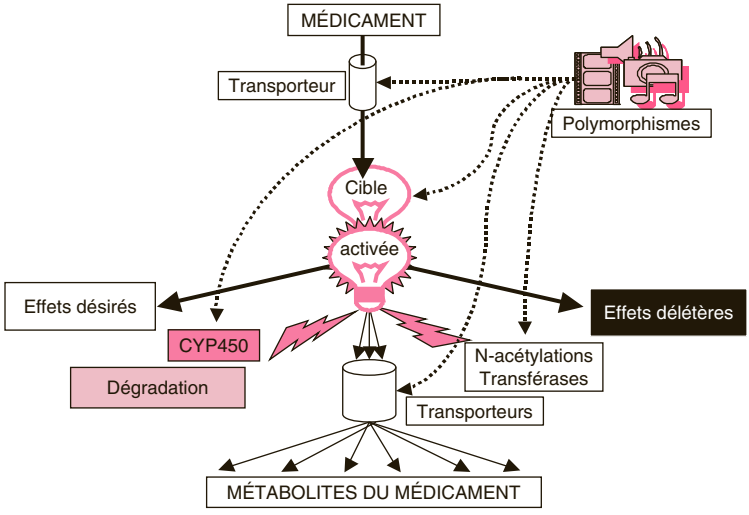


Figure 62 Le métabolisme et l'efficacité d'un médicament sont des traits multigéniques.

d'imaginer qu'une mutation qui modifierait la structure du récepteur bêta-adrénergique puisse modifier l'affinité des bêta-bloqueurs et donc leur efficacité, ce pour les effets bénéfiques.

(**2**) Les polymorphismes, disons de voisinage, sont un tout autre problème. La cible elle-même reste à définir, et, en finale, l'ambiguïté des résultats concernant l'effet thérapeutique ne fait que refléter la faiblesse de ceux qui concernent la physiopathologie.

(**3**) L'utilisation par l'industrie pharmaceutique des données fournies par le Programme Génome et par la pharmacogénomique est un troisième problème, souvent mélangé aux deux premiers. C'est un problème aux incidences industrielles et économiques considérables. L'innovation thérapeutique se doit de tenir compte à la fois de deux facteurs différents : (a) la connaissance de la structure des cibles déterminantes d'une affection fournit bien évidemment de nouvelles cibles à la recherche de médicaments; (b) on ne peut envisager de développer un nouveau composé si l'on sait que l'enzyme responsable de son élimination existe sous des formes alléliques non fonctionnelles fréquentes et susceptibles de créer des accidents thérapeutiques.

6.3 EMPREINTES GÉNÉTIQUES EN BIOMÉTRIE

La notion d'empreintes génétiques est due à A. JEFFREY qui, en 1985, a montré l'existence dans le génome humain de séquences identiques répétées en tandem un nombre variable de fois (VNTR = « *variable number of tandem repeats* ») d'un chromosome à l'autre, donnant lieu à un polyallélisme important (6 à 10 allèles de tailles différentes en moyenne pour un même locus). En raison de la lourdeur des techniques, les VNTR ont été rapidement abandonnés au profit des microsatellites (ou STR = « *short tandem repeats* »), c'est-à-dire de séquences de 2 à 5 nucléotides, également répétées en tandem et polyalléliques, très nombreuses (1 STR tous les 40-50 Kb), réparties de façon à peu près homogène dans le génome. Ces marqueurs sont étudiés par PCR en utilisant des amorces dont les séquences sont complé-

mentaires de celles situées de part et d'autre du microsatellite; les produits d'amplification sont ensuite séparés en fonction de leur taille par une technique électrophorétique. Le marquage des amorces par un fluorochrome permet de visualiser les fragments sous l'action d'un laser. En combinant les couleurs des fluorochromes et la diversité de taille des fragments amplifiés, plusieurs microsatellites peuvent être étudiés simultanément. Actuellement des « kits » commerciaux sont disponibles pour l'étude de 11 à 16 marqueurs différents, sélectionnés et standardisés au niveau international, ce qui permet des échanges directs d'informations entre les services de police de divers pays. Ces marqueurs ont été choisis essentiellement en raison de leur informativité (leur étude simultanée conduit à un résultat dont la probabilité d'être identique entre deux individus non jumeaux homozygotes n'est de l'ordre que de 1/100 milliards) et parce qu'ils ne présentent que peu ou pas de biais de fréquence allélique d'une population à l'autre. La technologie de ces analyses emploie des automates de plus en plus sophistiqués et repose sur une informatisation des lectures grâce à des logiciels performants. Les résultats obtenus conduisent à l'établissement pour chaque individu d'un génotype, qui sera ensuite éventuellement rentré dans une banque de données.

Les principales applications des empreintes génétiques sont d'ordre judiciaire (crimes, agressions diverses, viols, terrorisme…). Les performances obtenues sont celles de la PCR qui permet d'analyser des traces infimes d'échantillons biologiques (salive, sang…) et de les comparer aux empreintes génétiques d'un suspect. Elles sont parfois encore améliorées par différents aspects techniques : UV pour caractériser des traces de sang, extraction différentielle d'ADN en fonction des types cellulaires… Bien entendu dans les laboratoires compétents toutes les manipulations s'effectuent avec des précautions extrêmes (port de gants et de masque); en outre, pour éviter toute erreur due à une contamination, les génotypes de tous les employés du laboratoire sont déterminés pour comparaison. Dans le domaine judiciaire, les génotypes sont déposés dans une banque nationale de données ou Fichier national automatisé des empreintes génétiques. Ce dépôt concerne : les empreintes de tous les condamnés (sauf pour les délits dits « en col blanc »), les empreintes

trouvées sur les lieux d'un crime et transitoirement les empreintes de suspects en vue de comparaison. C'est ainsi qu'en Grande Bretagne 3,5 millions de profils génétiques sont conservés et que 8 à 10 000 rapprochements sont effectués chaque année.

Dans notre pays, sur décision d'un magistrat, les empreintes génétiques sont également utilisées pour des tests de paternité. Par ailleurs, outre les empreintes génétiques « standard » reposant donc sur l'examen d'une dizaine de microsatellites, deux autres types d'analyse d'ADN sont pratiqués. Le premier concerne l'ADN mitochondrial lorsque les échantillons biologiques sont très dégradés; il s'agit d'un séquençage de deux régions (HV 1 et 2) de la « *D-Loop* » à la recherche de variations nucléotidiques simples; bien entendu les résultats sont beaucoup moins discriminants que le polyallélisme des microsatellites, ils présentent en outre l'inconvénient de reposer sur une transmission génétique exclusivement maternelle. Le second est la reconstitution d'haplotypes du chromosome Y par l'analyse de 11 microsatellites, cette étude étant pratiquée essentiellement en cas de viols multiples. Il est aussi possible d'identifier le sexe sur un prélèvement ADN (voir Ch. 3.3).

Le marché de la Biométrie est actuellement en plein développement et représente une source d'emplois qualifiés importante. Outre l'aspect génotypique sus-décrit, il comprend tout un volet phénotypique, moins coûteux et plus facile d'accès comme la détermination des empreintes digitales ou l'analyse de l'iris. La corrélation entre ces deux types de données est un sujet important, de grande portée pratique qui pose de nombreux problèmes éthiques.

Références

7.1 LIVRES OU TRAITÉS

7.1.1 En anglais

- Clark DP ed. *Molecular biology. Understanding the genetic revolution*. Elsevier Academic Press pub. Burlington US 2005. 784 pp.
- Griffiths, Wessler, Lewontin et al. eds *Introduction à l'analyse génétique*. De Boeck pub. Bruwelles. 4e édition 2004
- Jobling MA, Hurles ME, Tyler-Smith C. *Human evolutionary genetics. Origins, peoples and disease*. Garland Science pub. New York & Abingdon 2004. 523 pages
- Lewin B. *Genes*. John Wiley. 1987. Il y a presqu'une dizaine d'éditions du Lewin, la dernière est de 2006
- Lodish H, Scott MP, Matsudaira P et al. *Molecular cell biology*. 5th ed. Freeman pub, NwY. 2000
- Ridley M ed. *Evolution*. Blackwell Publ. Malen US/Oxford UK. 2004
- Stearns SC, Hoekstra RF. *Evolution : an introduction*. Oxford University press. 2nd ed. 2005
- Wright A, Hastie N. *Genes and common diseases*. Cambridge University Press. Cambridge UK. 2007

7.1.2 En français (original ou traduction)

- Darwin CR, *L'origine des espèces*. GF Flammarion. Paris1992 (format poche)
- Gibson G, Muse SV eds. *Précis de génomique*. De Boeck Pub. Bruxelles. 2004
- Kaplan J.-C., Delpech M., *Biologie Moléculaire et Médecine*, 3e éd., Médecine-Sciences Flammarion. Paris 2007
- Maftah A, Julien R eds. *Biologie moléculaire*. 2e édition. Dunod pub. Paris 1999
- Noble D. *La musique de la vie. La biologie au-delà du génome*. Éditions du Seuil. Paris 2007

7.2 RÉFÉRENCES CITÉES

- Blackburn EH. Telomeres and telomerase : their mechanisms of action and the effects of altering their functions. *FEBS Letters*
- Brent R. Genomic biology. *Cell 2000*, 100, 169-183
- Brown MS, Goldstein JL. A receptor-mediated pathway for cholesterol homeostasis. *Science*, 232, 34-47, 1986
- Dobzhansky T. Nothing in biology makes sense except in the light of evolution. *American Biology Teacher* 1973, 35, 125-129
- Dykxhoorn DM, Lieberman J. The silent revolution : RNA interference as basic biology, research tool, and therapeutic. *Annu Rev Med* 2005, 56, 401-423
- Feuk L, Carson AR, Scherer SW. Structural variation in the genome. *Nature Reviews* 2006, 7, 85-97
- Fire A, Xu M, Montgomery M et al. Potent and specific genetic interference by double-standed RNA in Caenorhabditis elegans. *Nature* 1998, 391, 806-811
- Gerstein MB, Bruce C, Rozowsky JS et al. What is a gene, post-ENCODE ? History and updated definition. *Genome Res* 2007, June 18, 669-681
- Gewirtz AM. On future's doorstep : RNA interference and the pharmacological of tomorrow. *J Clin Invest* 2007, 117, 3612-3613
- Gill SR, Pop M, DeBoy RT et al. Metagenomic analysis of the human distal gut microbiome. *Science* 2006, 312, 1355-1359

- International HapMap Consortium. A haplotype map of the human genome. *Nature* 2005; 437 : 1299-1320
- Hedges SB. The origin and evolution of model organisms. *Nature Reviews Genetics* 2002, 3, 838-849
- Héléne C, Saison-Behmoaras E. La stratégie anti-sens : nouvelles approches thérapeutiques. *Médecine/Sciences*. 10, 253-273, 1994
- Kell DB Metabolomics and systems biology : making sense of the soup. *Current Opinion in Microbiology* 2004, 7 : 296–307
- Khaja R, Zhang J, MacDonald JR et al. Genome assembly comparison identifies structural variants in the human genome. *Nature Genetics* 2006, 38, 1413-1418
- Koshland D. The seven pillars of life. *Science*. 2003, 295, 2215
- Kruglyak L. Power tools for human genetics. *Nature Genetics* 2005, 37, 1299-1300
- Levy S, Sutton G, Ng PC et al. The diploid genome sequence of an individual human. PLoS Biol 2007, 5(10), e254 doc10.>1371/journal.pbio.0050254
- Noll H. The digital origin of human language-a synthesis. *BioEssay* 2003, 25, 489-500
- Rich DP, Anderson MP, Gregory RJ et al. Expression of CFTR corrects defective chloride channel regulation in cystic fibrosis airway epithelial cells. *Nature*, 374, 358-363, 1990
- Rosenberg SA, Aebersold P, Cornetta K et al, Gene transfer into humans. Immunotherapy of patients with advanced melanoma, using tumor-infiltrating lymphocytes modified by retroviral gene transduction. *N. Engl. J. Med.*, 323, 570-578, 1990
- Scherrer LJ, Rossi JJ. Approaches for the sequence-specific knockdown of mRNA. *Nature Biotech* 2003, 21, 1457-1465
- Woltereck R. Weitere experimentelle untersüchungen über artveränderung, speziell über das wesen quantitativer artunterschede bei daphniden. *Verhandlungen der Deutschen Zooligischen Gesllschaft* 1909, 19, 110-172
- Wray GA, Hahn MW, Abouheif E et al. The evolution of transcriptional regulation in eucaryotes. *Mol Biol Evol* 2003, 20, 1377-1419

Addenda

Addendum 1 LES DONNÉES NOUVELLES QUI OBLIGENT À REVOIR LA DÉFINITION DU GÈNE (D'APRÈS GERSTEIN 2007)

Données concernant la localisation et la structure
Gènes situés dans l'intron d'un autre gène
Une même séquence ADN peut coder pour deux protéines différentes
La séquence régulatrice d'un gène est très éloignée de la séquence codante
Il existe des transcrits provenant du brin antisense
Données structurelles
Un des éléments d'un gène est mobile (prix Nobel McClintock)
Les modes d'épissage ne sont pas transmis
Le nombre de copies d'un gène varie d'un individu à l'autre

Addendum 1 LES DONNÉES NOUVELLES QUI OBLIGENT À REVOIR LA DÉFINITION DU GÈNE (D'APRÈS GERSTEIN 2007)

Données épigénétiques
Empreintes parentales, le phénotype n'est pas directement déterminé par le génotype
Structure de la chromatine
Modifications post-transcriptionnelles
Épissage alternatif
Deux produits d'un même épissage alternatif produisent deux protéines différentes
Trans-épissage : une protéine résulte de la combinaison d'informations provenant de transcrits différents
L'ARN est modifié enzymatiquement et l'information de l'ADN n'est plus codante
Régulation post-traductionnelles
Épissage des protéines, le début et la fin d'une protéine n'est plus déterminé par le simple code génétique
Modifications directes des protéines avec changement de la fonction initiale
Pseudogènes et rétrogènes
Transcription réverse d'un ARNm, suivie par l'incorporation de l'ADN dans le génome (rétrogène)
Transcription d'un pseudogène

Addendum 2 COMPOSITION DU GÉNOME HUMAIN (D'APRÈS CLARK 2005)

Séquences uniques
Gènes codant en y incluant les régions régulatrices, les exons et les introns
Gènes codant pour des ARN non-traduits : sn (small nuclear) RNA, de très nombreux microARNs, ARN télomérasique
Séquences ADN non répétitives intragéniques
Séquences ADN répétitives non codantes intercalées
Pseudogènes
SINEs, dont éléments Alu (~ 1 000 copies)
LINEs, entre 200 et 800 pb (~ 750 000 copies)
Séquences rétrovirales, entre 500 et 1 300 pb (~ 250 000 copies)
Transposons (Séquences ADN mobiles), en moyenne 250 pb (~ 200 000 copies)
Séquences ADN répétitives en tandem
Gènes des ARN ribosomaux, environ 50 clusters répartis sur 5 chromosomes
Gènes des RNA transferts
Séquences télomèriques de 6 pb
Minisatellites (VNTRs) en blocs de 0,1 à 20 kpb, près du centromère
Séquences centromèriques de 171 pb, liées aux protéines du centromère
ADN satellite en blocs de 100 kpb, près du centromère
ADN megasatellitaire en blocs de 100 kpb ou plus

Addendum 3 Principales dates critiques concernant l'histoire de la vie sur Terre

Date, en milliard d'années	Période	Événement
4,6 – 2,5	**ARCHÉEN**	Origines de la vie
2,5 – 0,54	**PROTÉROZOÏQUE**	Oxygénation graduelle de l'atmosphère
2,5 – 1,6	Paléoprotérozoïque	Origine des eucaryotes et du sexe
1,6 – 1,0	Mésoprotérozoïque	Radiation des protistes
1,0 – 0,54	Néoprotérozoïque	Premiers multicellulaires
0,54 – 0	**PHANÉROZOÏQUE** (au moins 20 extinctions massives dont 3 majeures)	
0,54 – 0,25	***Paléozoïque***	
0,54 – 0,50	Cambrien	Apparition massive de la plupart des phyllae Mollusques, Annélides, Arthropodes, Brachiopodes, Échinodermes
0,50 – 0,43	Ordovidien (extinction de masse [22 %], glaciations)	Poissons osseux, plantes à spores, ammonites Invasion de la terre
0,43 – 0,41	Silurien	Champignons, scorpions, plantes vasculaires
0,41 – 0,35	Dévorien	Origine des membres, radiation des poissons, premières forêts, feuilles et bois
0,35 – 0,29	Carbonifère	Charbon. Tétrapodes et Amniotes (Amphibiens, Mammifères et Diapsides), Fleurs, radiation des Insectes

Addendum 3 PRINCIPALES DATES CRITIQUES CONCERNANT L'HISTOIRE DE LA VIE SUR TERRE

0,29 – 0,25	Permien (extinction de masse [83 %])	Reptiles Coévolution avec le splantes à pollen
0,25 – 0,065	***Mésozoïque***	
0,25–0,203	Triassique (extinction marine [20 %])	Gymnospermes, Téléostes, Crocodiles, Dinosaures, premiers Mammifères
0,203 –0,135	Jurassique	Oiseaux, Dinosaures, Tortues, radiation des Reptiles, Homéothermes
0,135 – 0,065	Crétacé 3(extinction de masse)	Angiospermes, Lézards, disparition des Dinosaures et Ammonites
0,065 – 0	***Cénozoïque***	
	Paléogène	
0,065 – 0,053	Paléocène	Réapparition des Mammifères, Oiseaux et Plantes actuels. Moustiques, requins, herbe, rongeurs
0,053 –0,034	Éocène	Primates, Ongulés, Chiens, Chauve-souris, Hêtres
0,034 – 0,023	Oligocène	Prairie, radiation des Mammifères
	Néogène	
0,023 – 0,0053	Miocène	Ruminants et carnivores
0,053 – 0,0016	Pliocène	Hominidés, bipédalisme (4 Ma), Orchidées
	Quaternaire	
1,75 – 0,01	Pléistocène	Homo erectus (1,5 Ma), Sapiens (0,3 Ma), langage (0,06 Ma), art, funérailles
0,01 – 0	Holocène	Agriculture, écriture, domestication

Index

B

C

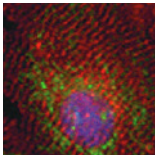
M

N

O

SCIENCES SUP

Série Aide-mémoire



Bernard Swynghedauw

avec la collaboration de Jean-Sébastien Silvestre

BIOLOGIE ET GÉNÉTIQUE MOLÉCULAIRES

Cette 3e édition en 2 couleurs de l'*Aide-mémoire de biologie et génétique moléculaires* est destinée aux étudiants en L1 ou L2 de Sciences de la vie ainsi qu'aux étudiants de PCEM1 et PH1.

En deux parties de même importance – **biologie moléculaire** et **génétique** –, ce livre rend compte des dernières avancées : parution de la séquence complète du génome humain ; développement de la génomique fonctionnelle, avec d'abord les techniques de protéomique, puis les techniques plus physiologiques ; également diffusion des méthodes dites de « microarrays » qui permettent de pratiquer des analyses globales de l'expression génique de tissus ; apparition d'une biologie intégrée dont l'outil essentiel est le transfert génique ; enfin développement de la génétique moléculaire vers la génétique des populations ou la génétique évolutionniste.

BERNARD SWYNGHEDAUW

est directeur de recherche émérite à l'INSERM (hôpital Lariboisière)

MATHÉMATIQUES

PHYSIQUE

CHIMIE

SCIENCES DE L'INGÉNIEUR

INFORMATIQUE

SCIENCES DE LA VIE

SCIENCES DE LA TERRE

ISBN 978-2-10-053798-3

www.dunod.com

DUNOD