Académie de Stanislas – 25 mars 2022

Découverte du code génétique et quelques conséquences sur la lecture des génomes

Jean-Pierre Jacquot

Dans le domaine des sciences, on est souvent confronté à l'incertitude, en particulier en physique (voir le principe d'incertitude de Heisenberg et la théorie de la relativité et ses conséquences sur le temps), mais aussi toutes les mesures physiques sont sujettes à une incertitude plus ou moins grande liée à l'expérimentation. Il n'en demeure pas moins qu'un esprit cartésien préfère se baser sur des certitudes absolues. Dans ce registre, nous pouvons mentionner la vitesse de la lumière dans le vide (3.10^8 m/s), le nombre d'Avogadro en chimie ($6{,}02\ 10^{23}$ atomes dans une mole), la température minimale absolue (-273°K), etc. En chimie, l'équivalent des tables de la loi pour les Hébreux s'appelle le tableau périodique des éléments de Mendeleïev. Ce chimiste russe exceptionnel a su faire sens de milliers d'expériences effectuées avant lui (en partie par Lavoisier) et a élaboré une classification des éléments chimiques suivant leur masse, leurs propriétés électroniques et nucléaires et leur réactivité. Ce tableau a été vérifié des milliards de fois expérimentalement et constitue un socle extrêmement solide et inviolable de la chimie moderne.

En biologie, l'équivalent du tableau de Mendeleïev s'appelle le code génétique. L'objet de cette communication est de vous décrire la signification de ce code qui permet de relier les séquences des gènes à celles de protéines qui en découlent et la façon dont il a été déchiffré. Je montrerai en conclusion que les expériences initiales qui ont mis à jour le code méritent d'être revisitées à la lumière de développements plus récents qui pourraient conduire à des lectures alternatives des génomes.

Pour mettre en perspective cette découverte du code génétique, il convient tout d'abord de rappeler quelques notions essentielles en biologie. Les êtres vivants contiennent quatre catégories de macromolécules : les glucides, les lipides, les protéines et les acides nucléiques. Pour cette communication, seuls les protéines et acides nucléiques nous intéresseront. Un petit mot préalable sur l'importance de ces deux catégories de macromolécules. Les « protéines » comportent des protéines de structure essentielles au maintien de la conformation de nos cellules mais aussi les enzymes nécessaires pour effectuer toutes les réactions chimiques de la cellule dans des conditions de pression et de température acceptables pour les êtres vivants. Les acides nucléiques (ADN et ARN pour l'essentiel) sont les supports de l'hérédité et nécessaires en particulier pour la synthèse des protéines.

Le cheminement expérimental et intellectuel pour déterminer que les protéines sont codées par des gènes présents dans l'ADN (lui-même restreint au noyau des cellules eucaryotes), que ces gènes sont transcrits en ARNm (ARN messager) transféré au niveau du cytosol des cellules puis traduits en protéines au moyen du système ribosomal, est très complexe[1]. Je relaterai ici seulement les expériences ayant servi à décoder les séquences nucléotidiques et à les relier aux séquences protéiques, un travail qui n'est pas sans évoquer celui de Champollion pour déchiffrer les hiéroglyphes.

[1] De nombreuses représentations de ces réactions de transcription (ADN vers ARNm) et de traduction (ARNm vers protéines) sont disponibles en accès gratuit sur le web, par exemple à l'adresse : https://www.nature.com/scitable/topicpage/translation-dna-to-mrna-to-protein-393/, ou encore, pour une version animée extrêmement remarquable : https://www.youtube.com/watch?v=gG7uCskUOrA.

Quelques explications préalables sur la nature des acides nucléiques et protéines

Pour comprendre ces expériences, il nous faut approfondir un petit peu nos connaissances sur les acides nucléiques et les protéines. Il est de notoriété publique que la molécule de l'ADN est une double hélice avec deux brins antiparallèles et, en schématisant à l'extrême, on peut décrire ces séquences comme une succession en apparence aléatoire de quatre bases nucléotidiques : adénine (A), cytosine (C), guanine (G) et thymine (T). Une règle de base est que dans l'ADN un A fait toujours face à un T et vice-versa, et un G face à un C. En conséquence, une séquence d'ADN peut être schématisée de la façon suivante (nous donnerons à la fin de cet article la signification protéique de cette séquence apparemment aléatoire mais qui a été choisie avec soin et constitue un clin d'œil à l'actuel président de l'académie Denis Grandjean, mais vous pouvez la déduire par vous-même en utilisant le code génétique dans le tableau 2 en traduisant le brin codant) :

Brin « codant » :

5'ATGGCGATTAGCGATGAAAACATTAGCTGCGAAAGCACCCTGGAACCGGCGCG
CTTTGCGATTACCGCGTGCGCGGATGAAATGATTTGCATTGAAAAC3'

Et son brin complémentaire :

5'GTTTTCAATGCAAATCATTTCATCCGCGCACGCGGTAATCGCAAAGCGCGCCGG
TTCCAGGGTGCTTTCGCAGCTAATGTTTTCATCGCTAATCGCCAT3'

L'ARN diffère de l'ADN en deux aspects : la base adénine est remplacée par un uracile (U) et la molécule est simple brin, il n'y a pas de brin complémentaire comme dans l'ADN. Les protéines en revanche sont constituées très différemment de l'ADN ou l'ARN, elles peuvent être définies très simplement comme des enchaînements d'acides aminés liés les uns aux autres lors de la synthèse protéique. On trouve dans les protéines vingt acides aminés différents dont la liste est fournie dans le tableau 1. Remarquons qu'il y a dans la nature bien plus de vingt acides aminés, mais seuls vingt d'entre eux sont utilisés de façon routinière pour constituer les protéines et ceci dans tous les êtres vivants.

Alanine	A	Leucine	L
Arginine	R	Lysine	K
Asparagine	N	Méthionine	M
Aspartate	D	Phénylalanine	F
Cystéine	C	Proline	P
Glutamate	E	Sérine	S
Glutamine	Q	Thréonine	T
Glycine	G	Tryptophane	W
Histidine	H	Tyrosine	Y
Isoleucine	I	Valine	V

Tableau 1 : les 20 acides aminés constitutifs des protéines

La découverte du code génétique par Nirenberg et Khorana

Près de quatre-vingt ans plus tard, il est difficile de déterminer quels indices ont amené les chercheurs en biologie à postuler qu'il y avait une connexion entre l'ADN, l'ARN et les protéines, mais cette notion était apparemment déjà présente vers la fin des années 1950. On subodorait donc qu'il y avait un message codé dans l'ADN/ARN qui devait se traduire dans la

séquence des acides aminés d'une protéine. Le problème était le suivant : dans les acides nucléiques il n'y a que 4 bases (A, C, G, T pour l'ADN, et U, C, G, T pour l'ARN) mais les combinaisons de ces quatre bases doivent coder pour vingt acides aminés différents. Il est fort à parier que les chercheurs qui se sont lancés dans le déchiffrage du code génétique sont partis de considérations mathématiques :
1. Il est impossible que l'information soit contenue dans une seule base (4 possibilités seulement pour 20 acides aminés.
2. Si l'information est contenue dans des doublets de bases, les possibilités de codage sont 4^2 soit 16 possibilités, donc un peu courtes pour 20 acides aminés, à moins d'envisager des mécanismes de « rattrapage ».
3. Si l'information est contenue dans des triplets de bases, les possibilités de codage sont de 4^3 soit 64 triplets. Vu qu'il n'y a que 20 acides aminés, cela revient à dire que le code génétique doit être en partie dégénéré, *ie* plusieurs codons peuvent signifier le même acide aminé.
4. Si l'information est codée par des paquets de 4 bases, la possibilité théorique est de 4^4 soit 256, une valeur qui paraît maintenant très élevée au regard du nombre d'acides aminés utilisés en routine.

Les groupes de recherche qui se sont attelés au déchiffrage du code génétique sont les professeurs Marshall Nirenberg[2] et Gobind Khorana[3] et leurs laboratoires respectifs aux États-Unis (NIH Bethesda, University of Wisconsin Madison). Pour ce travail, ils ont obtenu le prix Nobel de Physiologie et Médecine[4]. Nirenberg et ses collègues étaient des biologistes qui savaient comment isoler les ribosomes et préparer des extraits acellulaires possédant tous les composants nécessaires à la synthèse protéique. Khorana en revanche était un chimiste de haut vol capable de synthétiser chimiquement des acides nucléiques. A cette époque il ne devait pas y avoir beaucoup de laboratoires dans le monde possédant cette capacité… Dans un premier temps, Nirenberg et Khorana ont joint leurs connaissances respectives pour déterminer quel acide aminé est incorporé en présence d'une séquence ARN poly U (par exemple 5' UUUUUUUUUUUUUUU 3'). Evidemment Khorana s'est occupé de la synthèse et Nirenberg a mis au point la suite de l'expérience. Elle se déroulait comme suit : il a préparé vingt réactions différentes ; dans chacune il a introduit le poly U de son collègue, le système de traduction acellulaire et 19 acides aminés « froids » (*ie* non radioactifs) et seulement un acide aminé chaud (radioactif) en variant cet acide aminé dans chacun des tubes. Après un temps d'incubation suffisant, les protéines ont été précipitées dans chacun des essais, et la radioactivité a été comptée. Je peux à peine imaginer l'excitation de ces chercheurs lorsqu'ils se sont aperçus qu'un seul acide aminé était incorporé dans ces conditions, la phénylalanine (F). La suite était évidemment de construire des ARN synthétiques avec des séquences différentes et, pour citer un exemple, ils ont produit des séquences poly G qui induisaient l'incorporation de glycine (G), puis des répétitions du codon (triplet) GGA ou du codon GGT ou du codon GGC, ces quatre séquences permettant l'incorporation de la glycine. La fièvre qui les a animés pour ensuite tester les 64 codons possibles a dû être particulièrement intense ! Ces expériences ont été essentielles pour comprendre la signification des séquences dans l'ADN, et donc bien plus tard pour interpréter l'analyse des génomes. Avant d'aller plus loin, je voudrais souligner la prouesse technique et intellectuelle réalisée par ces chercheurs et leur impact sur la biologie moléculaire d'aujourd'hui. J'ai déjà mentionné que la génération d'ARN synthétiques était une performance remarquable dans les années 60, il est également important de faire remarquer que l'utilisation des isotopes radioactifs était loin d'être universelle à l'époque, seuls quelques

[2] https://fr.wikipedia.org/wiki/Marshall_Warren_Nirenberg.
[3] https://fr.wikipedia.org/wiki/Har_Gobind_Khorana.
[4] Marshall W. Nirenberg - Biographical (nobelprize.org).

grands centres de recherche atomique étaient capables de produire et purifier ces éléments (Oak Ridge, Lawrence Berkeley Laboratories aux USA, Cadarache et Saclay en France à une bien plus petite échelle). En d'autres termes, ces expériences ne pouvaient pas être faites n'importe où et c'est un euphémisme… Leur immense signification a été immédiatement comprise par la communauté scientifique, et il se dit que lorsque Nirenberg a rapporté ces premiers résultats dans un important congrès de biochimie en 1961, un des participants influents s'est levé pour lui donner l'accolade (Marshall Warren Nirenberg - Wikipédia). Assez curieusement aucun événement de la sorte ne s'est produit lors de ma propre carrière scientifique, c'est un peu décevant, mais toutefois à mon petit niveau dans la rubrique émotion de la découverte, je vous invite à visionner la séquence *YouTube* préparée par la faculté des sciences à l'Université de Lorraine La régulation de la photosynthèse grâce aux thiorédoxines[5]. Ce témoignage décrit la découverte des thiorédoxines et de leur rôle dans la photosynthèse, le gène codant pour cette protéine est donné en exemple dans la figure 2.

Le code génétique a été vérifié des millions de fois au travers de l'analyse de milliers de génomes maintenant, et il a brillamment résisté à l'épreuve du temps au cours de ces soixante dernières années. De plus, on s'est aperçu qu'il est universel, c'est-à-dire qu'il est le même pour toutes les espèces allant des bactéries comme *Escherichia coli* jusqu'à d'autres espèces modèles comme *Homo sapiens* pour le règne animal ou *Arabidopsis thaliana* pour le règne végétal. Des rapports sporadiques ont indiqué quelques variations dans l'ADN mitochondrial, mais elles semblent résulter de transformations discrètes au niveau de l'ARN (A et G sont des bases très proches dans leur structure ainsi que C et T) et dans certains cas, une petite modification chimique peut transformer l'une en l'autre. Le code génétique au niveau de l'ADN est montré dans ce document dans le tableau 2.

Définition moderne du gène en biologie moléculaire suite aux expériences de Nirenberg et Khorana

Évidemment un travail immense a été effectué depuis ces expériences remarquables en se basant sur des données de biochimie et sur l'analyse des séquences d'ADNc (ADN complémentaires, un équivalent de l'ARNm mais plus facilement utilisable pour les techniques de génie génétique). On a abouti au consensus suivant : un gène est une séquence ininterrompue de triplets (codons) (donc un multiple de 3) située entre un codon initiateur de la traduction (ATG, toutes les protéines commençant par une méthionine (M) et l'un des trois codons Stop du code génétique. Pour qu'un gène soit totalement fonctionnel, il est nécessaire qu'il y ait des séquences régulatrices relativement conservées en amont de l'ATG et en aval du Stop (*cf.* figure 1).

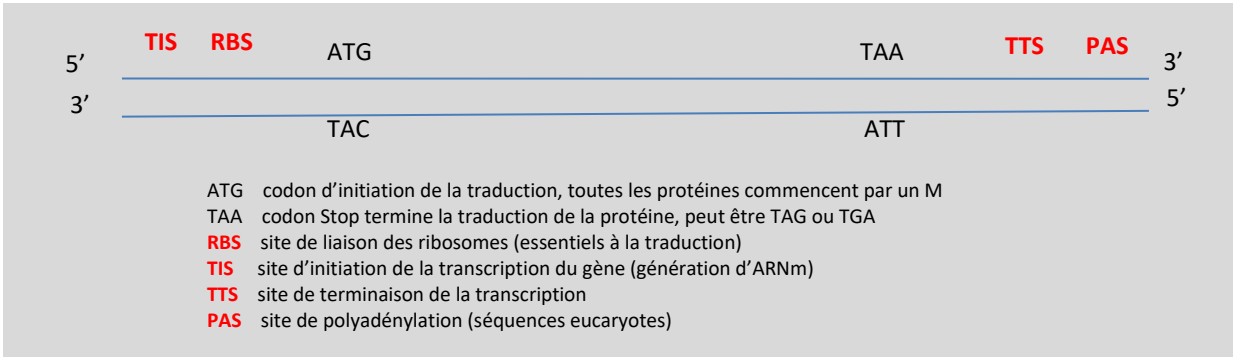

Figure 1 : structure schématique d'un gène

[5] [Eurêka !] – YouTube.

Ala	A	**GCT, GCA, GCC, GCG**
Arg	R	**CGT, CGA, AGA, CGC, CGG, AGG**
Asp	D	**GAT, GAC**
Asn	N	**AAT, AAC**
Cys	C	**TGT, TGC**
Gln	Q	**CAA, CAG**
Glu	E	**GAA, GAG**
Gly	G	**GGT, GGA, GGC, GGG**
His	H	**CAT, CAC**
Ile	I	**ATT, ATC, ATA**
LEU	L	**CTT, CTA, TTA, CTC, TTG, CTG**
LYS	K	**AAA, AAG**
MET	M	**ATG**
PHE	F	**TTC, TTT**
PRO	P	**CCA, CCC, CCG, CCT**
SER	S	**AGT, TCT, TCG, TCA, TCC, AGC**
THR	T	**ACT, ACC, ACA, ACG**
TRP	W	**TGG**
TYR	Y	**TAC, TAT**
VAL	V	**GTT, GTG, GTC, GTA**
STOP		**TAA, TAG, TGA**

Tableau 2 : Les 64 codons (triplets) du code génétique

De nos jours, les programmes de recherche de gènes dans les séquences génomiques brutes sont automatisés, mais ils sont tous basés sur le code génétique de Nirenberg et Khorana, et ils recherchent essentiellement des phases de lecture (*ie* des triplets de bases situés entre un ATG, le codon d'initiation de la lecture et un codon stop). La complexité de ces recherches est due au fait que l'on ne connaît pas *a priori* la phase dans laquelle la séquence a une signification (doit-on commencer la lecture à partir de la première base ou de la seconde ou de la troisième ?), ni même sur quel brin de l'ADN se situe la séquence qui fait sens au niveau protéique. Ces programmes traduisent donc dans six phases possibles (3 phases sur chaque brin d'ADN). Un niveau de complexité supplémentaire, seulement chez les eucaryotes (organismes dont les cellules possèdent un noyau), est que les gènes sont le plus souvent morcelés avec des séquences ayant un sens (exons) et d'autres appelées introns incorporées au milieu des exons et qui n'ont pas de sens au niveau protéique. Pour l'analyse des génomes, il est donc essentiel de se baser sur les critères énoncés précédemment, mais aussi sur toutes les données de biochimie (étude directe des protéines) et de séquençage des

ADNc (équivalent de l'ARN) pour positionner ces introns. Pour expliciter les difficultés de l'analyse des génomes, au cours du temps, le génome humain a été estimé tout d'abord contenir 50 000 gènes codant pour des protéines et, au fil du temps, ce nombre a diminué de moitié. Les estimations les plus récentes indiquent 21 000 gènes dans le génome humain comparé à 33 000 gènes dans le maïs et plus de 90 000 dans le blé, une situation relativement humiliante pour notre espèce[6]. Cette différence peut s'expliquer en partie par la plasticité métabolique des plantes qui sont autotrophes vis à vis du carbone (photosynthèse), de l'azote et du soufre, alors que nous ne le sommes pas. Un exemple de séquence génomique isolée dans les travaux de mon laboratoire à Orsay est fourni dans la figure 2. Cette séquence de près de 1600 bases code finalement pour une petite protéine chloroplastique de 106 acides aminés appelée thiorédoxine qui est impliquée dans la régulation de la photosynthèse (Stein *et al*, 1995). Dans cette figure, vous pourrez voir les exons en rouge et vert et l'intron positionné entre les deux exons en bleu.

AGGGTAGTGA	GTGCGCGTGC	CAACATGCGC	TGGGTGTCAA	AGTGCCGACT	ACCCGCCCTC	61
GCCCTCCCAG	TGCATGCCTC	TGAAACGTTT	GAGCAATGCA	CTGTGTACAT	CTTGACATGC	121
AAACCTGGTT	CTGGGCTAGA	AACGGTTCTG	GGTCGCTTTG	GATCGAGCCC	TTGCCCAGAT	181
GTCCAGGGGC	TGCAAAGCAA	AAGCATCCCT	CGGGCACCTG	ATCAGCCCTG	TCGAGCACCT	241
GATCCGATAC	TGGAAAACAC	ACAGCTGGTA	TCGCCAGTGT	GGGGGGAGGC	AGAAACGGGA	301
CATGCGTGGC	GAAGCAAACA	GAGCTGCGAG	CGAACCTCTA	GGCCTTTGGT	GCTCTGACGT	361
TGCTATGTAT	TGAGCGCTCG	GAGGCCAGCA	AGCCACACAC	CTCCATTCCC	AGGGAATGGG	421
AACCCATGCG	ATAGGGCCCG	CAAGACGCTC	TACTACGATT	TTGTTTCTAC	TCGTGCTCCC	481
TGATCATCTA	GCTACCCTAC	TCTTGTTTAG	TTCACACTGT	TGCTAAAAAT	GGCCCTCGTT	541
GCTCGCCGCG	CCGCTGTGCC	CTCGGCCCGC	TCCAGCGCTC	GCCCCGCTTT	CGCTCGCGCT	601
GCCCCGCGCC	GCTCGGTCGT	TGTGCGCGCC	GAGGCCGGTG	CCGTGAACGA	TGATACCTTC	661
AAGAACGTTG	TTCTGGAGAG	CAGCGTGCCC	GTCCTTGTGG	ATTTCTGGGC	TCCCTGGTGC	721
GGTCCTTGCC	GCATCATCGC	CCCCGTCGTC	GACGAGATTG	CGGGCGAGTA	CAAGGATAAG	781
CTCAAGTGCG	TGAAGCTGAA	CACGGACGAG	AGCCCCAACG	TGGCGTCCGA	GTACGGTATC	841
CGCTCCATCC	CCACCATCAT	GGTGAGGCGG	GGTACTGTGA	CCAGGCATGC	GAGAATCATG	901
TTGCTTGCTG	TCGCCGGGGC	CAGCTGACCA	GCCGCGGAGA	AGTGGGATAG	AGTTATAAGA	961
CAACAATTAT	AGTTCATTCG	CGCCAGCTCG	TTTTTCAACT	GGCCTGGTAA	ACGTCGGTCG	1021
CCCACAGGTG	TTCAAGGGTG	GCAAGAAGTG	CGAGACGATC	ATTGGCGCTG	TGCCCAAGGC	1081
GACCATCGTG	CAGACCGTGG	AGAAGTACCT	GAACTAAGCA	GCCTGGCGCA	GGCAATTCAG	1141
CCGGCGGGAG	TGGGGTTCCC	CGCTCGCCCA	TCCGCGGCGG	AGCGATGTGA	CAGGGGCACA	1201
GTAGCACACG	GGACTAGGGG	GCACCACACC	CAGAAGGGTG	CGGCCGCATT	GTTTCGTTTC	1261
AAGGGGACGC	TTGCCACCGT	ATGGGTGGCA	GTTTTGTAAG	CACTTAAATA	AGTAGGAGCA	1321
CGGCCCCCCG	GCCGGTGGGT	TCGTGTGACG	GACCGGGGCC	AGTGTGTGTC	GTGAACGTTT	1381
TGGACCCCGG	GCCGCCATTC	GGCCTGAGTT	TTGGAACGCT	TTTGGATGAA	GCGGAAAACG	1441
GATTGTGATT	GAACAATTAT	CGGGGTTGGC	CCATGGCAAA	GTGCATGGAG	CCGCGAGTGA	1501
GCGAGTAGTG	ACGTCCGGAG	CTGTCGGCGG	GTGCGGCCGC	GGGCCAGTGC	GTGAGTTGTA	1561
CTGCAGGATG	GAGAGGTTGT	AACCGAGAGG	CTCG			

Figure 2 : séquence génomique de la thiorédoxine m de Chlamydomonas reinhardtii

La figure 3 montre la traduction de cette séquence en protéine. On notera que la séquence de l'intron en bleu est absente ici car l'intron est excisé lors de la transcription ADN/ ARN.

MALVARRAAVPSARSSARPAFARAAPRRSVVVRA**EAGAVN**DDTFKNVVLESSVPVLVDFWAPWCGPCRIIAPVVD
EIAGEYKDKLKCVKLNTDESPNVASEYGIRSIPTIMVFKGGKKCETIIGAVPKATIVQTVEKYLN

Figure 3 : traduction protéique de la séquence génomique
Après intégration dans le chloroplaste, la préprotéine est clivée et la protéine mature commence au niveau souligné.

[6] » How many genes are in a genome? (bionumbers.org)

Est-il nécessaire de revisiter le code génétique ?

Comme indiqué précédemment, le code génétique a résisté brillamment aux assauts du temps et de l'expérimentation ; aussi on peut se demander quel serait l'intérêt de le revisiter. Une première remarque concerne le génome nucléaire humain : s'il est vrai qu'il contient un nombre de gènes apparemment très modeste, par contre la taille globale de ce génome est énorme (3 200 000 000 de paires de bases). Si l'on compare le nombre de gènes codant pour des protéines à la capacité totale de codage théorique du génome, on aboutit à la conclusion surprenante que seul 1 à 2% du génome code pour des séquences protéiques. Un certain nombre de raisons pourraient expliquer cette conclusion. L'une d'entre elles est que certaines séquences codent pour des ARN nécessaires au fonctionnement des ribosomes, y compris les ARNt, une seconde est qu'il existe dans le génome des séquences répétées qui clairement ne peuvent pas coder pour des protéines. Une troisième est la constatation que de nombreuses séquences d'origine présumée virale ou autre appelées transposons ont largement « infecté » le génome humain. Nonobstant, la valeur de 1% paraît tout de même incroyablement faible, suggérant qu'il est peut-être possible de lire le génome de façon différente par rapport à l'immuable code génétique.

À l'appui de cette proposition, je voudrais souligner le point suivant : dans les expériences réalisées initialement par Nirenberg et Khorana et leurs collègues, les séquences d'ARN synthétique utilisées ne comportent ni le codon d'initiation ATG ni aucun des codons Stop ! Et pourtant, en utilisant ces séquences dans ces systèmes certes un peu artificiels, on arrive à démarrer de la synthèse protéique en l'absence du codon d'initiation ATG, et elle se déroule même en l'absence d'un codon Stop qui signifie normalement la fin de la synthèse. Cette observation indique donc que, dans certaines conditions expérimentales, la synthèse protéique peut démarrer à partir de n'importe quel codon et qu'un codon ATG n'est pas absolument nécessaire. Cette conclusion est tout à fait orthogonale à l'utilisation actuelle des programmes de prédiction de gènes qui se basent, rappelons-le, sur la présence de séquences lisibles (multiple de 3) entre un ATG et un Stop. On peut donc imaginer qu'il est possible de lire le génome différemment si l'on utilise des variations du code génétique.

Et en effet, quelques évidences expérimentales plus récentes que les travaux de Nirenberg et Khorana viennent appuyer cette possibilité de lectures alternatives des séquences. La première figure dans un article de Blattner *et al.* en 1997, qui, travaillant sur le génome de la bactérie modèle *Escherichia coli*, ont recensé les gènes codant pour des protéines par ailleurs déjà caractérisées dans cet organisme modèle. Ils sont arrivés à la conclusion que dans *E. coli*, le codon d'initiation de la traduction n'est pas toujours ATG. Les pourcentages d'utilisation de codon de démarrage (start) sont les suivants : ATG 83%, GTG 14%, TTG 3%. Ils ont également recensé l'utilisation de codons ATT et CTG. En d'autres termes, pour environ 1/5 des séquences, le codon d'initiation diffère du codon utilisé par les programmes de recherche de gènes. Bien que ce type de conclusion soit apparemment pour le moment plutôt restreint aux modèles bactériens, il sera extrêmement intéressant de comparer les séquences génomiques aux séquences protéiques obtenues maintenant à grande échelle depuis l'avènement de la spectrométrie de masse dans la dernière décennie.

Une deuxième série d'expériences réalisées un peu plus récemment, indiquent une autre possibilité de codage alternatif pour un 21^{e} acide aminé très rarement utilisé dans les protéines, la sélénocystéine. Comme son nom l'indique, la sélénocystéine ressemble beaucoup à la cystéine (C) avec une seule variation, l'atome de soufre présent dans la cystéine est

remplacé par un atome de sélénium beaucoup plus réactif. Cet acide aminé est retrouvé chez de rares protéines chez les bactéries, les algues unicellulaires et les animaux. Notons au passage que le sélénium est du coup essentiel dans l'alimentation animale et humaine pour cette raison. La sélénocystéine est en revanche absente chez les plantes supérieures pour des raisons que nous ne discuterons pas ici. L'une des protéines qui possède une sélénocystéine est la thiorédoxine réductase qui joue un rôle essentiel dans la synthèse des désoxyribonuléotides (les briques schématisées A C G T dans l'ADN). Il en résulte que le sélénium est essentiel pour la synthèse de l'ADN chez la plupart des organismes biologiques. Les travaux essentiels sur l'incorporation de la sélénocystéine dans les protéines ont été faits dans le laboratoire de Earl et Teresa Stadtman avec une contribution importante de Vadim Gladyshev (voir par exemple Tamura *et al*., 1995). En comparant les séquences protéiques des sélénoenzymes aux séquences d'ADN codant pour ces protéines, ces auteurs ont pu déterminer que la sélénocystéine est codée par un codon TGA (UGA dans l'ARNm) qui normalement doit être un Stop dans le code génétique (*cf* tableau 2). Un mécanisme un peu complexe résulte en l'altération de la signification de ce codon lorsqu'il est suivi d'une structure secondaire dans l'ARN appelée tige-boucle puis d'un authentique codon stop ultérieurement[7].

Le message à retenir de ces travaux est que, dans certaines conditions, un codon stop TGA ne veut pas dire stop mais sélénocystéine. Des travaux similaires ont abouti à une conclusion du même type en ce qui concerne un 22^{e} acide aminé, la pyrrolysine, dérivé comme son nom l'indique de la lysine (K). Cet acide aminé est lui-même codé par un autre codon pseudo stop TAG (*cf.* figure 2)[8]. De plus, des travaux très récents font état de l'existence de mécanismes abrégés sons le nom de YARIS et qui permettent de lire au travers des codons Stop chez la levure (Beznoskova *et al*., 2019). Dans ce cas, encore une fois, un Stop ne veut plus dire réellement Stop dans certains contextes.

Conclusion

En conclusion, tout un faisceau d'indices indiquent que le codon d'initiation ATG peut être remplacé par d'autres codons avec des fréquences non négligeables, et que, dans des conditions particulières, au moins deux des codons réputés Stop codent en réalité pour des acides aminés inhabituels. En conjonction avec les travaux pionniers de Nirenberg et Khorana qui montrent que la synthèse protéique peut en réalité démarrer sur n'importe quel codon, je pense qu'il sera du plus grand intérêt de réévaluer les capacités de codage des génomes au fur et à mesure des découvertes en biologie. Je rappelle ici que les programmes de recherche des gènes sont basés sur l'existence d'un codon d'initiation unique ATG et de codons Stop bien définis. Il est donc largement possible au vu des développements que j'ai expliqués que nous ne reconnaissions pas un certain nombre de gènes à cause de connaissances fondamentales insuffisantes. À l'appui de cette proposition, lors d'une récente conversation avec Francis Martin, un collègue de mon laboratoire du côté INRAe et spécialiste du déchiffrage des génomes, je lui ai exposé mes réflexions, et il m'a indiqué qu'en ce qui concerne les génomes auxquels il s'intéresse, 30% des séquences identifiées à partir des ARNm (messagers) n'ont pas été identifiées au niveau de l'ADN génomique. De nouveaux événements extraordinaires nous attendent-ils en recherche ? La question est ouverte, nous ne sommes peut-être pas au bout de nos surprises. Je souhaite conclure cet article en soulignant que cette présentation peut

[7] Une représentation de ce mécanisme figure à l'adresse suivante :
https://www.researchgate.net/publication/43161523_Differing_views_of_the_role_of_selenium_in_thioredoxin_reductase/figures?lo=1

[8] https://en.wikipedia.org/wiki/Pyrrolysine.

paraître quelquefois un peu trop technique pour l'académie de Stanislas, mais, suite aux épidémies de Covid que nous avons vécues, nous avons tous été confrontés à l'utilisation de la vaccination par ARNm par exemple. Il me semble que si l'on veut se faire une idée des avantages ou inconvénients de ces nouvelles technologies, il est tout à fait utile pour tout un chacun de faire un petit effort de compréhension de la biologie moléculaire, et c'est à quoi tend cet article.

Pour finir et pour aider le lecteur qui n'aurait rien compris à tout ce galimatias, la figure 4 montre la traduction en acides aminés de la séquence nucléotidique de la page 2 en hommage à notre président Denis Grandjean.

MAISDENISCESTLE PARFAITACADEMICIEN

Figure 4 : signification protéique de la séquence nucléotidique en page 2
Soit : **Mais Denis c'est le parfait académicien !**

Références

https://www.nature.com/scitable/topicpage/translation-dna-to-mrna-to-protein-393/
https://www.youtube.com/watch?v=gG7uCskUOrA
https://fr.wikipedia.org/wiki/Marshall_Warren_Nirenberg
https://fr.wikipedia.org/wiki/Har_Gobind_Khorana
Marshall W. Nirenberg - Biographical (nobelprize.org)
Marshall Warren Nirenberg - Wikipedia
La régulation de la photosynthèse grâce aux thiorédoxines [Eurêka !] – YouTube
Stein M, Jacquot JP, Jeannette E, Decottignies P, Hodges M, Lancelin JM, Mittard V, Schmitter JM, Miginiac-Maslow M. Chlamydomonas reinhardtii thioredoxins: structure of the genes coding for the chloroplastic m and cytosolic h isoforms; expression in Escherichia coli of the recombinant proteins, purification and biochemical properties. *Plant Mol Biol.* 1995, 28(3):487-503. doi: 10.1007/BF00020396.
Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y. The complete genome sequence of Escherichia coli K-12. *Science* 1997 277(5331):1453-1462. PubMed ID 9278503
Tamura T, Gladyshev V, Liu SY, Stadtman TC. The mutual sparing effects of selenium and vitamin E in animal nutrition may be further explained by the discovery that mammalian thioredoxin reductase is a selenoenzyme. *Biofactors* 1995-1996 5(2):99-102.
https://en.wikipedia.org/wiki/Pyrrolysine
Beznosková P, Pavlíková Z, Zeman J, Echeverría Aitken C, Valášek LS. Yeast applied readthrough inducing system (YARIS): an *in vivo* assay for the comprehensive study of translational readthrough *Nucleic Acids Research* 2019 47(12):6339–6350.