INITIATION À LA BIOLOGIE MOLÉCULAIRE

Jean-Bernard et Laurence Quiot
2 avenue A. Briand - 35400 Saint-Malo
jblfr@aol.com

Sommaire

Résumé :

Depuis 40 ans, les progrès dans la connaissance du vivant ont été rendus possibles par un fort développement de la collaboration entre des disciplines différentes mais complémentaires (écologie, biologie, biochimie, biométrie, bio-informatique), par le lancement de projets internationaux et par l'espoir de voir le progrès des connaissances aboutir à des applications, en particulier en médecine humaine. Ces progrès ont été aussi rendus possibles par l'apparition et la généralisation rapide de techniques d'études extrêmement performantes permettant l'étude de la biologie au niveau moléculaire.

On peut citer parmi les résultats les plus marquants :

- la révision de la notion de gène qui n'est plus considéré comme monolithique ;

- la découverte du rôle essentiel d'innombrables petits ARN dans la régulation cellulaire ;

- la révélation de la formidable biodiversité des microorganismes.

1 – DÉFINIR LA BIOLOGIE MOLÉCULAIRE

On peut tenter de définir la biologie moléculaire (BM) comme une discipline qui est issue de la compréhension de l'information présente dans les acides nucléiques cellulaires et qui a pour but principal de décrire le rôle de cette information dans le fonctionnement intime des systèmes biologiques de toutes sortes.
Depuis trente ans et grâce aux progrès fantastiques des techniques d'étude, la biologie moléculaire a permis la diffusion progressive d'outils d'analyses moléculaires qui entrainent, pour de très nombreuses disciplines biologiques, des progrès importants dans la compréhension de leurs domaines d'étude.

Le succès de la BM a entrainé ainsi l'apparition ou le développement de nouvelles disciplines :

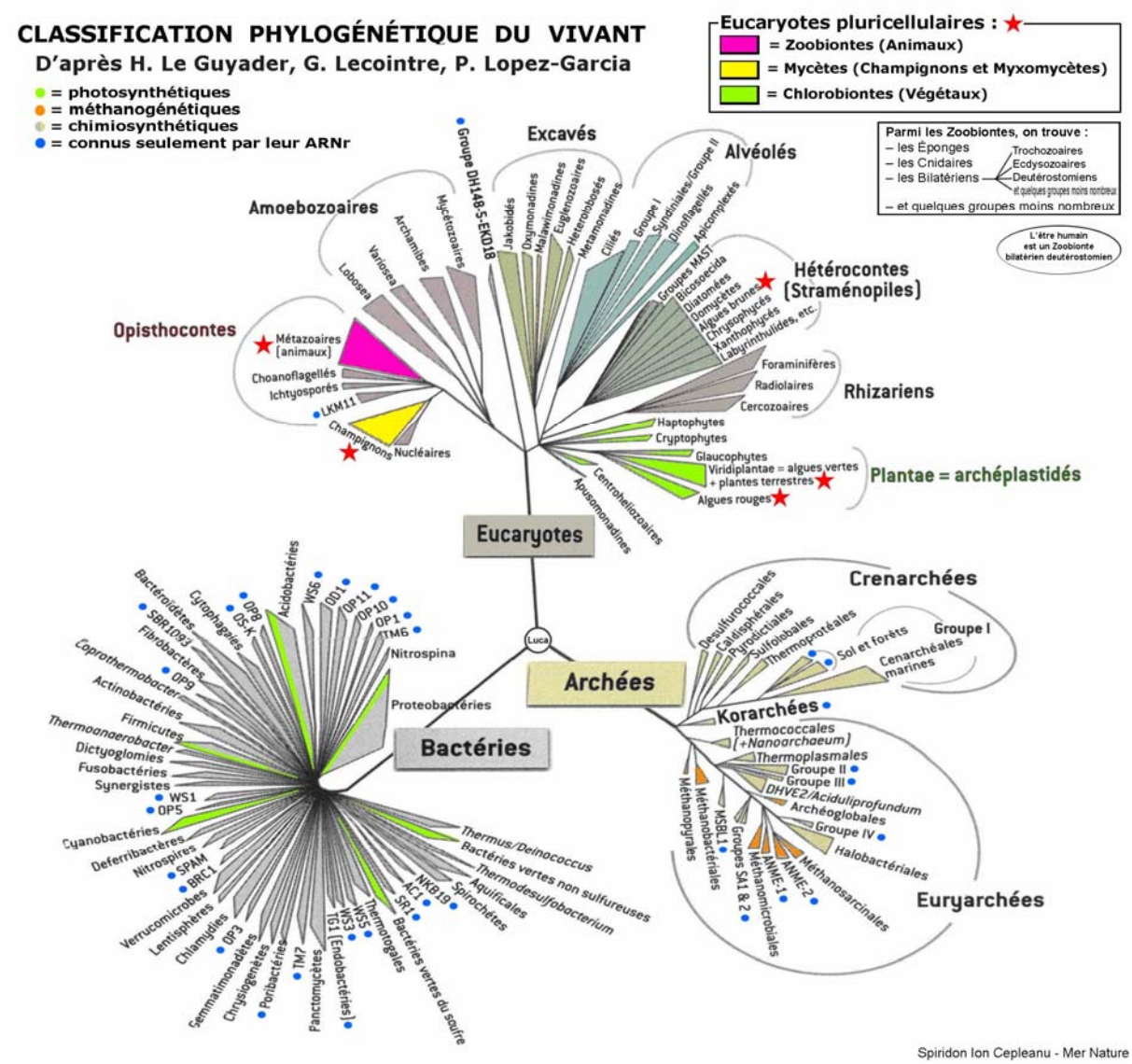
- La **Génomique structurale** qui séquence et compare les génomes entiers ;
- la **Génomique fonctionnelle** qui détermine la fonction et l'expression des gènes, à travers l'étude du **Transcriptome** qui s'intéresse aux divers ARN présents dans la cellule et à leurs rôles, et au travers du **Protéome** qui identifie les protéines codées par le génome et détermine leurs rôles ;
- **-** la **Métabolomique** qui relie génome et métabolisme ;
- **-** la **Phylogénie moléculaire** qui situe les organismes vivants les uns par rapport aux autres en se basant sur la comparaison des séquences d'ADN ;
- **-** l'**Écologie ou épidémiologie moléculaire** qui étudie en conditions naturelles ou par modélisation le comportement des composants des populations naturelles ;
- la **Transgenèse** qui vise à modifier des fonctions biologiques par action sur l'ADN ;
- **-** la **Biologie de synthèse** qui cherche à concevoir des êtres vivants adaptés à de nouvelles fonctions.

2 - RAPPELS DE BIOLOGIE

2-1 – On regroupe les êtres vivants en trois règnes

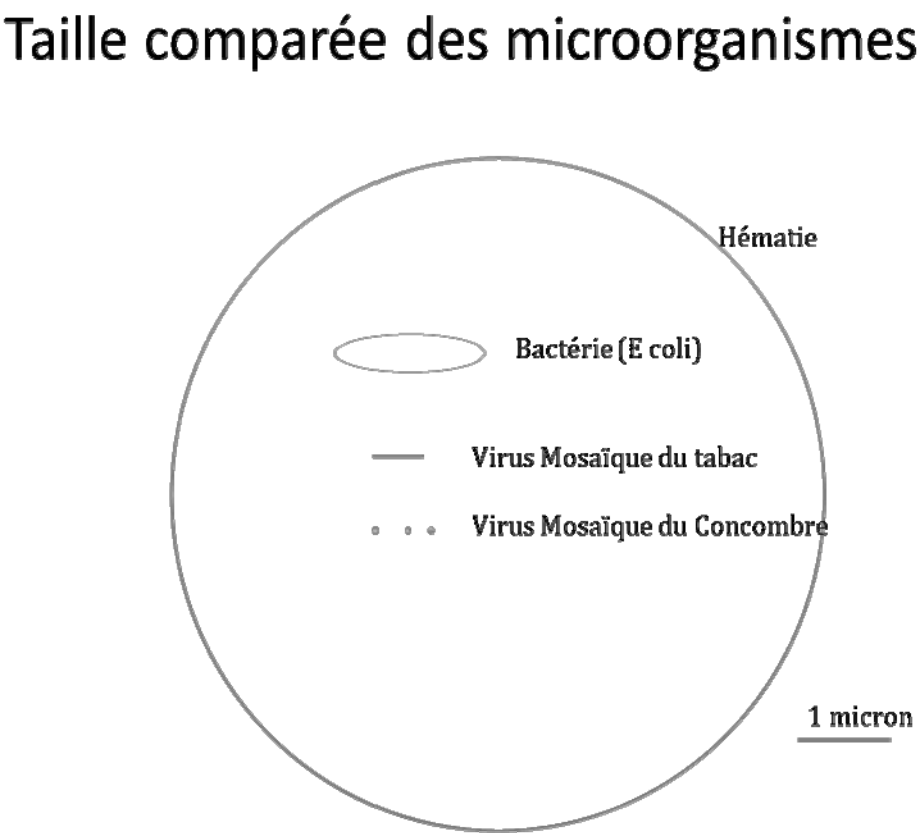
Les données biologiques et moléculaires récentes conduisent actuellement à diviser le monde du vivant en trois règnes selon la méthode cladistique (HENNIG, 1950) :

- Les **Bactéries** caractérisées par l'absence de noyau.
- Les **Archées** de formes analogues à celles des bactéries.
- Les **Eucaryotes** qui regroupent tous les organismes vivants formés d'une ou plusieurs cellules de type animal, végétal ou champignon.



Il reste encore des questions ardemment débattues telle celle concernant la place des virus dans le monde du vivant.

Des travaux récents remettent en question l'évolution en trois règnes et considèrent qu'il n'y a, au départ, que deux règnes, les bactéries et les archées. Quant aux eucaryotes, ils dériveraient des archées complétés par une évolution symbiotique avec une ou des bactéries « partenaires » (WILLIAMS et al., 2013).



Les Bactéries

Les bactéries ont des dimensions de l'ordre de 1 à 3 microns, et, à l'exception des **mycoplasmes**, elles possèdent une paroi rigide qui définit leur forme. Elles présentent en général un ou plusieurs flagelles et des cils.
Les cellules bactériennes n'ont pas de noyaux figurés mais possèdent un grand ADN circulaire qui porte l'essentiel de l'information génétique et de petits ADN circulaires complémentaires (plasmides) facilement échangeables entre individus ou intégrables dans le génome.
Elles renferment des ribosomes contenant en particulier un ARN 16S (Svedberg) très utilisé dans les caractérisations phylogénétiques. *(Les ribosomes sont des organites cellulaires qui servent à fabriquer des protéines à partir de l'information apportée par les ARN messagers).*
En conditions favorables, les bactéries se multiplient en très grand nombre par scissiparité. Elles peuvent aussi varier car leur génome mute très facilement (ex : système SOS qui déclenche un taux de mutation accéléré en cas de stress subi) et possède plusieurs mécanismes pour transférer facilement des gènes entre espèces différentes *(transfert latéral de gènes)*.
Dans les années 1960, on ne connaissait que quelques milliers d'espèces de bactéries identifiées grâce à leurs formes et à leurs réactions sur une batterie de milieux sélectifs lors de leur mise en culture. Les nouvelles techniques moléculaires (métagénomique, séquençage de l'ARN 16S) montrent que des centaines de milliers, et peut-être des millions, d'espèces non cultivables par les techniques conventionnelles existent dans la nature.

La classification des bactéries est en constante évolution avec la découverte de nouveaux **phylums** (40 connus en 2005). Un livre, le « BERGEY's Manual of systematic bacteriolology », régulièrement remis à jour, tente de recenser les espèces reconnues.
Les bactéries colonisent de très nombreux milieux naturels ou artificiels et peuvent former des **biofilms**, ensembles de bactéries qui se reconnaissent entre elles par des signaux chimiques (quorum sensing) et qui synchronisent leur comportement (production de toxines, de substances protectrices, etc.).
Certaines espèces possèdent une capacité de photosynthèse (cyanobactéries).

Les Archées

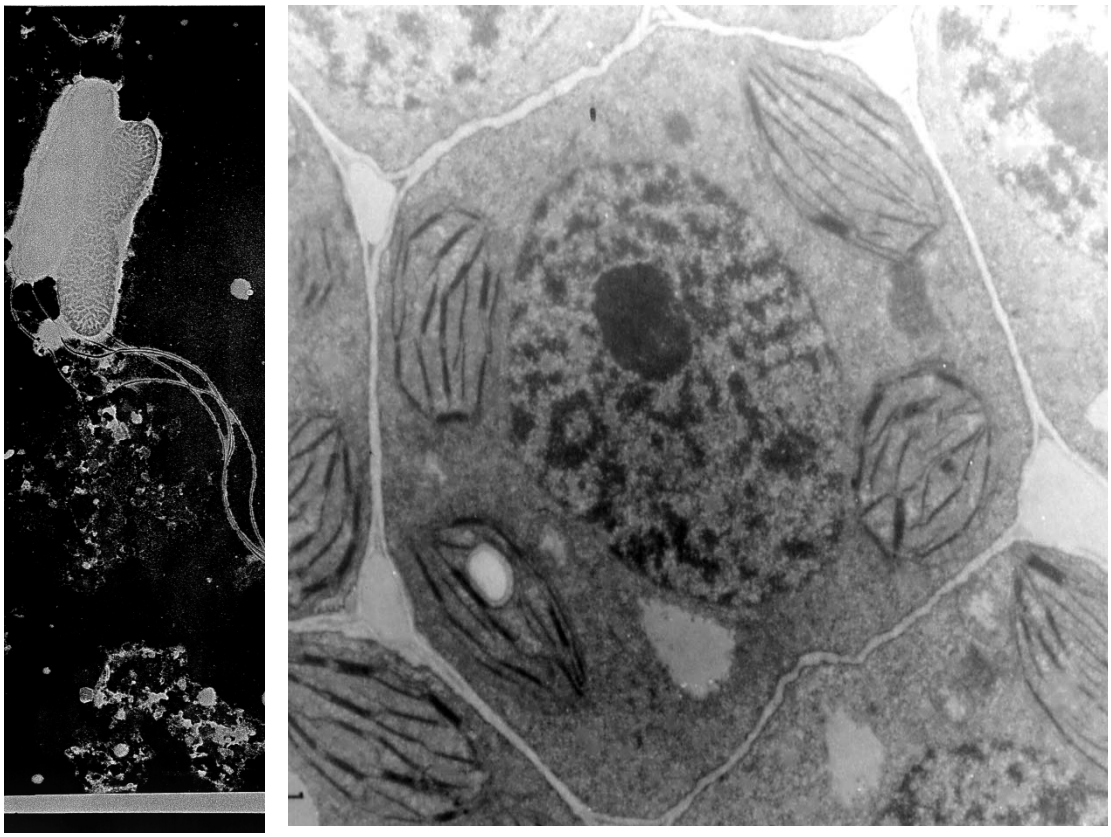
Les **Archaea** ou **archéobactéries** ou **archées** ont été décrites par Carl WOESE et George FOX en 1977. Comme les bactéries, les Archaea ont des formes sans noyau figuré. Elles présentent des gènes et des structures de génome différents des bactéries qui conduisent à les considérer comme un règne à part entière.
Elles se multiplient par scissiparité et renferment des ribosomes de type bactérien.
Leur génome peut contenir des portions proches du génome des bactéries et d'autres proches des eucaryotes. Comme les bactéries, elles présentent un génome circulaire, des plasmides et des gènes avec peu d'introns souvent organisés en opérons.
Comme les eucaryotes, elles possèdent des **nucléosomes** avec des **histones** et utilisent des promoteurs de types eucaryotes.
Elles peuvent présenter des originalités métaboliques : croissance à plus de 100°C, méthanogenèse…
Le règne ne comprend actuellement que quatre phylums : Crenarchaeota, Euryarchaeota, Korarchaeota, Nanoarchaeota recensés dans le « BERGEY's Manual ».
Les Archaea, d'origine très ancienne mais de découverte récente, colonisent le plus souvent des milieux extrêmes (sources chaudes, geysers, fumeurs sous-marins, écoulements toxiques de mines et de zones industrielles, etc.) mais elles sont aussi présentes dans des sols classiques.

Les Eucaryotes

Les eucaryotes regroupent les organismes vivants constitués d'une ou de plusieurs cellules possédant en particulier un noyau limité par une membrane ponctuée. Les chromosomes sont constitués par l'ADN enroulé autour de protéines (histones). La division de la cellule se fait par un mécanisme équationnel (mitose) ou réductionnel (méiose) permettant la formation de gamètes et une reproduction sexuée qui peut assurer une certaine diversification.
Le cytoplasme contient des ribosomes responsables de la fabrication des protéines codées par le génome nucléaire. Le séquençage d'un ARN ribosomal 18S est très utilisé pour la différenciation des espèces. Le cytoplasme contient aussi des organites (mitochondries et chloroplastes) qui sont des vestiges de bactéries absorbées par la cellule au cours de l'évolution. Ces organites renferment encore un peu d'ADN et des ribosomes de type bactérien. Les mitochondries assurent l'alimentation en énergie de la cellule (cycle de Krebs) et, chez les algues et les cellules végétales, les chloroplastes sont le lieu de la photosynthèse.
La classification phylogénétique réunit chez les eucaryotes la majorité des êtres vivants répertoriés : protozoaires, champignons, algues, plantes, arthropodes, vertébrés…

Les Virus

Les virus sont constitués le plus souvent par un seul acide nucléique (ADN ou ARN) codant quelques gènes, lui-même protégé par une « capside » formée de protéines et parfois d'une enveloppe qui constitue une protection complémentaire.
Actuellement, les virus ne sont pas considérés comme des êtres vivants car ils ne peuvent pas se multiplier par eux-mêmes. Leurs gènes doivent être lus par les ribosomes de la cellule hôte.
Certains virus, comme le HIV, responsable du SIDA, codent une enzyme (transcriptase reverse) qui leur permet d'intégrer leur propre ARN dans les chromosomes de la cellule hôte. Une fois intégrés dans le génome de l'hôte, ces virus peuvent devenir inactifs ou produire des rétrotransposons capables de se déplacer sur le génome de l'hôte. Ils peuvent aussi redonner des virions infectieux générateurs d'épidémies.
Le rôle pathogène des virus est bien connu mais on s'intéresse de plus en plus à leur action, en conditions naturelles, comme transporteurs de gènes ou de microARN entre individus.



Les cellules bactériennes à gauche ne contiennent pas de noyau figuré. Elles se divisent par scissiparité. Une cellule eucaryote, ici une cellule végétale à droite, présente un noyau qui contient l'ADN sous forme de chromosomes relaxés en période de quiescence. Certains organites comme ici les chloroplastes ou les mitochondries peuvent aussi renfermer un peu d'ADN car ce sont d'anciennes bactéries qui ont été intégrées par la cellule au cours de l'évolution (clichés DELÉCOLLE et QUIOT).

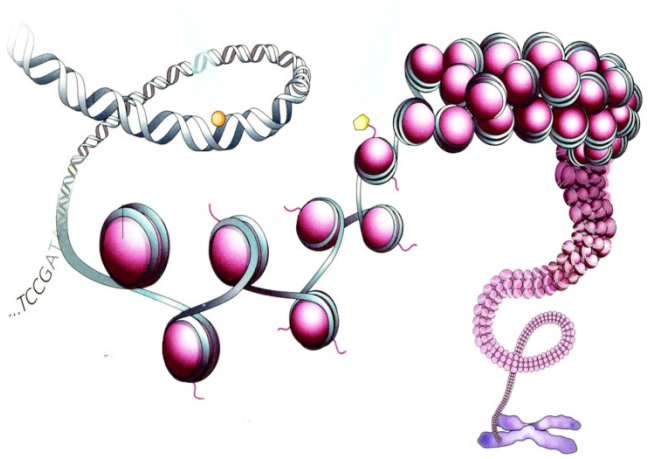
2-2 – Les acides nucléiques

Il existe deux acides nucléiques : **l'acide désoxyribonucléique ou ADN (ou DNA)** et **l'acide ribonucléique ou ARN (ou RNA)**.
Ce sont des macromolécules linéaires souvent très longues, constituées par une succession ordonnée de quatre nucléotides.
Un nucléotide est formé par l'association d'un sucre, le désoxyribose dans l'ADN ou le ribose dans l'ARN, d'un résidu phosphate et d'une base purique ou pyrimidique.
Cinq bases sont utilisées par les acides nucléiques: deux bases puriques (**adénine, A et guanine, G)** et trois bases pyrimidiques (**cytosine, C ; thymine, T et uracile, U**).
Les nucléotides sont nommés selon la nature de la base qu'ils contiennent : A, T, G, C, U.
Dans un acide nucléique, la suite des nucléotides est appelée **séquence**.

L'ADN

L'ADN est formé par une succession des nucléotides A, T, G et C.
Selon la célèbre démonstration de WATSON et CRICK (1953), l'ADN se présente sous la forme de deux brins antiparallèles appariés qui s'enroulent en double hélice. L'appariement est réalisé par des liaisons chimiques faibles entre les deux bases A et T (deux liaisons) et G et C (trois liaisons). L'appariement peut être rompu par des enzymes ou par des contraintes thermiques ce qui a une grande importance biologique.

L'ADN double brin (ou dicaténaire) est présent sous forme de très longues chaînes (environ 2 m chez l'Homme) dans les chromosomes des cellules eucaryotes. Il est enroulé par paquets de 164 nucléotides autour de groupes de huit protéines appelées **histones**, l'ensemble formant un **nucléosome**. Le ruban de nucléosomes forme la **chromatine**. La chromatine peut être très condensée, c'est l'**hétérochromatine** ou, au contraire, plutôt relaxée, c'est l'**euchromatine**.
Cette différence de structure joue un rôle très important dans la disponibilité de certaines portions de l'ADN pour le fonctionnement de la cellule. L'ADN présent dans les zones condensées n'est pas accessible et ne peut pas être copié en ARN messager sans modification du chromosome.



Représentation schématique d'un chromosome et de la chaine d'ADN double brin enroulée autour de protéines histones (en rouge). L'enroulement plus ou moins compact joue un rôle dans l'accessibilité des séquences d'ADN pour la copie en ARN.

Dans les bactéries, l'ADN est circulaire et superenroulé sans la présence d'histones.

L'ARN

L'ARN est formé par une succession de nucléotides A, U, G et C.
Il est présent le plus souvent sous forme de simple brin, de longueur variable. De ce fait, il est moins stable et peut être détruit par des enzymes spécifiques, les ribonucléases (ou **Rnases**). Il peut circuler dans la cellule, aussi bien dans le noyau que dans le cytoplasme.
Les premiers travaux montraient le rôle essentiel des ARN dans la synthèse des protéines. Actuellement les ARN font l'objet de recherches actives qui révèlent que leur rôle est beaucoup plus important que pensé au départ.

2-3 - Décryptage de l'information contenue dans les acides nucléiques : le code génétique

Les études de MENDEL (1865) sur la descendance des croisements entre pois ridés et pois lisses avaient jeté les bases de l'hérédité. On doit à l'équipe de MORGAN (1910), travaillant sur les populations de mouches Drosophiles, d'avoir montré que l'hérédité de certains caractères appelés gènes se situait au niveau des chromosomes.
Au début des années 1960, on démontrait que l'ADN est le support de l'hérédité et que la séquence des 4 nucléotides ATGC sur sa chaîne est reproduite sur les ARN sous forme AUGC. Ces ARN copies de portions de l'ADN sont appelés **ARN messagers (ARNm)**. Ils transportent l'information du noyau vers le cytoplasme où ils sont interprétés par une machinerie cellulaire appelée **ribosome** pour fabriquer des protéines en associant de façon ordonnée les vingt types d'acides aminés connus.

Le code génétique

Premier nucléotide	Deuxième nucléotide U		Deuxième nucléotide C		Deuxième nucléotide A		Deuxième nucléotide G		Troisième nucléotide
U	UUU, UUC	phényl-alanine	UCU, UCC, UCA, UCG	sérine	UAU, UAC	tyrosine	UGU, UGC	cystéine	U C A G
	UUA, UUG	leucine			UAA, UAG	STOP	UGA	STOP	
							UGG	tryptophane	
C	CUU, CUC, CUA, CUG	leucine	CCU, CCC, CCA, CCG	proline	CAU, CAC	histidine	CGU, CGC, CGA, CGG	arginine	U C A G
					CAA, CAG	glutamine			
A	AUU, AUC, AUA	isoleucine	ACU, ACC, ACA, ACG	thréonine	AAU, AAC	asparagine	AGU, AGC	sérine	U C
	AUG	méthionine			AAA, AAG	lysine	AGA, AGG	arginine	A G
G	GUU, GUC, GUA, GUG	valine	GCU, GCC, GCA, GCG	alanine	GAU, GAC	acide aspartique	GGU, GGC, GGA, GGG	glycine	U C A G
					GAA, GAG	acide glutamique			

http://raymond.rodriguez1.free.fr/Documents/Cellule-genome/codeG.jpg

Cette découverte essentielle a permis de définir le **code génétique**. Des séquences de trois nucléotides présents sur l'ARN forment des triplets ou **codons** qui définissent les acides aminés selon le code ci-dessus.

Par exemple, le triplet UUU définit l'acide aminé phénylalanine.

On constate que 61 des 64 triplets possibles codent des acides aminés et que la plupart des acides aminés sont codés par plusieurs codons. Ce sont souvent les deux premiers nucléotides qui sont déterminants dans la définition de l'acide aminé. On dit que le code est dégénéré.
Observons que ce code est universel et se retrouve avec quelques très rares modifications chez la totalité des êtres vivants, ce qui plaide pour une origine commune de la vie.

Des différences dans la succession des nucléotides sont fréquemment rencontrées lorsqu'on compare des séquences homologues de plusieurs individus. Elles mettent en évidence la variabilité pouvant exister au sein d'un groupe.

ALIGNEMENT DE LA SÉQUENCE NUCLÉOTIDIQUE DES FRAGMENTS DE PCR DE PLUSIEURS ISOLATS D'UN MÊME VIRUS

```
Isolat 1    TTCAACGATACGGGCACGTGAAGCTCATATCCAGAT
isolat 2    - - - - - - - - - - - - - - - A - - - - - - - - - - - - - - - - -
isolat 3    - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -A- -
isolat 4    - - - - - - - - - - - - - - - - - - - - - C - - - - - - - - - - -
isolat 5    - -A -C - -T - - CC- -T - - - - - - - - - - - - - A - - - -
isolat 6    - -A -C - -T - - CC- -T - - - - - - - - - - - - - AA - - - -
```

Différences dans les séquences nucléotidiques d'un fragment d'acide nucléique de plusieurs isolats d'un même virus. Seules les différences par rapport à la séquence du haut sont représentées sur le tableau.

VARIABILITÉ DANS LES SÉQUENCES D'ACIDES AMINÉS PRÉDITES À PARTIR DES SÉQUENCES DES ACIDES NUCLÉIQUES

```
Isolat 1    UUCAACGAUACGGGCACGUGAAGCUCAUAUCCAGAU
              F   N   D   T   G   T   W   S   S   Y   P   D
Isolat 2    - - - - - - - - - - - - - - - -A - - - - - - - - - - - - - - - - -
                                          K
Isolat 3    - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -A- -
                                                                            N
Isolat 4    - - - - - - - - - - - - - - - - - - - - - - - - - C- - - - - - - - - -
                                                            T
Isolat 5    - - A- C- -U - - -CC - -U - - - - - - - - - - - - - -A- - - -
              L   T   V   T   R   S                           Q
Isolat 6    - - A- C- -U - - -CC - -U- - - - - - - - - - - - - -AA- - - -
              L   T   V   T   R   S                           K
```

L'utilisation du code génétique permet de prédire la séquence d'acides aminés (indiqués par leurs initiales en bleu et vert). Comme le code génétique est dégénéré, certaines mutations sur l'acide nucléique n'entraînent pas de modification de l'acide aminé. Ex : la séquence ACG devenue ACC ne produit pas de changement de l'acide aminé thréonine (T).

En pratique, des logiciels de bio-informatique permettent d'aligner les séquences provenant de mêmes zones collectées dans différents individus. On peut ainsi visualiser le niveau de variation d'une population. On peut aussi traduire en acides aminés les codons de ces isolats et repérer les mutations pertinentes qui modifient l'acide aminé codé et les mutations silencieuses qui n'ont pas d'effet sur la séquence d'acides aminés.

2-4 – La synthèse des protéines

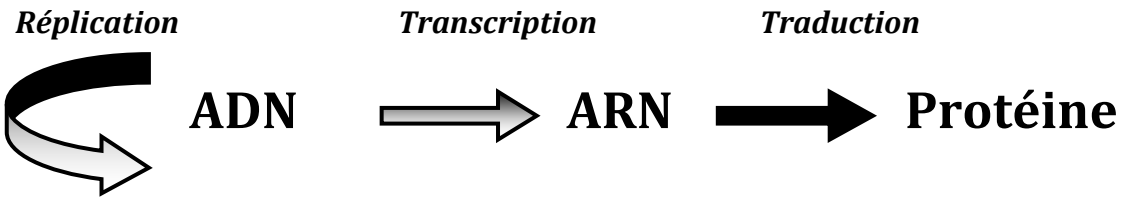
Le décryptage du code génétique et les premières analyses de séquences ont permis de comprendre le mode de circulation de l'information dans la cellule.

La séquence de nucléotides présente sur l'ADN chromosomique constitue la mémoire programme de la cellule.
Au moment de la division d'une cellule, qu'elle soit bactérienne ou eucaryote, l'ADN des chromosomes est copié par une enzyme de type DNA polymérase DNA dépendante pour former deux chromosomes fils qui iront dans les deux cellules filles. Ce mécanisme de copie de l'ADN est appelé **réplication**.

Pour pouvoir être utilisée, l'information présente sur le chromosome doit être copiée sur un **ARN messager, ARNm**, qui va ensuite transporter l'information depuis le noyau où se trouve le chromosome vers le cytoplasme. C'est la **transcription.**

Dans le cytoplasme, les codons présents sur l'ARNm vont être traduits par un **ribosome** sous la forme d'une chaîne d'acides aminés qui constituera la structure primaire d'une protéine. C'est la **traduction**.
Les ribosomes, en grand nombre dans la cellule, sont une association **d'ARN ribosomaux, ARNr**, et de protéines qui vont présenter à chaque codon de l'ARNm l'acide aminé correspondant apporté par des types spécifiques d'ARN appelés **ARN tranfert, ARNt**. Le ribosome va alors constituer une chaîne d'acides aminés en les détachant des ARNt et en les unissant entre eux par des liaisons peptidiques de type chimique. Par la suite, un processus de maturation donnera à la chaîne d'acides aminés sa structure tridimensionnelle de protéine active.

En 1956, ce mécanisme fut baptisé « dogme central » par Francis CRICK. Il peut être résumé par le schéma suivant :



À partir des années 1980, ce dogme sera progressivement modifié et complété, mais les trois termes (réplication, transcription et traduction) constituent toujours les fils directeurs des mécanismes régissant la circulation de l'information génétique.

3 - LES PROGRÈS DANS LES TECHNIQUES D'ÉTUDE DEPUIS 40 ANS

L'essor rapide de la biologie moléculaire ne peut pas se comprendre si l'on ignore les progrès fulgurants des techniques de laboratoire et d'analyse des génomes depuis les années 1970.

Dans les années 1970

- Commercialisation de produits chimiques ultra purs (ex : produits ANALAR exempts de ribonucléases).
- Commercialisation d'enzymes purifiées agissant sur les acides nucléiques (ex : polymérases permettant de copier un ADN, enzymes de restriction qui sont capables de couper un ADN au niveau d'une séquence de nucléotides définie …). On parle de ciseaux moléculaires qui permettent de disséquer précisément, de copier ou de rabouter des brins d'ADN.
- Miniaturisation du matériel de laboratoire en raison du prix des réactifs (pipettes de type Gilson permettant de distribuer de façon fiable des quantités de l'ordre du microlitre), utilisation progressive de micro-vaisselle de laboratoire en plastique jetable (ex : tube Eppendorf de 0,5 mL).
- Développement de méthodes de sélection et de multiplication de petits fragments d'acide nucléiques dans des bactéries ou des levures puis extraction et purification par des méthodes chimiques (miniprep).
- Progrès de l'électrophorèse et de ses dérivés (Western Blot, Northern Blot, Southern Blot) qui permet de purifier et de caractériser de petites quantités de protéines ou d'acides nucléiques (*Voir Annexe 1*).
- En immunologie, développement du test ELISA (Enzyme-Linked-ImmunoSorbent-Assay) qui permet en 24 h de détecter spécifiquement et de quantifier toutes sortes d'antigènes si l'on dispose de l'anticorps spécifique.

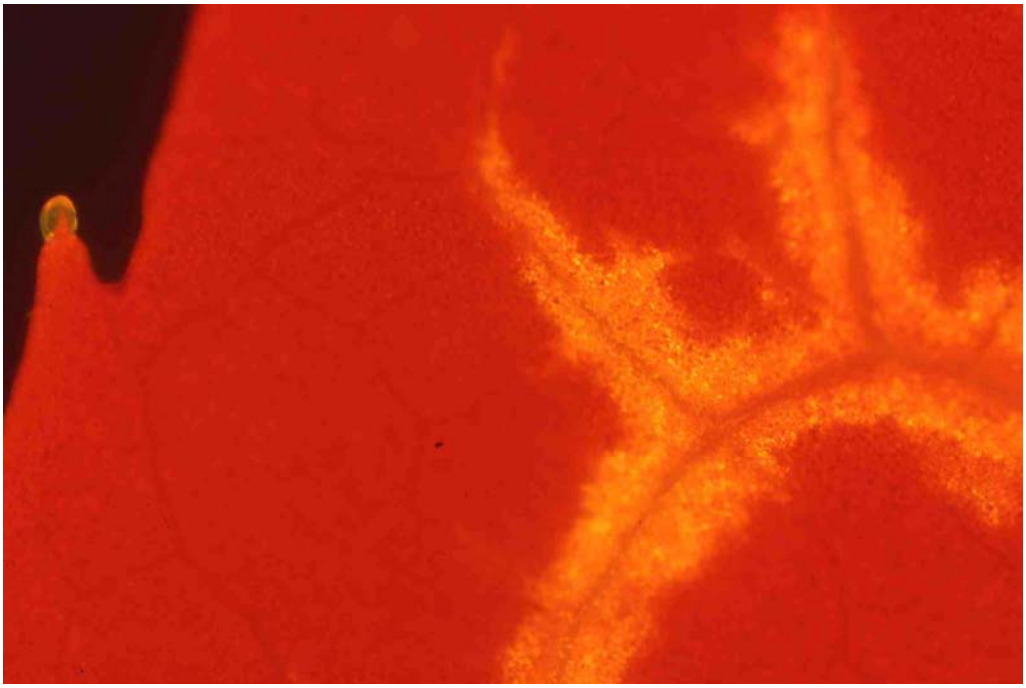
Dans les années 1980

- Publication de « Molecular cloning » (Cold Spring Harbour ed. 1982) qui vulgarise les nouvelles techniques dans les laboratoires du monde entier.
- Progrès dans les techniques de purification des acides nucléiques et des protéines suivis par la commercialisation de kits d'extraction et de purification rapides des ADN et ARN.
- Mise au point des **anticorps monoclonaux** permettant de produire avec une grande reproductibilité des anticorps spécifiques de zones précises (**épitopes**) présentes sur une macromolécule cible.
- Développement dans les laboratoires du séquençage selon les méthodes de SANGER (la plus usitée) ou de MAXAM et GILBERT. Il devient possible de séquencer en quelques mois des génomes de l'ordre de 10 000 nucléotides (*Voir annexe 3*).

Dans les années 1990

- Arrivée des séquenceurs automates remplaçant dans la technique de SANGER les marqueurs radioactifs par des marqueurs colorés.

- Mise au point et généralisation mondiale de la **PCR** (Polymerase Chain Reaction) et de ses variantes. Cette technique simple et très performante est basée sur l'usage d'une enzyme polymérase thermorésistante et d'un bain-marie à sec automate appelé **thermocycleur.** Elle permet, en quelques heures, de multiplier par près d'un milliard une portion ciblée d'un acide nucléique (*Voir annexe 2).*
- Début du séquençage de grands génomes entiers souvent par des collaborations internationales de plusieurs années (ex : génome humain, plus de 3 milliards de nucléotides). Le séquençage mené par un consortium international sous la direction de J. WATSON commença en 1989 et se termina en 2001/2003 pour un coût de 3 milliards de $. En 1998, Craig VENTER à la tête d'une société privée entreprit aussi de séquencer un génome humain (le sien) grâce à l'appui de la Sté Perkin Helmer, le fabricant des premiers séquenceurs automatiques. Disposant de centaines de séquenceurs fonctionnant en parallèle, il publia aussi une séquence brute en 2001.
- Développement des banques de séquences où chaque chercheur qui a séquencé une portion d'ADN peut déposer son résultat. Celui-ci est en accès libre sur Internet avec des outils d'analyse pour la communauté scientifique (ex. : GenBank, EMBL, pour les ADN, Swissprot pour les protéines) La croissance est très rapide et commence à poser des problèmes de gestion. En 2006, GenBank contenait déjà 61 millions de séquences !
- Disponibilité sur Internet de programmes de bio-informatique permettant d'aligner, de comparer, classer ou hiérarchiser ces séquences.
- Développement de biomarqueurs colorés qui permettent de marquer spécifiquement des macromolécules. Appelées sondes moléculaires, ces biomarqueurs sont surtout utilisés dans des techniques d'observation *in situ*, ex : immuno-empreinte, FISH (Fluorescent In Situ Hybridation). On sait aussi fixer la séquence d'une protéine fluorescente à l'ADN que l'on veut observer, ex : GFP (Green fluorescent Protein).



Utilisation du marquage à la GFP (Green Fluorescent Protein) : Observation en microscopie à fluorescence de la progression d'un virus (PPV), marqué à la GFP, le long des nervures d'une feuille de Nicotiana benthamiana. Le gène GFP, issu au départ d'une méduse, a été ajouté au génome du virus. Chaque fois que le virus se réplique, il produit pour chaque copie un exemplaire de la protéine fluorescente qui révèle la position du virus dans l'hôte (cliché Quiot).

Dans les années 2000

- Apparition à un rythme rapide de nouvelles techniques de séquençage ultra performantes, puis de nouvelles générations, qui permettent de séquencer en quelques jours des génomes de très grandes tailles à des coûts de plus en plus réduits. En 2014, un génome humain pourrait être séquencé en une semaine pour un coût de l'ordre de 1000 $ (*Voir annexe 3*).
- Développement des « puces à DNA » ou **microarrays**. Il s'agit de regrouper sur des plaques de verre ou de silicium un nombre parfois très grand de sondes spécifiques (brins d'ADN). Mises en présence d'un extrait biologique, ces sondes peuvent s'apparier avec un brin d'ADN complémentaire s'il est présent dans l'extrait et former un ADN double brin. Un marqueur coloré avec une fluorescence spécifique en UV a été fixé auparavant aux brins d'ADN de l'extrait. Un point lumineux sur la plaque lue sous un éclairage UV signe la présence, dans l'extrait de l'ADN recherché.
Cette technique est utilisée en particulier pour l'étude du transcriptome. Elle permet de détecter les ARN présents dans une cellule à un instant donné si l'on a fait d'abord agir une enzyme (la transcriptase reverse) pour transformer en ADN complémentaires les ARN de l'extrait cellulaire.
- L'analyse des masses gigantesques de données collectées est possible en combinant équipes pluridisciplinaires et collaborations internationales sur des programmes ciblés, utilisation des nouveaux automates, des ordinateurs de très grande puissance et des programmes de bio-informatique de plus en plus performants.

4 - QUELQUES RÉSULTATS MARQUANTS DE LA BIOLOGIE MOLÉCULAIRE

4-1 – Le séquençage des génomes entiers

À la suite des travaux de SANGER et de MAXAM et GILBERT sur les techniques de séquençage publiés en 1977, des laboratoires commencent à séquencer des portions de génomes puis des génomes de petites tailles comme ceux des virus de plantes et d'animaux.
La première séquence d'un génome entier présent dans une cellule vivante, la levure *Saccharomyces cerevisiae*, est publiée en 1996. Puis suivent, entre autres, la bactérie *Escherichia coli* en 1997, la Drosophile en 2000, une plante adoptée comme modèle mondial de référence, *Arabidopsis thaliana*, en 2000 ; la souris, *Mus musculus*, en 2002 et l'Homme en 2003.
Par la suite, le progrès des techniques de séquençage permet une forte accélération pour atteindre, en septembre 2013, le nombre de 6 887 génomes terminés et accessibles sur Internet.
D'autre part, la très forte baisse du coût du nucléotide séquencé permet de fiabiliser les résultats en répétant plusieurs fois les mêmes séquençages (**Deep sequencing**) pour compenser les erreurs éventuelles de l'enzyme polymérase utilisée.
L'accumulation de ces séquences pose des problèmes de coût de stockage, de gestion et de disponibilité dans les grandes banques de données (en septembre 2013 la banque européenne EMBL doit gérer une librairie de 670 milliards de nucléotides), il en est de même pour GenBank aux USA et DDBJ au Japon. Depuis quelques années, on voit

apparaître des banques de données locales spécialisées sur un organisme ou sur une thématique précise (178 en 2013) (PERRIÈRE 2013).

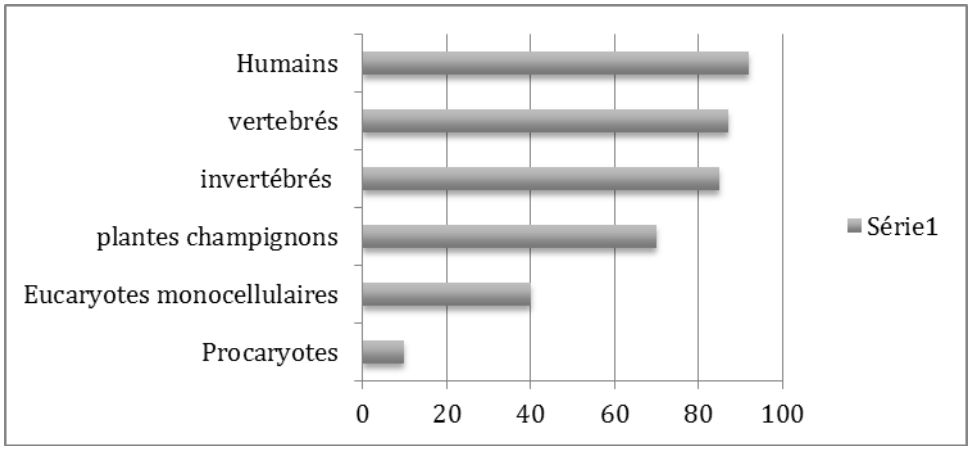
L'analyse par la communauté scientifique de cette masse d'informations explique en grande partie le progrès étonnant des connaissances ces dernières années.

4-2 - Génome et junk DNA ou ADN poubelle

Les premières analyses de séquences ont réservé de grosses surprises aux scientifiques.

La plupart des génomes séquencés ne parait coder réellement qu'un nombre de protéines voisin de 20 000. Le reste de l'ADN n'étant pas traduit en protéine. Cela a conduit, dans un premier temps, à baptiser **ADN poubelle** ou **Junk DNA** la partie du génome constitué de DNA non traduit en protéines.

D'autre part, la comparaison des génomes de différents groupes évolutifs a montré que la proportion d'ADN poubelle augmentait à mesure que l'on s'élevait dans l'échelle de complexité. Cet ADN poubelle pouvait représenter plus de 90 % du génome chez les espèces les plus évoluées.
Ces observations ont suscité d'intenses recherches au cours des dernières années.



Pourcentage approximatif d'ADN poubelle (JunkDNA) selon les types d'êtres vivants. On observe que la quantité augmente à mesure que l'on monte dans l'échelle de complexité. (inspiré de KHALIL, 2013).

4-3 - Révision de la notion de gène

La notion de gène s'est profondément modifiée au cours des quinze dernières années. La représentation classique du gène correspondait à un segment d'ADN délimité par des codons spécifiques qui était recopié dans le noyau en ARNm par une enzyme de type ARN-polymérase-ADN-dépendente. L'ARNm migrait ensuite dans le cytoplasme où il était pris en charge par des ribosomes. Les ribosomes sont des assemblages d'ARN et de protéines capables de lire la séquence des codons de l'ARNm, pour y apparier des acides aminés apportés par des **ARN transfert (ARNt)** et pour former finalement une chaine d'acides aminés. Cette chaine d'acides aminés peut être ensuite mise en conformation tridimensionnelle par des protéines spécialisées, les **chaperones**. Elle devient alors une protéine ayant une fonction précise : enzyme, protéine de structure, etc.

L'analyse du transcriptome (les ARN présents dans la cellule) montre que des gènes, que l'on croyait monolithiques, peuvent se cliver et s'associer pour produire un nombre d'ARNm beaucoup plus important que le nombre de gènes identifiés au départ (PEARSON, 2006).
On établit qu'un très grand nombre de gènes est constitué d'une mosaïque de blocs de séquences dont certains appelés **exons** sont nécessaires à la construction de la protéine tandis que d'autres blocs appelés **introns** (*intrus, mnémotechnique*) sont excisés. Excision des introns et raboutage des exons se réalisent dans le noyau au niveau de l'ARNm brut, grâce à un ensemble complexe appelé **splicéosome** constitué d'environ 300 protéines.
En pratique, le splicéosome est capable d'exciser des introns et de rabouter des exons provenant d'un même gène mais aussi des exons assez éloignés provenant d'autres zones du même chromosome.
On aboutit à la notion d'épissage alternatif qui consiste à combiner, dans des protéines fonctionnelles, des exons pouvant provenir de différents gènes, chaque exon pouvant exercer une fonction différente (localisation de la protéine dans la cellule, activité enzymatique, interaction avec d'autres protéines…). De la sorte, on peut estimer que les 20 000 gènes repérés dans le génome peuvent en fait permettre la formation de 100 000 à 1 000 000 d'isoformes protéiques (AUBEUF, 2013).

Transcription, maturation et traduction d'un gène

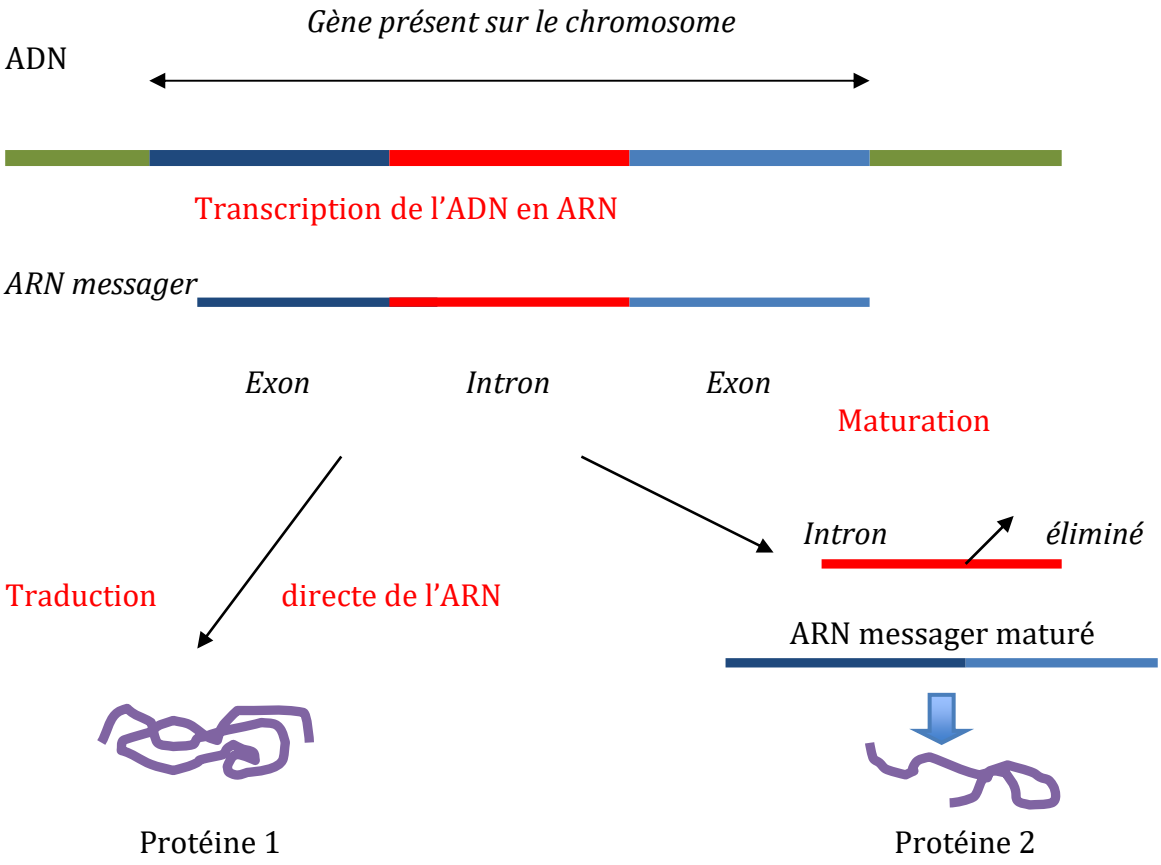


Schéma simplifié montrant la formation de deux protéines différentes à partir d'un même gène grâce au processus de maturation de l'ARNm dans le noyau.

Il restait à identifier les mécanismes qui incitent la cellule à fabriquer tel ou tel type d'assemblage d'exons, voire à modifier les protéines produites en fonction de leur localisation dans un organe donné.
Les recherches ont montré la très grande complexité des facteurs impliqués. On a cité des facteurs topologiques : niveau de condensation des histones sur l'ADN jouant sur la disponibilité pour la réplicase, répartition spatiale de l'ADN plus ou moins pelotonné dans le noyau (MISTELI, 2011).
On a aussi trouvé des facteurs épigénétiques (surtout méthylation de l'ADN ou des histones, *voir plus loin*) bloquant la transcription.
Un certain niveau de hasard pourrait aussi intervenir dans l'expression des gènes, expliquant une certaine variabilité et adaptabilité aux sollicitations du milieu extérieur (HEAMS, 2009).
C'est à partir de là que l'ADN poubelle, une fois décrypté a révélé son importance.

4-4 - La compréhension de l'ADN poubelle : un bouleversement majeur dans la compréhension de la régulation cellulaire

L'étude du transcriptome, c'est à dire des ARN présents dans une cellule, montre que de très nombreux ARN peuvent être présents dans une cellule à un moment ou à un autre de sa vie ou en fonction de sa spécialisation (ex : BARASH et al., 2010). Leurs rôles et leur importance commencent à être compris.

Chez les bactéries et les archées mais aussi chez des champignons et quelques plantes, des ARN appelés **riborégulateurs** (riboswitches) sont mis en évidence (BARRICK et BREAKER, 2007, 2014).
Les riborégulateurs sont des segments d'ARN faisant partie de la séquence non codante d'un ARN messager dont la forme tridimensionnelle peut varier. Selon la présence ou non d'une molécule cible, le riborégulateur adopte par **allostérie** une conformation qui autorise ou non la fabrication de la protéine codée par le reste de l'ARN messager. On est en présence d'un système très précis qui permet à un ARN messager de décider de sa traduction en protéine en présence d'une molécule cible.
D'autres types d'ARN régulateurs ont aussi été détectés chez les bactéries ; par exemple, plus de 100 transcrits antisens (ARN copies du brin d'ADN complémentaire non codant d'un gène) ont été trouvés chez *E. coli* (BUC et MELLIN, 2011).

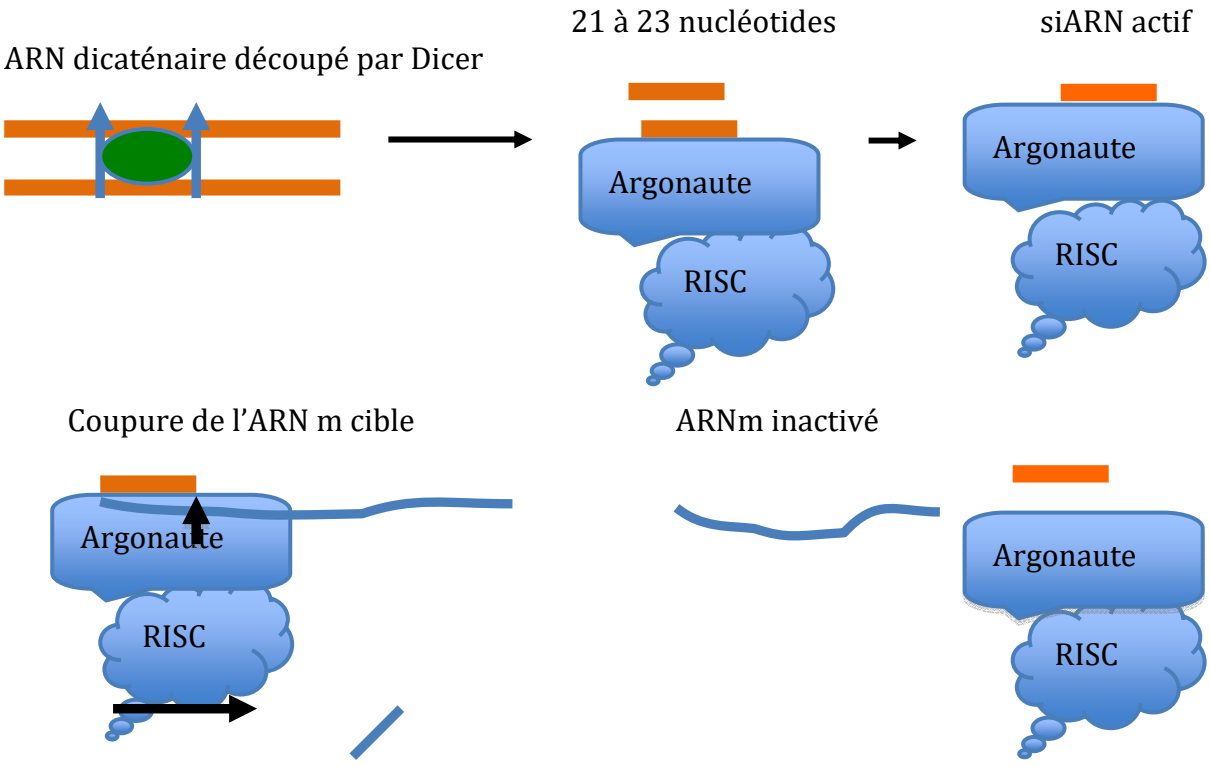
Chez les eucaryotes qui renferment un pourcentage plus important d'ADN poubelle, les systèmes de régulations sont plus complexes.
Actuellement on considère que, chez l'Homme, environ 1,5 % de l'ADN nucléaire code pour des exons qui sont traduits en protéines mais que environ 80 % de l'ARN poubelle peuvent être transcrits en ARN non codants, présents dans la cellule sans devoir être traduits en protéines. Des résultats analogues sont trouvés pour les autres groupes d'êtres vivants.
Les recherches de ces dernières années montrent que ces ARN transcrits mais non codants interviennent dans la régulation du génome selon plusieurs mécanismes : soit en coupant un ARNm pour le rendre illisible par un ribosome, soit en bloquant sa lecture sans l'inactiver, soit encore en agissant directement sur l'ADN du gène source en bloquant sa transcription en ARNm.
On est en présence de mécanismes très fins permettant de moduler très rapidement la traduction d'un gène donné. Cela peut permettre d'éviter l'accumulation d'une protéine

devenant toxique par excès, ou de bloquer la multiplication d'un virus pathogène ou peut-être de modifier rapidement au cours de son développement les fonctions d'une cellule donnée.

Plusieurs types de petits ARN régulateurs longs d'une vingtaine de nucléotides sont décrits : **siARN** et **miARN,** qui diffèrent par leurs modes d'action, telles que coupures d'ARNm ou répression de leur traduction.
Autre groupe de petits ARN, les **piARN** interviennent dans le développement des cellules germinales (GROBHANS et FILIPOWICZ, 2008 ; PINZON RESTREPO et MARTINEZ, 2014).

À titre d'exemple, on peut décrire en détail le fonctionnement des siARN, un des petits ARN intervenant dans le mécanisme de **PTGS (Post Transcripional Gene Silencing**).
(La PTGS est un mécanisme qui détruit spécifiquement un ARN cible avant qu'il ne soit traduit en protéine). Les siARN (siRNA, Small Interfering RNA) sont de petits ARN de 21 à 23 nucléotides qui sont produits à partir d'ARN dicaténaires (double brin). Ils sont découpés par une enzyme nommée **Dicer**. Ensuite, un des brins est pris en charge par une protéine de type **argonaute,** liée à un complexe de protéines **RISC** (**RNA Induced Silencing Complex**). En utilisant ce siARN comme détecteur de séquence complémentaire, l'argonaute et le complexe RISC vont reconnaître l'ARNm cible et vont le détruire en le coupant (*voir schéma).*

Mode d'action d'un siARN



Genèse et fonction d'un siARN : Un ARN dicaténaire est découpé en fragments d'une vingtaine de nucléotides par le Dicer, puis une protéine argonaute associée à d'autres protéines RISC prend en charge un brin et l'utilise pour détecter un ARNm cible monocaténaire complémentaire. L'ensemble RISC inactive l'ARNm cible en le coupant.

Ce mécanisme est important car des ARN double-brin peuvent exister dans la cellule. Il permet de lutter, par exemple, contre les virus à ARN+ en détruisant les formes réplicatives à ARN dicaténaires présentes dans la cellule infectée. Lorsqu'un petit virus à ARN entre dans une cellule, son ARN code d'abord une ARN-polymérase-ARN-dépendante qui va copier l'ARN viral en formant des ARN dicaténaires. C'est à ce stade que le virus est sensible à des coupures par le Dicer. Les siARN permettent aussi d'empêcher la mobilité de rétrotransposons tels que le virus VIH du SIDA, capable de s'intégrer dans le DNA du génome cellulaire, mais qui doit lui aussi passer par une phase d'ARN dicaténaire (sensible au siARN) pour se reproduire ou se déplacer.
Ce mécanisme a été découvert en 1990 par A. FIRE et C. MELLO chez le *Caenorhabditis elegans,* ce qui leur valut un prix Nobel en 2006.

Une autre famille d'ARN non codants, d'une longueur supérieure à 70 voire 200 nucléotides selon les auteurs, est décrite. Ce sont les **lncARN (lncRNA** ou **Long Non Coding RNA**) (KHALIL, 2013). Plusieurs milliers sont décrits, provenant soit de la transcription de parties intergéniques du génome, soit de la transcription du brin d'ADN complémentaire d'un gène normalement transcrit (transcription antisens). Leurs rôles pourraient comprendre une action ciblée sur le chromosome permettant de localiser l'action d'un complexe de protéines dont le rôle consiste à agir sur les histones pour passer de l'état d'euchromatine à celui d'hétérochromatine et inversement. Un exemple démontré consiste à bloquer l'activité du deuxième chromosome X chez les femelles qui portent deux chromosomes XX alors que les mâles portent les chromosomes XY.

Récemment, des **ceARN (ceRNA, Competing Endogenous RNA**) ont été décrits (HANSEN et al., 2013). Il s'agit d'ARN parfois circulaires produits naturellement dans les cellules et qui portent de nombreuses séquences complémentaires d'un microARN donné. Ces ceARN se comportent comme des éponges à microARN. Leur rôle paraît être de nettoyer la cellule d'un excès d'un microARN donné pour permettre le redémarrage immédiat de la traduction du gène correspondant. Un système de régulation qui se surajoute aux précédents et explique l'exquise adaptabilité des eucaryotes (TAY et al., 2014).

Principaux types d'ARN connus

ARN intervenant dans la synthèse des protéines :

- **ARNm** (ARN messager), copie de l'ADN du gène. Lu par un ribosome, il va définir la séquence des acides aminés de la protéine.
- **ARNr** (ARN ribosomal), groupe d'ARN qui s'associent à des protéines pour former les ribosomes où ont lieu la lecture de l'ARNm et l'association des acides aminés apportés par les ARNt.
- **ARNt** (ARN transfert), groupe d'ARN portant un acide aminé et un codon spécifique de cet acide aminé, complémentaire du codon à reconnaître sur l'ARNm. Ils apportent les acides aminés au ribosome dans l'ordre défini par l'ARNm.

ARN non codants régulateurs :

- **siARN** (Small Interfering RNA), petits ARN de 20 nucléotides environ provenant d'ARN dicaténaires découpés par le Dicer.
- **miARN** (microARN), ARN de 20 nucléotides environ découpés par le Dicer dans des zones dicaténaires en épingles à cheveux de long ARN. Ils ont pour rôle de bloquer la traduction des ARNm cibles.

- **piARN** (ARN associés au PIWI), petits ARN de 25 à 30 nucléotides qui sont générés à partir de longs ARN précurseurs. Ils fonctionnent en association avec des types d'argonautes appelés PIWI et sont essentiels, entre autres, pour le développement des cellules germinales.

- lncARN (long ARN non codant), ARN de 70 à plusieurs milliers de nucléotides. Ils interviennent à divers niveaux dans la cellule, probablement aussi dans des phénomènes épigénétiques.

- **ceARN** (Competing Endogenous ARN), long ARN régulateurs, parfois circulaires qui jouent le rôle d'éponges à petits ARN dont ils portent de nombreuses séquences complémentaires.

4-5 - La multiplication des séquences disponibles permet d'évaluer la variabilité des génomes

Les recherches se concentrent surtout sur le génome humain en raison de leurs applications potentielles dans le domaine médical. À partir du moment où la séquence complète du génome humain est devenue disponible, une coopération internationale a été lancée en 2003 et se poursuit de nos jours. Baptisée **ENCODE** (**ENC**yclopedy **O**f **D**na **E**lements) elle a pour but d'identifier tous les éléments fonctionnels présents sur les 3 milliards de nucléotides constituant le génome humain.

En 2008, un projet parallèle de séquençage de 1000 génomes humains a été lancé pour évaluer leur niveau de variabilité. Les résultats commencent à être publiés dans les revues internationales.

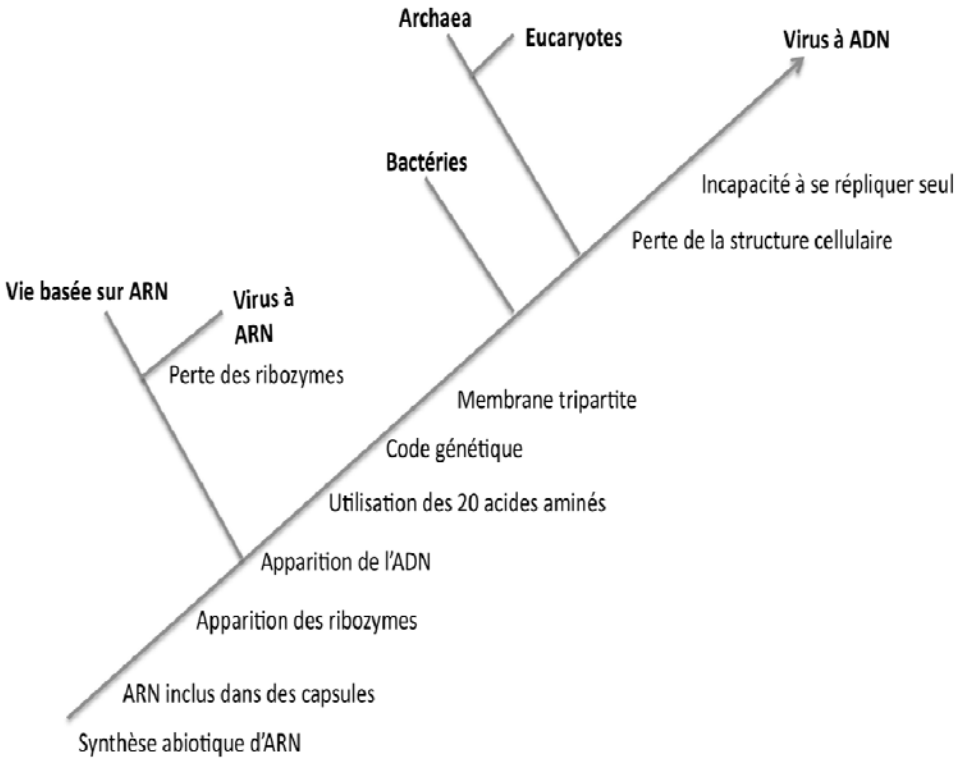
Par exemple, l'analyse des génomes de 1092 individus provenant de 14 populations a permis de dresser une carte de 38 millions de polymorphismes liés à la modification d'un seul nucléotide à un emplacement donné du génome (**SNP ou Single Nucleotide Polymorphism**). Ont été aussi détectées 1,4 millions d'insertions ou délétions de courtes séquences et plus de 14 000 délétions plus importantes mettant en jeu des chaînes plus ou moins longues de nucléotides (1000 genomes project consortium, 2012). Un rôle identifié des SNP paraît être de modifier la conformation spatiale des ARN qui les portent ce qui entraîne une modification des propriétés biologiques de ces ARN (WAN et al., 2014).

À partir de cette foule de données, la comparaison de séquences ADN avec les ARN correspondants identifiés par l'analyse des transcriptomes permet de détecter les variations observables sur les gènes et de les relier à des caractères fonctionnels tels que la prédisposition à certaines maladies, l'origine géographique, etc. (LAPPALAINEN et al., 2013).

Des études associées sont aussi menées sur d'autres modèles biologiques : bactéries, Drosophile, *Caenorhabditis elegans,* ou *Arabidopsis thaliana* et vertébrés.

4-6 - Découverte des ribozymes et hypothèses sur l'ancienneté des ARN

Le rôle d'une enzyme est de catalyser une réaction biochimique pour permettre à celle-ci de se réaliser avec une moindre dépense d'énergie. De très nombreuses protéines d'eucaryotes ou de procaryotes sont connues pour avoir une activité enzymatique. Au cours des trente dernières années, on s'est aperçu que certains ARN sont aussi capables de catalyser des réactions biochimiques.
Nommés **ribozymes**, ces ARN ont des séquences et des structures spatiales particulières. On les a identifiés à des places importantes pour la vie cellulaire telles qu'au sein des ribosomes ou des mécanismes d'excision des introns (DOUDNA et CECH, 2002).



Arbre de la vie proposée par WARD (2006 / 2015) supposant l'apparition d'une vie propre d'organismes à ARN maintenant disparus

L'existence des ribozymes est un indice souvent présenté en faveur d'une origine de la vie basée sur les seuls ARN. En absence de protéines et avant l'apparition des ADN, des ARN auraient pu servir à la fois de biocatalyseurs et de supports capables d'enregistrer de l'information. Des expériences récentes montrent que des ribozymes peuvent avoir une tendance à s'associer en réseaux pour accroître leur efficacité (VAIDYA et al, 2012). Le sujet reste encore conjecturel (WARD et KIRSCHVINK, 2015).

4-7 - L'épigénétique

L'épigénétique se définit comme un ensemble de modifications moléculaires agissant au niveau du génome et de sa régulation et qui peuvent être influencées par la physiologie de la cellule ou par son environnement (HEARD, 2012).
Contrairement aux mutations classiques de l'ADN, ces modifications ne changent pas la nature d'un nucléotide dans la séquence de l'ADN.
Ces modifications peuvent néanmoins se transmettre d'une génération à l'autre. Elles ont aussi la propriété d'être réversibles même après plusieurs générations.
Ces phénomènes existent chez les bactéries, les archées et les eucaryotes.
Le phénomène met en jeu des mécanismes qui peuvent être induits par des ARN non codants, petits ou des longs. On les relie à des méthylations des histones (ajout d'un radical –CH3), ce qui transforme l'euchromatine en hétérochromatine condensée et donc non transcriptible.
Des méthylations peuvent aussi se produire directement sur l'ADN et pourraient entraîner aussi des blocages de transcription.
L'épigénétique pourrait expliquer la spécialisation des cellules en tissus au sein des organismes pluricellulaires et leurs modifications au cours de l'embryogénèse.
Des évènements extérieurs (stress, hautes températures) pourraient aussi entrainer des phénomènes épigénétiques.
Quelques exemples sont bien démontrés tel un changement dans la forme de la fleur d'une Linaire relié à une méthylation bloquant un gène identifié.
Toutefois, les limites de l'épigénétique demandent à être définies. Un certain nombre d'études par corrélation font état de phénomènes épigénétiques sans que la preuve en ait été apportée par des études de causalité (DERMITZAKIS, 2013).

5 – DE NOMBREUSES APPLICATIONS UTILISENT LES CONNAISSANCES OU LES TECHNIQUES DE BIOLOGIE MOLÉCULAIRE

Les connaissances nouvellement acquises en biologie moléculaire fournissent un cadre solide pour définir de façon raisonnée de nouveaux axes de recherche dans de nombreuses disciplines biologiques.

- En phylogénétique et systématique, les premiers séquençages de masse permettent de revoir les relations et la classification des êtres vivants (LECOINTRE et LE GUYADER, 2001 ; DE REVIERS, 2002).
- En médecine, l'identification des maladies d'origine génétique et les possibilités de traitement par thérapie génique ont fortement progressé ces dernières années.
- En police scientifique, l'utilisation de techniques moléculaires est largement médiatisée.
- La transgénèse, contestée dans le cas des OGM cultivés, est utilisée en médecine pour produire des composés biologiques de traitement dans des maladies chroniques via l'utilisation d'animaux modifiés.
- La biologie de synthèse modifie le génome de microorganismes pour les amener à devenir des micro-usines biologiques plus économes en énergie que les processus traditionnels.

- Depuis une dizaine d'années est apparue la métagénomique, un procédé de connaissance du vivant combinant amplification génique (PCR), séquençage à haut débit et analyse bio-informatique des données, qui trouve des applications dans de multiples domaines.

5-1 - La métagénomique et ses multiples applications

La **métagénomique** est une application à succès des techniques de PCR et de séquençage à haut débit développées depuis une dizaine d'années.
Son but est de détecter l'ensemble des variants d'un gène ou d'une séquence donnés qui peuvent être présents dans un échantillon prélevé directement dans un milieu naturel et cela sans avoir à isoler et à cultiver les organismes porteurs de ces gènes (QUIOT, 2013).

En pratique, la métagénomique est basée sur une analyse de l'ADN présent dans un extrait brut de l'échantillon à étudier.
Un gène délimité par des amorces bien choisies va être amplifié par PCR puis séquencé dans la foulée en utilisant des techniques de séquençage à haut rendement (*voir annexes 2 et 3*).
Les données de séquences sont ensuite analysées grâce à des programmes de bio-informatique disponibles sur La Toile pour établir leur niveau de diversité.
Le plus souvent ces données de séquences obtenues expérimentalement sont ensuite complétées avec les séquences de la même zone génomique disponibles dans les banques de données pour bâtir des arbres de relations phylogénétiques. Elles sont aussi utilisées pour comparer des niveaux de diversité en fonction de l'origine ou de l'histoire des échantillons…
Les points délicats sont le mode d'obtention sans biais de tout l'ADN que l'on veut étudier et l'optimisation du choix des amorces.

Les études de métagénomique ont surtout concerné le monde bactérien et elles ont abouti à prendre conscience de l'extrême diversité de ce règne.
Dès 2004, Craig VENTER, après avoir terminé le séquençage de son propre génome, entreprend une croisière dans la mer des Sargasses et collecte des échantillons d'eau de mer de l'ordre de 200 l. Il rassemble les bactéries par passage sur des filtres avec des pores de 0,1 à 0,3 microns puis réalise un séquençage de masse de type « **shotgun** ». Il détecte ainsi 18 000 espèces de bactéries incluant 148 **phylotypes** jusque-là inconnus.
En 2007 il publie dans PLOS une autre étude tout aussi surprenante sur la biodiversité des virus présents dans l'eau de mer.
Au cours des dernières années, des études de ce type se sont généralisées conduisant à une nouvelle vision de la biodiversité microbiologique.
Il est maintenant admis qu'un gramme de sol de jardin renferme couramment plus de cent millions de bactéries appartenant à une centaine d'espèces. Un litre d'eau de mer renferme en moyenne un milliard de particules virales.
Le plus étonnant concerne la biodiversité des bactéries présentes dans le tube digestif humain. QIN et al, 2010, notent que le corps humain contient cent mille milliards de bactéries. En étudiant ce métagénome chez 124 européens, il détecte 3,3 millions de gènes bactériens correspondant à 1000 à 1150 espèces différentes.
Au cours des années suivantes, la baisse du coût des techniques permet d'élargir l'échantillonnage et de répéter ces études dans le temps et dans l'espace. Chez l'homme, ces études sont répétées et permettent de corréler, par exemple, des différences de

diversité bactérienne du tube digestif à des symptômes comme une tendance à l'obésité (LE CHATELIER et al., 2013) ou encore de suivre l'effet de modifications de régimes alimentaires (DAVID et al., 2014).

5-2 - Transferts latéraux de gènes

L'exploitation des données de séquences disponibles dans les banques de données démontre que des gènes en assez grand nombre se retrouvent quasi à l'identique dans des organismes pluricellulaires appartenant à des lignées parfois très éloignées d'un point de vue évolutif. À côté de la transmission de gènes classiques au fil des générations, il doit exister des mécanismes de transferts latéraux de gènes qui ne passent pas forcément par la voie généalogique.

Chez les bactéries, ces transferts latéraux de gènes sont bien connus et les mécanismes ont été identifiés depuis le milieu du XX^{e} siècle (transformation par absorption d'ADN libre dans la milieu, transduction par des bactériophages et conjugaison permettant le transfert d'ADN au travers de « pilus » reliant deux bactéries).

Chez les organismes supérieurs pluricellulaires, ces mécanismes de transfert latéral de gènes sont encore mal identifiés et leur fréquence actuelle peu documentée. Selon SELOSSE, 2011, des phénomènes d'endosymbiose, d'hybridation et de transfert de gènes ont abouti souvent à des fusions de lignées évolutives au cours de l'évolution.

6 - Utilisation de la biologie moléculaire en lichénologie

Depuis une vingtaine d'années, Les concepts et techniques de biologie moléculaire ont permis des avancées dans les connaissances des lichens. On peut en donner quelques exemples non exhaustifs.

- De quand datent les lichens ? L'analyse de séquences permet d'avancer que la plupart des espèces de champignons n'a jamais été lichénisée. Par contre, les espèces lichénisées aujourd'hui l'étaient déjà au niveau ancestral avec quelques pertes ou acquisitions (SCHOCH et al., 2009). Pour obtenir ce résultat, six gènes ont été séquencés sur plus de 420 représentants des principaux groupes de champignons. L'utilisation de programmes de bio-informatique a permis de comparer ces séquences et de bâtir des arbres phylogénétiques concernant différents caractères et, en particulier, la place de l'état lichénisé dans les différentes branches de l'arbre phylogénétique.

- La systématique des lichens intègre maintenant l'analyse des séquences. Au fur et à mesure que des données de séquence sont acquises sur les différentes espèces de lichens, il est possible de réaliser des analyses phylogénétiques dont les résultats sont associés aux données morphologiques et biologiques classiques pour proposer des classifications plus précises. Cette approche est appliquée à la plupart des groupes de lichens et permet parfois de définir de nouveaux genres à l'exemple de la nouvelle classification des Verrucariacées (GUEIDAN et al., 2007, 2009).

La nature même des lichens est mieux appréhendée :

- La présence dans les lichens de bactéries jusque-là inconnues, car non cultivables sur milieux gélosés, est démontrée par des techniques de métagénomique. Certains types bactériens retrouvés sous différents climats paraissent inféodés aux

lichens. L'analyse des séquences montre que certains lignages bactériens de type rhizobiale sont associés aux lichens et possèdent un gène permettant la fixation de l'azote de l'air ce qui peut leur permettre de contribuer à l'alimentation du lichen (HODKINSON et LUTZONI, 2009). En utilisant la technique de SSCP (voir annexe 1) GRUBE et al (2009) montrent que d'autres espèces bactériennes sont associées à des lichens et qu'elles peuvent avoir aussi d'autres fonctions : production d'enzymes lytiques, d'hormone ou de substances antimicrobiennes. L'approche métagénomique permet de comparer de nombreux échantillons et se prête bien aux études écologiques telles que la répartition dans différents environnements des types bactériens associés aux lichens (HODKINSON et al., 2012).

- La partie mycélienne des lichens est aussi analysée à l'aide des techniques de métagénomique ou d'outils moléculaires comme la SSCP (Single Strand Conformation Polymorphism). On constate que le mycosymbiote abrite souvent d'autres espèces fongiques, très différentes ou pouvant provenir en particulier du mycosymbiote d'autres espèces de lichens présentes dans le voisinage. Le lichen n'est pas un monde clos.

- Toujours par SSCP, GRUBE et MUGGIA (2010) détectent aussi une diversité au niveau de l'algue du groupe *Trebouxia* parfois dans les lobes d'un même lichen.

- Une recherche de virus par métagénomique a permis de détecter assez curieusement des virus de plantes supérieures (un Rabdovirus et de l'Apple Mosaic Virus) dans des *Trebouxia* de plusieurs lichens provenant de diverses origines géographiques (PETRZIK et al., 2014). Si ce résultat se confirme, cela suggère que les phycosymbiotes ne sont pas isolés au sein du stroma fongique et peuvent recevoir de l'information génétique.

Au final, les lichens apparaissent comme des structures favorables à l'évolution des organismes qui les composent. LUTZONI et PAGEL (1997) constatent que, comparé à des espèces non lichénisées, le mutualisme accroit le taux d'évolution de l'ADN ribosomal, en particulier chez les champignons lichénisés. DEL CAMPO et al., (2009) analysent la séquence de l'ARN ribosomal 23S de l'algue *Trebouxia* et observent la présence de séquences proches de celles trouvées chez des champignons et des bactéries, ce qu'ils expliquent par des phénomènes de transferts latéraux de gènes. Un phénomène qui peut être mis en rapport avec plusieurs caractéristiques des lichens :

> Les lichens peuvent avoir des longévités très importantes et occupent souvent des sites où les pressions de sélection peuvent varier de façon importante au cours du temps.

> Ils ne possèdent pas de cuticule et peuvent absorber des micro-organismes, pollens ou poussières véhiculés par le vent et la pluie.

> Ils assurent pendant de très longues durées des contacts étroits entre champignons, algues et bactéries et leur structure « collégiale » peut peut-être permettre à des variants de se maintenir, voire de supplanter progressivement les symbiotes d'origine.

On peut sans grand risque dire que l'utilisation des techniques de biologie moléculaire devrait faire encore progresser significativement les connaissances sur les lichens dans les prochaines années.

CONCLUSION

Les travaux réalisés en biologie moléculaire au cours des vingt dernières années apportent une vision nouvelle et parfois surprenante sur la complexité des mécanismes

d'adaptation, de régulation ou de transferts de gènes existant au sein ou entre des cellules vivantes.
L'épigénétique apporte quelques réponses et beaucoup de pistes de réflexion.
La métagénomique se révèle un outil extrêmement puissant pour mettre au jour la biodiversité du monde des microorganismes au sens large et ses interrelations. Elle peut fournir des données jusqu'ici inaccessibles pour des études écologiques et épidémiologiques.
Des techniques performantes, pouvant exiger du matériel coûteux, se généralisent aussi bien dans le monde de la recherche que chez des prestataires de services. Mais, le coût unitaire de beaucoup d'analyses est en forte baisse.
La disponibilité gratuite sur Internet de nombreux articles scientifiques, de gigantesques bases de données et de logiciels d'analyse peut permettre à tout esprit curieux et informé, disposant d'un simple microordinateur, de trouver des réponses à des questions que les chercheurs professionnels n'ont plus le temps d'étudier.
Il devient clair que les naturalistes vont voir leur expérience et leurs connaissances du terrain devenir nécessaires pour exploiter les nouvelles connaissances acquises.
Combien de questions ne demandent qu'à être correctement formulées pour trouver une réponse immédiate, voire pour orienter de nouvelles recherches ?
Les lichens, présents sur terre depuis plus de 440 millions d'années, couvrant plus de 8 % de la surface des terres émergées sous toutes les latitudes, souvent installés en communautés complexes, constituent des modèles d'étude particulièrement attirants.

GLOSSAIRE

Acide aminé (*amino acid*) : constituant élémentaire des protéines. Il en existe 20 types différents chez les êtres vivants. Leur enchaînement dans un ordre précis (ou séquence) est déterminé par la séquence des triplets de l'ADN (donc des gènes).

Adénine : base purique. L'une des 4 bases constitutives de l'ADN ou de l'ARN capable de s'associer à la thymine ou à l'uracyle par des liaisons chimiques faibles.

ADN = acide désoxyribonucléique (*DNA*) : macromolécule, constituant principal et support chimique de l'hérédité. L'ADN se présente sous la forme d'une double hélice : 2 longs filaments anti parallèles unis par des liaisons chimiques faibles. Ces filaments sont formés par l'enchaînement de milliers de nucléotides de types A, T, G, C qui renferment du désoxyribose (voir nucléotides).

ADN polymérase (*DNA polymerase*) : enzyme qui catalyse la polymérisation des nucléotides lors de la réplication de l'ADN.

Allostérie : changement de conformation tridimensionnelle d'une protéine qui modifie ses propriétés biologiques.

Anticorps (*antibody*) : protéine (immunoglobuline) spécifique produite par un animal au contact d'une substance étrangère à l'organisme (antigène). Les anticorps reconnaissent et participent à l'élimination de cet antigène.

Argonaute : protéine du complexe « RISC » qui présente à sa surface un petit ARN et recherche une séquence complémentaire sur un ARN cible à annihiler.

ARN = acide ribonucléique (RNA) : macromolécule constituant principal du transport de l'information dans la cellule et de la régulation cellulaire. Il est constitué par une succession de quatre nucléotides A, U, G, C, qui renferment du ribose (voir nucléotide).

ARN antisens : ARN possédant la séquence complémentaire de tout ou partie d'un ARN messager (ARNm) avec lequel il peut s'apparier.

Cellule eucaryote : cellule possédant un vrai noyau et de nombreux organites subcellulaires. Elle se divise par un processus complexe, la mitose.

Chaperon (*chaperone*) : protéine capable, entre autres, de s'associer à d'autres protéines pour leur donner leur conformation tridimensionnelle.

Clade : selon le principe de classification de HENNIG (1950), le clade est basé sur l'identification de groupes monophylétiques réunissant tous les descendants d'un ancêtre donné.

Code génétique : correspondance entre les triplets de nucléotides (codons) et les 20 acides aminés.

Codon : association de 3 nucléotides successifs codant un acide aminé donné.

Cytosine : base pyrimidique. L'une des 4 bases constitutives de l'ADN ou de l'ARN capable de s'associer à la guanine par des liaisons chimiques faibles.

Dicer : une enzyme de la famille des ribonucléases qui coupe les ARN double-brins en fragments de 21 à 23 nucléotides pour former des miARN ou des siARN impliqués dans la régulation des gènes.

Électrophorèse sur gel (*gel electrophoresis*) : technique qui permet de séparer rapidement un mélange de molécules par leurs migrations différentielles à travers un milieu stationnaire poreux et aqueux (gel) soumis à un champ électrique.

Enzymes de restriction (*restriction enzyme*) : série d'enzymes capables de couper l'ADN au niveau de certaines séquences particulières.

Epigénétique : mécanismes moléculaires intervenant au niveau du génome et de la régulation de son expression sans en modifier la séquence.

Épissage (*splicing*) : processus de maturation d'un ARN messager réalisé par les splicéosomes (ensemble de 300 protéines dans le noyau) entrainant l'excision des introns et le raboutage des exons.

Euchromatine : portions moins condensées de la chromatine, comprenant les régions les plus facilement transcrites.

Exons : portions de gène qui s'associent pour être traduites en protéine fonctionnelle.

Extinction de gène (*gene silencing*) : mécanismes empêchant l'expression de gènes.

Gène : portion d'ADN traduit en protéines caractéristiques d'un génotype donné. La notion de gène est en évolution depuis une quinzaine d'années.

Génotype : ensemble des gènes d'un organisme.

Guanine : base purique. L'une des 4 bases constitutives de l'ADN ou de l'ARN capable de s'associer à la cytosine par des liaisons chimiques faibles.

Hétérochromatine : zone où l'ADN est fortement condensé autour des histones et inaccessible pour la transcription en ARN messager (ARNm).

Histone : protéine présente dans le chromosome des eucaryotes. La chromatine est formée par l'enroulement de l'ADN autour de groupes de 8 histones formant les nucléosomes.

Intron : séquence d'une portion de gène située entre des exons qui peut être excisée lors de la maturation de l'ARN messager (ARNm).

Mitochondrie : organite cellulaire présent chez les eucaryotes formé probablement par endosymbiose et dont le rôle principal est de réaliser le cycle de Krebs qui apporte l'énergie à la cellule.

Mycoplasme : bactérie pléomorphe, ne possédant pas de paroi rigide de peptidoglycanes.

Nucléosome : ensemble formé par de l'ADN dicaténaire enroulé autour d'un groupe de huit protéines, les histones.

Nucléotide : les nucléotides sont des unités formées de l'une des 5 bases (adénine A thymine T, uracile U, guanine G, cytosine C), reliée à un sucre (ribose ou désoxyribose), et à un radical phosphate. L'enchaînement des nucléotides forme un acide nucléique (ADN ou ARN).

Phénotype : ensemble des propriétés observables d'un organisme vivant.

Plasmide : petit ADN circulaire présent dans une bactérie ou une levure et distinct de son chromosome principal. Il se reproduit indépendamment de celui-ci. Il peut se propager d'une bactérie à une autre (transmission génétique horizontale).

Promoteur : séquence d'ADN sur laquelle se fixe l'ARN polymérase, déterminant le site de démarrage de la transcription.

Réplication : mécanisme assurant la reproduction des molécules d'ADN. Elle peut impliquer différentes enzymes (polymérases, topo-isomérases, ligases, etc.).

Rétrotransposon de type viral (*virus-like retrotransposon*) : l'un des 3 grands types de transposons. Cette classe inclut les nombreux rétrovirus insérés dans l'ADN. Ces

transposons peuvent se déplacer vers une nouvelle localisation dans l'ADN en utilisant un intermédiaire transitoire composé d'ARN.

Rétrotransposon poly-A : l'un des 3 grands types de transposons. Ces éléments sont également appelés « les rétrotransposons non viraux ». Ces transposons se déplacent vers une nouvelle localisation dans l'ADN en utilisant un intermédiaire transitoire composé d'ARN.

Riborégulateur (riboswitch) : chez les bactéries : segment à l'extrémité de la séquence non codante d'un ARNm. En présence d'une molécule cible, le riborégulateur adopte une conformation tridimensionnelle qui autorise, ou non, la fabrication de la protéine codée par le reste de l'ARNm.

Ribosome : les ribosomes sont composés d'ARN associés à des protéines. Ils servent de sites d'attachement et de machinerie pour réaliser la synthèse d'une protéine, à partir de la séquence d'un ARNm. La composition des ribosomes est différente chez les eucaryotes d'une part, et chez les bactéries et archées d'autre part.

Ribozyme : molécule d'ARN ayant une activité catalytique.

RISC = complexe d'extinction induit par l'ARN (*RNA Induced-Silencing Complex*) : un complexe protéinique associé à de petits ARN régulateurs qui inactive un ARNm cible en le coupant.

Shotgun : technique de séquençage qui consiste à découper un génome de façon aléatoire puis à séquencer les brins obtenus avec des séquenceurs à haut débit. Un traitement des résultats par des programmes de bio-informatique permet de reconstituer le génome par alignement des séquences en partie chevauchantes.

Sonde (*probe*) : macromolécule d'ADN qui a été marquée (par un isotope radioactif ou par un fluorophore) et qui peut donc être détectée par autoradiographie ou en fluorescence. Les sondes d'ADN sont utilisées pour détecter par hybridation et repérer souvent *in situ* des séquences d'ADN ou d'ARN dont elles sont complémentaires.

Southern blot : technique de biologie moléculaire inventée par Edwin SOUTHERN dans laquelle des fragments d'ADN, produits par une digestion de restriction et séparés par électrophorèse, sont transférés sur une membrane. Certains de ces fragments sont alors identifiés spécifiquement par hybridation avec une sonde d'ADN marquée.

Splicéosome : complexe nucléoprotéique d'environ 300 protéines situé dans le noyau des eucaryotes. Il a pour rôle d'exciser les introns et d'épisser les exons.

Svedberg (S) : unité de mesure de la vitesse de sédimentation de molécules lors d'une ultracentrifugation analytique. Souvent utilisé pour caractériser un acide nucléique, ex : ARN 16S ou ARN 18S.

Thymine : base pyrimidique. L'une des 4 bases constitutives de l'ADN (mais pas de l'ARN) capable de s'associer à l'adénine par des liaisons chimiques faibles.

Topoisomérases : enzymes catalysant une modification du degré de surenroulement des ADN. Ces enzymes se classent en 2 types selon qu'elles agissent sur un seul ou sur les 2 brins de la molécule d'ADN.

Traduction (*translation*): synthèse dans un ribosome d'un polypeptide selon la séquence des codons figurant sur un ARN messager (ARNm).

Transcription : processus permettant le copiage de l'un des 2 brins de l'ADN en ARN.

Transcriptase inverse (RT) (*reverse transcriptase*) : c'est une enzyme ADN polymérase-ARN dépendante, capable de synthétiser un ADN à partir d'une séquence d'ARN. On la trouve, en particulier, dans certains virus comme le VIH, ce qui leur permet d'intégrer une partie au moins de leur génome viral dans le chromosome de la cellule hôte.

Transcriptome : ensemble des molécules d'ARN transcrites, présentes dans une cellule à une phase donnée de son activité ou de son développement. Son étude permet de déterminer les parties du génome qui sont transcrites en ARN sans être traduites en protéines.

Transposon : élément d'ADN qui possède la capacité de changer de localisation dans le génome. On parle aussi d'élément transposable ou d'élément génétique mobile.

Uracile : base pyrimidique. L'une des 4 bases constitutives de l'ARN (mais pas de l'ADN) capable de s'associer à l'adénine par des liaisons chimiques faibles.

BIBLIOGRAPHIE

- Ouvrages de base

- GAUDRIAULT S. et R. VINCENT, 2009, *Génomique*. 129 p., de Boeck éd.
- GROS F., 2012, Les mondes nouveaux de la biologie. 253 pp, Odile Jacob éd.
- LECOINTRE G. et H. LE GUYADER, 2001, Classification phylogénétique du vivant, Tome 1. 543 pp, Belin éd.
- LECOINTRE G. et H. LE GUYADER, 2013, Classification phylogénétique du vivant, Tome 2, 607 pp, Belin éd.
- MADIGAN M. et John MARTINKO, 2007, trad.. 2012, Brock, Biologie des micro-organismes. 1047 pp, Pearson éd.
- Pour la Science, 2013, L'hérédité sans gènes, dossier hors-série N° 81, octobre-décembre 2013, 119 pp.
- RAVEN P.-H., R.-F. EVERT, S.-E. EICHHORN, 2000, *Biologie végétale*. 944 pp, De Boeck éd.
- REVIERS B. de, 2002, *Biologie et phylogénie des algues*, Tome 1, 352 pp, Belin éd.
- SELOSSE M.-A., 2000. *La Symbiose*, 154 pp Vuibert éd.
- WARD P. et J. KIRSCHVINK. 2015, A new history of life. 392 pp, Bloomsbury ed.
- WATSON J. et al., 2012, Biologie moléculaire du gène (6° édition), 688 pp, Pearson ed.

- Références

- AUBEUF D., 2013, Un gène, combien de protéines ? *Pour La Science*, dossier 81 : 36-41.
- BARASH Y., John A. CALARCO et al, 2010, Deciphering the splicing code. *Nature* 465 : 53-59.
- BARRICK J.-E. et R.-R. BREAKER, 2007, The power of riboswitches. *Scientific american*, january 2007 : 36-43.
- BARRICK J.-E. et R.-R. BREAKER, 2014, Les ARN aux commandes. *Pour la science*, dossier 81 : 42-48.
- BUC H. et J. MELLIN, 2011, La régulation des gènes : trois avancées qui ont changé la donne. *Biofutur* 321 : 32-34.
- DAVID L.-A., 2014. Diet rapidly and reproducibly alters the human gut microbiome. *Nature* 505 : 559-563.
- DEL CAMPO E.-M., L.-M. CASANO, F. GASULLA, E. BARRENO. 2009. Presence of multiple group I introns closely related to bacteria and fungi in plastid 23S rRNAs of lichen-forming *Trebouxia*. *International Microbiology* 12 : 59-67.
- DERMITZAKIS E., 2013. Les secrets de la méthylation de l'ADN. Service de communication Université de Genève, www.unige.ch.
- DOUDNA J.-A. et T.-R. CECH, 2002. The chemical repertoire of natural ribozymes, *Nature* 418 : 222-228.
- GROBHANS H. et W. FILIPOWICZ, 2008. The expanding world of small RNAs. *Nature* 451 : 414-416.
- GRUBE M. et al. 2009, Species-specific structural and functional diversity of bactérial communities in lichen symbioses. *ISME Journal* 3 : 1105-1115.
- GRUBE M. et L. MUGGIA. 2010. Identifying algal symbionts in lichen symbioses. In Tools for Identifying Biodiversity, *Progress and Problems*. 296-299. Nimis P.L., Vignes Lebbe R. eds.
- GUEIDAN C. et al., 2009, Generic classification of the *Verrucariaceae* (*Ascomycota*) based on molecular and morphological evidence : recent progress and remaining challenges. *Taxon* 58 (1) : 184-208.
- GUEIDAN C., C. ROUX, F. LUTZONI, 2007. Using a multigen phylogenetic analysis to asseess generic delineation and character evolution in *Verrucariaceae (Verrucariales, ascomycota)*. *Mycological Research* 111 : 1145-1168
- HEAMS T., 2009, Du hasard dans l'expression des gènes. *Pour la Science* 385 : 80-86.
- HEARD E., 2013, Leçons d'épigénétique. Leçons du Collège de France, disponibles sur le Web.
- HODKINSON B. et al. 2012. Photoautotrophic symbiont and geography are major factors affecting highly structured and diverse bacterial communities in the lichen microbiome. *Environmental Microbiology* 14 (1) : 147-161.
- HODKINSON B. et F. LUTZONI, 2009. A microbiotic Survey of lichen-associated bacteria reveals a new lineage from the Rhizobiales. *Symbiosis* 49 : 163-180.
- KHALIL A.-L., 2013, Epigenetic regulation by large non coding DNA. Peanuts, *a biotechnical newsletter* 10 (1) : 4-6.
- KIN J. et al. 2010, Ahuman gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464 : 59-65.
- LAPPALAINEN T. et al, 2013. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501 : 506-511.
- LE CHATELIER E., 2013. Richness of human gut microbiome correlates with metabolitc markers. *Nature* 500 : 541-454.

- LUTZONI F. et P. PAGEL. 1997. Accelerated evolution as a consequence of transition to mutualism. *PNAS* 94 (21) : 11422-11427.
- Mille genomes project consortium, 2012, An integrated map of genetic variation from 1092 human genomes. *Nature* 491 : 56-65.
- MISTELI T., 2011, La vie agitée du génome. *Pour la Science* 403 : 76-83.
- MUGGIA L. et M. GRUBE, 2010. Fungal composition of lichen thalli assessed by single strand conformation polymorphism. *The Lichenologist* 42 (04) : 461.
- PEARSON H., 2006, What is a gene? *Nature* 44 : 399-401.
- PERRIERE G.. 2013, Génomes, protéomes et trancriptions à foison. *Pour la Science* 433 : 38-39.
- PETRZIK K. et al. 2014. Lichens – A new source or yet unknown host of herbaceous viruses ? Eur. J. *Plant Pathol*. 138 (3) : 549-559.
- PIZON-RESTREPO N. et L. MARTINEZ, 2014, Les petits ARN entrent dans l'arène. *Pour la Science*, dossier 81 : 50-56.
- QUIOT J.-B., 2013. Lichens et bactéries. *Bulletin Association Française de Lichénologie* 38 (1) : 140-146.
- RAMOS S.-B.-V. et A. LAEDERBACH, 2014, A second layer of information in RNA. *Nature* 505 : 621-622.
- SCHOCH C.-L. et al., 2009, The *Ascomycota* tree of life : a phylum-wide phylogeny clarifies the origin and evolution of fundamental reproductive and ecological traits. *Syst. Biol*. 58 (2) : 224-239.
- Sean NEE, 2004. More than meets the eye, *Nature* 429 : 804-805.
- SELOSSE M.-A., 2011, L'évolution par fusion. *Pour la Science* 400 : 50-56.
- TAY Y. et al. 2014, The multilayered complexity of ceRNA crosstalk and competition. *Nature* 505 : 344-352.
- TB HANSEN T.-B. et al., 2013. Natural RNA circles function as efficient microRNA sponges. *Nature* 495 : 384-388.
- VAIDYA N. et al., 2012. Spontaneous network formation among cooperative RNA replicators. *Nature* 491 : 72-77.
- VENTER C. et al., 2004. Environnmental genome shotgun sequencing of the Sargasse sea. *Science* 304 : 66-74.
- WAN Y. et al., 2014. Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* 505 : 706-709.
- WILLIAMS T.-A. et al., 2013. An archaeal origin of eukaryotes support only two primary domains of life. *Nature* 504 : 231-236.
- WOESE C.-R. et G.-E. FOX, 1977. Phylogenetic structure of the procaryotic domain : the primary kingdoms, Proceed *PNAS* 74 (11) : 5088-5090.

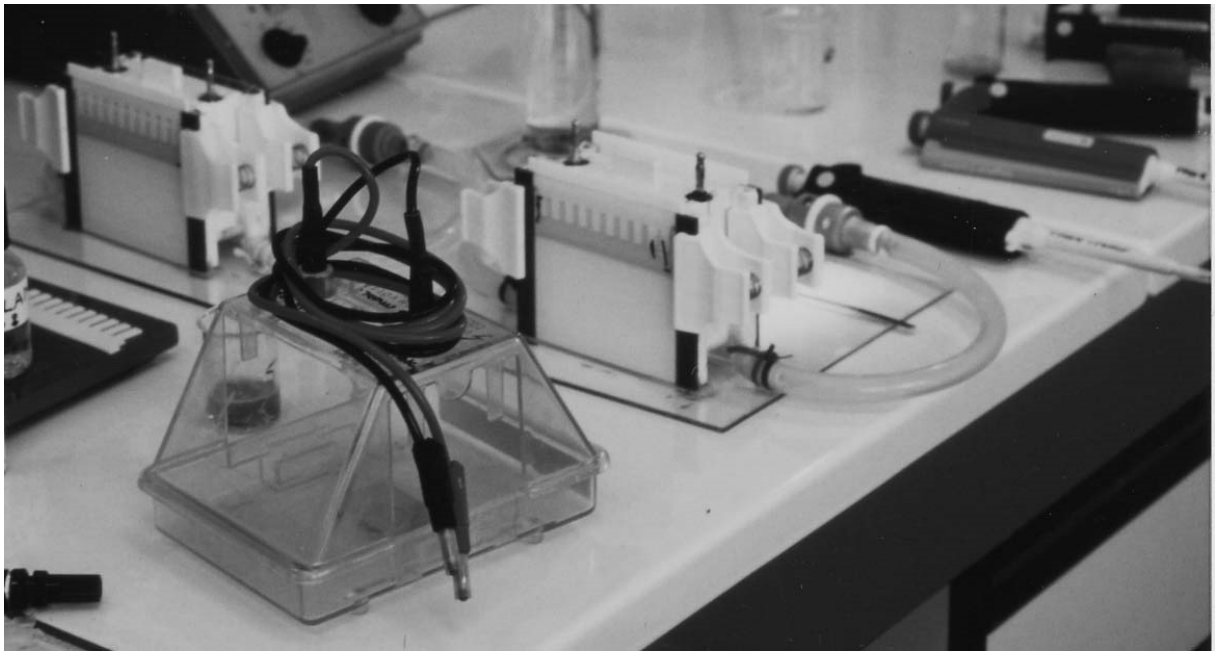
Annexe 1

L'ÉLECTROPHORÈSE ET SES DÉRIVÉS

1°) PRINCIPE DE L'ÉLECTROPHORÈSE

Le but de l'électrophorèse est de trier des molécules biologiques ionisées (protéines, acides nucléiques…) par passage forcé dans les mailles calibrées d'un gel poreux sous l'effet d'un courant électrique.

Après avoir été déposées en haut d'un gel d'agarose ou de polyacrylamide, ces molécules vont être triées en fonction de leur charge électrique, de leur taille et de leur conformation tridimensionnelle.



En électrophorèse verticale, le gel est placé entre deux plaques de verre. On dépose les échantillons dans les puits ménagés au sommet du gel puis on fait passer le courant grâce à des électrodes immergées dans du tampon conducteur. Une circulation d'eau froide permet de réguler la température.

A la fin de la migration, on visualise l'emplacement des échantillons par une coloration *in situ*. On peut aussi les extraire du gel par électrotransfert pour pouvoir réaliser plus facilement des colorations spécifiques d'une molécule donnée.

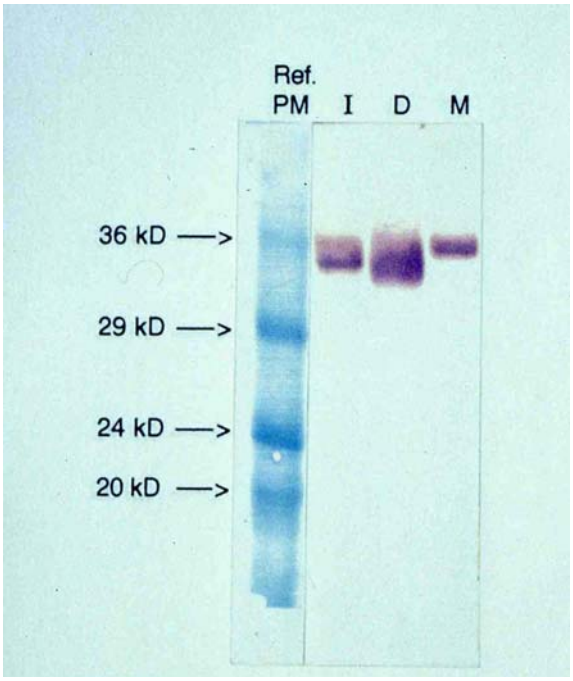
2°) L'ÉLECTROTRANSFERT

Après l'électrophorèse sur gel, si on veut pouvoir réaliser sur les *spots*, des réactions de caractérisation, on réalise un électrotransfert.

L'électrotransfert permet de faire passer les *spots* du gel sur un support plus facile à manipuler (feuille de nitrocellulose ou très fine toile de nylon), tout en conservant les positions respectives qu'ils avaient sur le gel en fin d'électrophorèse.

En pratique, l'électrotransfert consiste à réaliser une nouvelle électrophorèse perpendiculaire à la première. Une feuille de nitrocellulose ou de nylon est appliquée sur le gel et l'ensemble placé entre deux grandes plaques d'électrodes permettant au courant de faire migrer les "*spots*" perpendiculairement au gel tout en limitant la diffusion latérale.

La technique a été mise au point pour l'ADN par Edwin SOUTHERN. Elle prit le nom de « Southern Blot ». Par la suite, les adaptations pour l'ARN et pour les protéines ont été nommées tout naturellement Northern Blot et Western Blot.



Différence de migration entre protéines virales provenant d'extraits bruts de plante. Après l'électrotransfert, la feuille de nitrocellulose a été traitée avec des anticorps spécifiques de la protéine recherchée, puis par un colorant réagissant avec une molécule fixée sur l'anticorps, pour ne faire apparaître que la protéine étudiée. La colonne de gauche contient des références de poids moléculaires. Les autres protéines ne sont pas colorées (cliché Quiot).

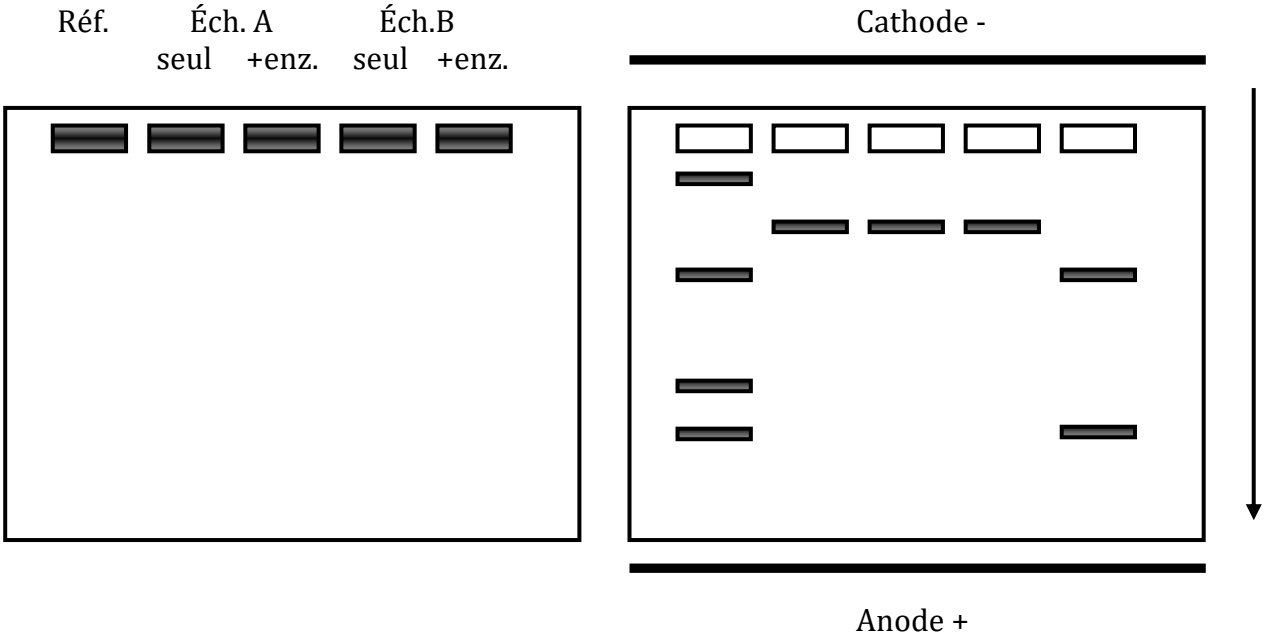
3°) RFLP (Restriction Fragment Length Polymorphism)

C'est une variante de l'électrophorèse, applicable à des ADN, qui a pour but de caractériser ces ADN par la présence de sites de restriction spécifiques.

Sur l'ADN, un site de restriction est une courte séquence de 4 à 8 nucléotides qui va être reconnue et coupée spécifiquement par une « enzyme de restriction » donnée.

Il existe des dizaines d'enzymes de restriction disponibles sur le marché (*exemple : l'enzyme Alu 1 coupe l'ADN en deux fragments lorsqu'elle rencontre la suite de nucléotides AG/CT*).

La recherche de RFLP se fait par une co-électrophorèse permettant de comparer les migrations d'un même échantillon sans et après traitement par l'enzyme de restriction choisie. En combinant des enzymes différentes, il est aussi possible d'établir des cartes de restriction d'un ADN donné.



Après passage du courant et coloration, le nombre et la position des bandes permettent de repérer les ADN porteurs du site de restriction par apparition de bandes multiples. L'échelle de référence (puits 1) permet d'estimer la taille des fragments de restriction. Ici, l'échantillon B possédait le site de restriction recherché car son ADN a été coupé.

Cette technique de caractérisation, simple et fiable, peut être appliquée à des extraits bruts si on dispose d'un moyen de marquage de l'ADN à étudier (sonde spécifique).
Elle peut aussi être appliquée après amplification spécifique d'un ADN donné (*par exemple après PCR*).

4°) SSCP (Single Strand Conformation Polymorphism)

La SSCP est une application de l'électrophorèse, qui permet de caractériser des ADN monocaténaires (simple brin) en fonction de leur forme tridimensionnelle.

À l'état monocaténaire, un ADN donné peut, à cause de zones d'appariement, présenter des boucles, des épingles à cheveux qui lui donnent, dans des conditions données, une forme tridimensionnelle spécifique.

Ces différences de conformations tridimensionnelles vont permettre de les distinguer les uns des autres lorsqu'ils sont soumis à une électrophorèse sur gel de polyacrylamide dans des conditions strictes de pH et de température.

Classiquement, la SSCP comprend trois étapes :
- la production d'ADN amplifiés monocaténaires par PCR « asymétrique » utilisant des quantités très différentes des deux amorces. En fin de cycle, l'un des deux brins sera en excès par rapport à l'autre et, ne pouvant se réapparier, restera sous forme monocaténaire ;
- l'électrophorèse sur gel de polyacrylamide ;
- la détection des bandes se fait par coloration directe du gel ou après électrotransfert. Des différences de distances de migration sont le signe des modifications tridimensionnelles entre les ADN et donc indiquent la présence d'une biodiversité.

Des variantes de la technique permettent de séparer des molécules mutées d'acide nucléique (ADN ou ARN) en les dénaturant au fur et à mesure de leur migration dans le gel. Il s'agit de la DGGE (Denaturing Gradient Gel Electrophoresis) qui se fait sur un gel de polyacrylamide contenant une concentration croissante d'agent dénaturant (ex : urée) et de la TGGE (Temperature Gradient Gel Electrophoresis) qui soumet les échantillons à des températures croissantes. Les deux brins d'acide nucléique se séparent plus ou moins vite en fonction de leur richesse en bases AT et GC (respectivement 2 et 3 liaisons hydrogène) qui ne se dissocient pas à la même température.

Annexe 2

LA PCR (Polymerase Chain Reaction)

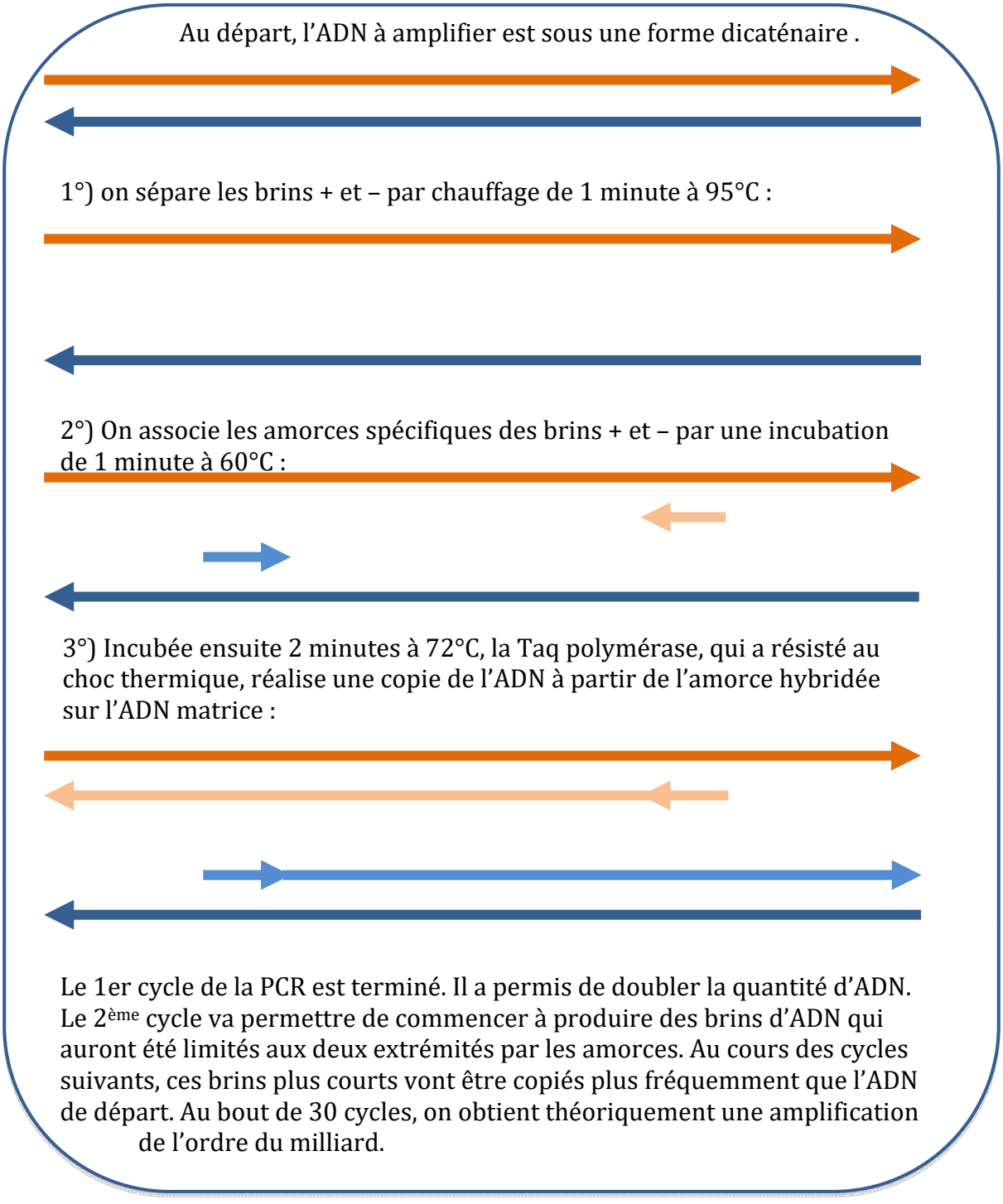
La PCR (Polymerase Chain Reaction) est une technique qui a pour but d'« amplifier » c'est-à-dire de multiplier, à l'identique et en très grande quantité, la séquence d'une portion délimitée d'un ADN. Le taux d'amplification peut être de l'ordre de 1 milliard.

La première PCR est réalisée par K. MULLIS en 1986 (prix Nobel en 1995) qui a bénéficié des travaux essentiels d'autres chercheurs, comme la découverte de l'ADN polymérase ADN dépendante par Kornberg en 1956 (prix Nobel en 1959).

La technique classique repose sur trois caractéristiques originales :

- la découverte et la commercialisation d'une enzyme, la **Taq polymérase**, isolée d'une bactérie (*Thermus aquaticus*) vivant dans les sources chaudes du Parc national de Yellowstone. Cette enzyme est une ADN-polymérase-ADN-dépendante qui a la propriété particulière de résister à des températures de 95°C (demi-vie de 40 minutes). Depuis les années 1990, d'autres polymérases thermorésistantes plus performantes sont utilisées.
- l'utilisation d'**amorce**s pour délimiter la zone de l'ADN à amplifier. Les amorces sont de petites séquences de 20 à 25 nucléotides complémentaires des séquences situées aux bornes de la zone de l'ADN à amplifier. C'est à partir de ces amorces que la Taq polymérase initiera ses copies de l'ADN.

- l'utilisation de cycles répétés permet en jouant sur la température du milieu réactionnel pour, successivement, séparer les ADN dicaténaires, permettre aux amorces de se fixer sur les brins à copier, puis permettre à la Taq polymérase de réaliser sa copie, chacune de ces opérations nécessitant une température différente.



Pour la réalisation de l'amplification, on utilise de très petits tubes (0,5 ml) à parois minces contenant un milieu réactionnel composé de : l'ADN à amplifier, les 2 amorces, la Taq polymérase et des nucléotides en quantité suffisante.

La PCR se déroule ensuite dans un **thermocycleur**, qui est un bain-marie à sec programmable, capable de réaliser en quelques secondes des changements de température d'incubation.

À 95°C, les deux brins d'ADN sont séparés l'un de l'autre ;

ensuite, aux alentours de 60°C (la température dépend de la séquence des amorces) les amorces vont se fixer sur les sites de l'ADN dont elles sont complémentaires ;

enfin, à 72°C, sa température optimale, la taq polymérase va, à partir des amorces, synthétiser un nouveau brin d'ADN, en additionnant le nucléotide complémentaire de celui qu'elle rencontre sur l'ADN qu'elle copie. Ces trois étapes constituent un cycle qui, répété un grand nombre de fois, réalise l'amplification de l'ADN.

La puissance de la PCR est telle (facteur de multiplication de l'ordre du milliard) que des précautions draconiennes doivent être prises lors de l'ouverture des tubes contenant le produit amplifié et pendant toutes les manipulations du produit de l'amplification pour éviter des contaminations de longue durée du laboratoire et du matériel.

La PCR quantitative : une technique plus sophistiquée

Basée sur le même principe d'amplification que la PCR classique, la PCR quantitative (ou PCR en temps réel) permet de suivre, au fur et à mesure des cycles, la quantité de copies d'ADN qui sont produites dans le tube dans lequel a lieu la réaction.

Le suivi de la réaction se fait grâce à un marqueur coloré fluorescent constitué de deux fluorophores associés à un oligonucléotide (Taqman) qui va s'hybrider sur l'ADN devant être copié. La fluorescence reste invisible tant que les deux fluorophores restent reliés par l'oligonucléotide donc proches l'un de l'autre. Le passage de la Taq polymérase entraîne la dissociation de l'oligonucléotide et la séparation des deux fluorophores. La fluorescence devient alors visible dans le milieu réactionnel. Un fluorimètre couplé avec le thermocycleur permet de faire apparaître sur un écran associé, les courbes de fluorescence au fur et à mesure de leur accroissement.

Cette technique permet de comparer la progression de la concentration d'ADN pour différents échantillons dans des tubes séparés. Ces comparaisons peuvent être relatives ou basées sur une courbe de référence construite à partir d'échantillons contenant, au départ, des quantités d'ADN connues.

Cette technique qui existe aussi sous d'autres variantes, est fiable et de plus en plus employée, en particulier en écologie moléculaire.

La PCR est une technique de base qui a révolutionné la biologie moléculaire et est très largement utilisée dans les laboratoires. Elle permet de disposer rapidement et en très grande quantité d'une portion d'ADN que l'on veut pouvoir étudier de façon plus approfondie, par exemple pour le séquencer.

Annexe 3

LE SÉQUENÇAGE

Dans un organisme vivant, l'information nécessaire à son bon fonctionnement et à sa reproduction est portée par l'ADN (acide désoxyribonucléique). L'ADN est stocké dans les chromosomes du noyau sous forme dicaténaire (double-brin), ce qui contribue à sa stabilité. Cette information s'inscrit sous la forme d'une succession de nucléotides attachés les uns à la suite des autres. Un nucléotide est constitué d'une base azotée, d'un sucre et d'un phosphate.

Quatre types de nucléotides, A, T, G, ou C selon la nature de la base azotée, se répartissent sur l'ADN pour former des « codons » de trois nucléotides qui seront lus par la machinerie cellulaire pour fabriquer, par exemple, des protéines. Ces séquences, mesurées en paires de bases, constituent les génomes caractéristiques des êtres vivants - Homme : 3 400 millions de paires de bases (Mpb) ; riz : 389 Mbp ; blé : 17 000 Mbp ; virus de la grippe : 13 000bp ; une diatomée, *Naviculla pelliculosa* : 690 000Mbp ; *Bacillus subtilis* : 4,2 Mbp -

Connaître la séquence, c'est-à-dire l'ordre de succession des nucléotides sur l'ADN, permet d'identifier les gènes de l'organisme étudié et aussi ses systèmes de régulation. C'est le but du séquençage.

Plusieurs techniques de séquençage existent, plus ou moins performantes, plus ou moins rapides, plus ou moins coûteuses. Nous en présentons deux : la méthode classique de Sanger datant de 1977 et la méthode de « pyrosequencing 454 » du laboratoire Roche lancée en 2004.

LA MÉTHODE DE SANGER

Mise au point par F. SANGER en 1977 (prix Nobel en 1980) cette technique permet de séquencer des fragments d'ADN de l'ordre de 300 à 700 nucléotides.

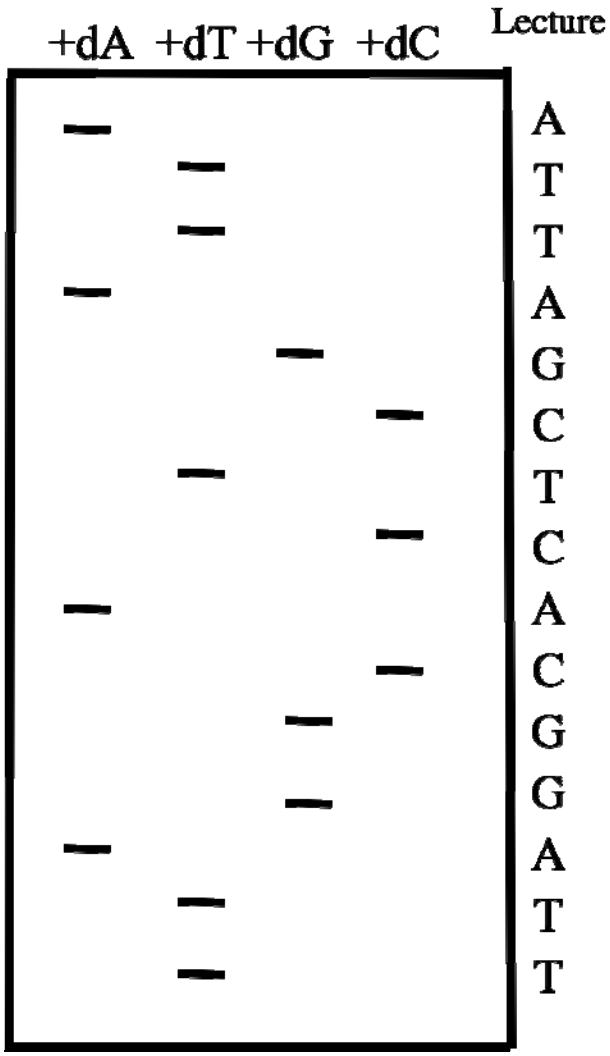
Le principe est de faire réaliser, par une ADN-polymérase-ADN-dépendante, une copie de l'ADN à séquencer en ajoutant dans le mélange réactionnel, en plus des nucléotides normaux (A, T, G et C), une petite proportion de l'un des didéoxynucléotides, (dA, dT, dG ou dC). Ces didéoxinucléotides sont des molécules qui vont remplacer de façon aléatoire le nucléotide correspondant sur la molécule d'ADN en cours d'élongation. Chaque didéoxynucléotide empêche la fixation à sa suite d'un nouveau nucléotide ce qui bloque la poursuite de cette élongation. On obtient donc une famille de séquences de différentes longueurs toutes terminées par le didéoxynucléotide, famille que l'on visualise sur une électrophorèse.

En pratique, on prépare dans quatre tubes un mélange réactionnel contenant : l'ADN à séquencer, l'ADN-polymérase-ADN-dépendante qui va copier cet ADN, un mélange des quatre nucléotides A, T, G et C qui vont permettre de réaliser les copies de l'ADN modèle

et une petite proportion d'un seul des quatre didéoxynucléotides. Pendant l'incubation, la polymérase va générer des familles d'ADN de différentes longueurs.

En soumettant ensuite ces mélanges réactionnels à une électrophorèse, on obtient une séparation des copies d'ADN en fonction de leurs longueurs (les fragments les plus courts migrant le plus loin).

On place côte à côte, sur un gel de polyacrylamide, les 4 mélanges contenant chacun un seul des 4 didéoxynucléotides et on réalise une électrophorèse en parallèle.



Après révélation des bandes, on peut lire, directement sur le gel, la succession des nucléotides qui constituaient l'ADN à séquencer. Soit TTAGGCACTCGATTA dans l'exemple.

Dans les années 1980, on utilisait des didéoxynucléotides marqués au phosphore radioactif pour accroître la sensibilité de la méthode. En plaçant une plaque photo contre le gel, on obtenait, après révélation, une photo du gel plus facile à lire et à conserver.

Au cours des années 1990, la technique a été modernisée et simplifiée par l'utilisation de didéoxynucléotides colorés avec 4 couleurs différentes, ce qui a supprimé le marquage radioactif. De plus, les 4 didéoxynucléotides pouvaient être placés ensemble dans un seul tube, la couleur signalant le type de nucléotide. Après électrophorèse, la succession des taches colorées sur la même colonne du gel est lue par un colorimètre, ce qui a permis d'automatiser la technique.

LES SEQUENÇAGES À HAUT DÉBIT :

Au cours des années 2000, de nouvelles techniques de séquençage en parallèle sont apparues, qui apportent des modifications très importantes en matière de rapidité et de coût.

Le prix du matériel est très élevé, pouvant dépasser les 300 000 € et le prix des produits consommables est aussi très élevé. Par contre, le très grand rendement de ces

techniques permet de ramener le prix de revient du séquençage d'un ADN de 1 million de bases à un coût près de 1 000 fois inférieur à celui du séquençage traditionnel.
La possibilité, dans un proche avenir, de réaliser le séquençage complet d'un organisme vivant en une semaine et pour moins de 1 000 € peut être sérieusement envisagé, ainsi que des études à grande échelle de variabilité entre individus.

Trois techniques sont implantées sur le marché : le système 454 de Roche, la technique illumina et la technique SOLiD de Applied Biosystems . Leurs performances sont résumées ci-dessous :

	Technique Sanger	454 Roche	illumina	SOLiD
Million de bases lues / essai	0,0006	100	1 300	3 000
Durée de l'essai	1 à 3 jours	10 h	4 jours	5 jours
Longueur des séquences lues	650 bp	250 bp	32 à 40 bp	35 bp
Coût d'un essai	environ 50 $	8 000 $	9 000 $	17 000 $
Coût de 1 million de bases lues	env. 500 000 $	env. 5 $	5,97 $	5,81 $

Adapté de MARDIS, 2008, Trends in genetics 24 (3) : 133-141

LA TECHNIQUE 454 DU LABORATOIRE ROCHE

La technique 454, la plus versatile en raison de la taille des fragments séquencés, est basée sur le « pyroséquençage ». Elle consiste à utiliser les phosphates libérés par les nucléotides au cours de la polymérisation de l'ADN pour déclencher un jeu de réactions enzymatiques couplées fabriquant de l'ATP (adénosine triphosphate) et activant une enzyme, la luciférase. Cette dernière transforme de la luciférine ajoutée au milieu réactionnel en oxyluciférine avec émission de luminescence qui peut être captée par une caméra CCD.

En pratique et de façon simplifiée, la technique 454 peut être divisée en 5 étapes :
1°) Préparation de l'échantillon : l'ADN en double hélice est coupé en fragments double brin plus courts d'environ 400 à 600 bp ; les fragments sont rendus monocaténaires par séparation des deux brins ; des adaptateurs A et B (oligonucléotides) sont attachés chimiquement aux fragments d'ADN.

2°) Formation de microréacteurs : les fragments d'ADN, une solution aqueuse contenant les réactifs de PCR et des microbilles d'agarose sont introduits dans des microtubes contenant une huile synthétique. Les concentrations des ADN et des microbilles sont calculées de façon à ce que chaque microbille ne fixe qu'un fragment de l'ADN à amplifier. Le mélange est agité vigoureusement de façon à former une émulsion. Les gouttelettes de réactifs enveloppant les microbilles se retrouvent isolées les unes des autres par l'émulsion huileuse. Chaque microbille se trouve ainsi isolée avec un ADN à amplifier et une gouttelette de réactifs (polymérase et nucléotides), ce qui constitue un microréacteur.

3°) Amplification par PCR sous émulsion (EmPCR) : les microbilles sont soumises aux successions thermiques d'une PCR. Chaque brin d'ADN est ainsi amplifié pour former des millions de brins identiques fixés sur la même bille. L'émulsion est ensuite cassée pour libérer les billes.

4°) Détermination de la séquence amplifiée : la séquence se révèle en identifiant un à un les nucléotides polymérisés. Pour cela, une ADN-polymérase-ADN-dépendante se déplace le long des ADN amplifiés toujours fixés sur les billes d'agarose et rendus monocaténaires.
En pratique, les billes viennent s'insérer sur une « picotiterplate» brevetée, constituée par les extrémités de centaines de milliers de fibres optiques accolées. Chaque extrémité de fibre ne peut recevoir qu'une seule microbille. Les enzymes et réactifs nécessaires à la réaction luminescente de pyroséquençage sont apportés sur la picotiterplate. Les nucléotides A, T, G ou C, nécessaires pour réaliser une copie de l'ADN amplifié, sont apportés un à un de façon séquentielle. À chaque étape (par exemple apport de T), la caméra en bout des fibres optiques enregistre la luminescence éventuelle déclenchée par la luciférase s'il y a eu incorporation du nucléotide par une microbille donnée. La force du pic de luminosité permet de savoir si 1 ou 2 T ont été incorporés en cas de répétition sur la séquence cible. Entre chaque apport de nucléotide, un traitement enzymatique élimine l'excès du nucléotide précédent.

5°) Analyse informatique des données : en fin d'expérience, le traitement informatique de l'ensemble des clichés pris par la caméra permet de générer les séquences nucléotidiques présentes sur l'ensemble des centaines de milliers de microbilles fixées sur la picotiterplate. La technique est considérée comme capable de fournir, en un seul « *run* » de 10 heures, plusieurs centaines de milliers de séquences utilisables, d'une longueur de l'ordre de 250 bases.

Des programmes informatiques complémentaires peuvent permettre de reconstituer des génomes par recouvrement des séquences des différents fragments amplifiés. C'est aussi une technique qui, combinée avec d'autres programmes informatiques, peut permettre de réaliser rapidement des études de métagénomique visant à connaître la biodiversité d'un gène donné dans un échantillon renfermant des microorganismes non connus au départ.

De nos jours la compétition entre les systèmes de séquençage continue, ce qui laisse espérer encore des baisses de prix du nucléotide séquencé.