Préface

Le monde de l'ARN moderne

Il y a une forte probabilité pour que la programmation génétique chez l'homme et d'autres organismes complexes ait été mal comprise, en partie à cause de l'hypothèse incorrecte que la plupart des informations génétiques sont portées par les protéines. Cette hypothèse provient des études historiques en biochimie enzymatique et sur l'opéron lac chez l'E. coli, au milieu du XX^{e} siècle, en accord avec le dogme mécanistique du « flux de l'information génétique » de l'époque, et a persisté malgré un certain nombre de découvertes qui auraient dû infléchir la tendance.

La première observation paradoxale était que les gènes, dans les organismes complexes, sont des mosaïques de séquences codantes pour des protéines d'une part, et non codantes d'autre part. Ces dernières, appelées « introns », ont été immédiatement et presque universellement rejetées comme des débris évolutifs, une survivance primitive du code des protéines, en dépit du fait qu'ils sont transcrits. L'alternative intéressante, à savoir que d'autres informations significatives, non codantes, puissent être transmises par l'ARN, est tout aussi, sinon plus, plausible.

La deuxième surprise est que près de la moitié du génome humain est constituée de séquences dérivées du rétrotransposon, rejetées à nouveau comme des résidus génétiques égoïstes, mais l'alternative la plus intéressante est que ces séquences puissent se comporter comme des cassettes régulatrices mobiles.

La troisième surprise est que le nombre et le répertoire des gènes codants pour les protéines sont similaires chez les nématodes et les humains, malgré plusieurs ordres de grandeur dans la complexité de ces deux organismes. La rationalisation qui s'ensuivit fut que le potentiel explosif de la combinatoire des « facteurs de transcription » fournissait plus que suffisamment de marge de régulation pour diriger l'ontogenèse d'un ver ou d'un humain. Cette hypothèse, cependant, n'était

pas mathématiquement justifiée, ni par référence à la théorie déterministe ni par référence à la théorie mécanistique, mais a été acceptée sans critique, car elle était confortable. La découverte de petits ARN régulateurs et d'ARN interférents a été traitée comme un ajout au paradigme établi de la régulation de l'expression des protéines, d'autant plus que les miARN régulent la traduction et la stabilité de l'ARNm, plutôt que de constituer la pointe d'un iceberg reflétant le potentiel régulateur des ARN non codants.

Contrairement aux gènes codants pour les protéines, la partie non codante du génome augmente avec la complexité du développement, atteignant plus de 98 % du génome chez les humains. De plus, les études transcriptomiques à haut débit effectuées au cours de la dernière décennie ont montré que la majeure partie du génome des mammifères est transcrite de manière très régulée, produisant, en plus des petits ARN mentionnés ci-dessus, une pléthore d'ARN antisens, intergéniques et exoniques, collectivement appelés longs ARN non codants (lncARN).

Certains lncARN sont des précurseurs de petits ARN, mais la plupart d'entre eux sont fortement spécifiques d'un type cellulaire, largement exprimés dans des tissus plus restreints que les gènes codants pour des protéines, bien qu'il y ait des exceptions. Certains ont remis en question la pertinence des lncARN, car leurs séquences ne sont pas hautement conservées, par rapport à celles codant des protéines (bien qu'au moins 20 % du génome des mammifères soit conservé au niveau de la structure secondaire des ARN), et de nombreux lncARN semblent exprimés à bas niveau dans les données transcriptomiques.

L'évolution rapide des lncARN (et même des promoteurs des gènes) n'est pas surprenante, étant donné les différentes contraintes de structure/fonction des séquences régulatrices et la probabilité qu'elles soient sujettes à une sélection positive pour le rayonnement adaptatif. De plus, il est bien accepté que, étant donné le protéome de base relativement stable, la plupart des modulations adaptatives chez les animaux sont obtenues par variation de l'architecture régulatrice qui contrôle les modèles d'expression génique plutôt que les modifications des protéines elles-mêmes.

La faible expression perçue des lncARN est une conséquence du sous-échantillonnage de transcrits particuliers qui sont exprimés dans des cellules spécifiques dans des tissus complexes. L'hybridation *in situ* et le séquençage de l'ARN à haut débit ont montré que les lncARN sont rarement mais précisément exprimés dans une population cellulaire, et non une sorte de « bruit transcriptionnel ». En effet, peut-être liés à leur spécificité cellulaire, les promoteurs de lncARN sont, en moyenne, plus fortement conservés que ceux des gènes codant pour des protéines.

Bien qu'il y ait beaucoup à faire pour comprendre toutes leurs dimensions, il est clair que les ARN non codants remplissent un large éventail de fonctions dans la biologie cellulaire et développementale. Il existe de nombreux types de petits ARN, notamment les miARN auxquels on se réfère déjà, et les piARN qui semblent contrôler la mobilisation des transposons. Ceux-ci ont attiré beaucoup d'attention, mais il existe d'autres classes moins bien comprises de petits ARN qui dérivent des sites d'initiation de la transcription et des jonctions d'épissage, et qui peuvent jouer un rôle dans le positionnement des nucléosomes. De plus, tous les snoARN H/ACA (de la levure à l'homme) produisent des molécules de type miARN et tous les ARN de boîte C/D produisent des fragments de la taille d'un piARN. Les ARNt sont également clivés pour produire des fragments spécifiques qui sont exportés à partir de cellules, dont les orthologues décorent les extrémités de certains ARN viraux et humains.

Il y a des mondes de régulation à découvrir ; les liens fonctionnels et les réseaux parmi ces ARN restent à déterminer. Bien que la plupart ne le soient pas, certains lncARN sont largement exprimés, comme XIST, responsable de l'extinction de l'un des 2 chromosomes X chez les femelles. Un autre est MALAT1, l'un des ARN les plus fortement exprimés chez les vertébrés, qui est également associé à la chromatine. Sa fonction est inconnue et sa suppression ne produit que des conséquences développementales subtiles. Un autre est NEAT1, qui est exprimé et associé à d'énigmatiques organelles subnucléaires spécifiques de mammifères, appelées « paraspeckles », dans des types particuliers de cellules différenciées, et dont l'absence ne produit à nouveau que des phénotypes subtils, principalement en rapport avec la reproduction placentaire. Une interprétation que je privilégie est que ces ARN sont impliqués dans la mise en place des plateformes de l'apprentissage. Un exemple est l'ARN du cerveau BC1 fortement dérivé d'un rétrotransposon, dont la délétion ne produit aucune conséquence manifeste sur le développement, mais provoque la perte du comportement exploratoire – invisible dans la cage, mais létale dans la nature. Un autre est le lncARN Gomafu, qui décore les spliceosomes modifiés dans des neurones particuliers, et a des liens mécanistes avec la schizophrénie. D'autres sont associés à des structures doubles, inconnues, dans les noyaux des cellules de Purkinje. De nombreux lncARN semblent être impliqués dans la détermination de l'identité cellulaire. La plupart, mais pas tous, sont localisés dans le noyau et beaucoup sont associés à des complexes modificateurs de la chromatine. Cela suggère que leur fonction première est de guider la centaine d'enzymes modificatrices de l'ADN ou des histones pour marquer différemment les nucléosomes à des millions d'endroits différents dans du génome de différentes cellules, à différents moments au cours du développement. Les lncARN peuvent également servir d'échafaudages pour l'assemblage de complexes ADN-ARN-

protéine, organisant l'architecture de la chromatine. Les séquences enhancer sont transcrites dans les cellules où ils sont actifs. Ces ARN sont supposés être un sous-produit de l'activation de l'enhancer, mais ils semblent plus susceptibles d'être impliqués dans le guidage de la formation de la boucle de chromatine associée à l'action de l'enhancer. Un nombre étonnamment élevé de lncARN a été caractérisé dans le cytoplasme, avec des preuves émergentes que certains d'entre eux sont impliqués dans les processus de transduction du signal. D'autres peuvent créer des domaines subcellulaires dans le cytoplasme et dans le noyau, interagissant éventuellement avec des régions intrinsèquement désordonnées des protéines pour créer des granules d'ARN ou des régions cristallines liquides. Par exemple, un lncARN décore un mystérieux domaine dans les cellules de Purkinje.

La force des lncARN est leur capacité à traverser le monde numérique et analogique en biologie : il relie les structures tridimensionnelles (formées par des liaisons hydrogène sur les brins Watson-Crick, la face de Hoogstein et la face ribose par le 2'OH), peut interagir avec des protéines, avec des séquences d'autres ARN ou d'ADN. Les lncARN étaient probablement les molécules primordiales de la vie, qui ont transféré leurs fonctions analogues aux protéines, plus versatiles chimiquement, et leurs fonctions informationnelles à l'ADN, plus stable et facilement réplicable. Il est probable que les lncARN aient subi une renaissance, en tant qu'intermédiaire des processus épigénétiques, guidant le développement d'organismes complexes. Ces considérations impliquent une structure modulaire, hypothèse soutenue par des preuves récentes de l'épissage universel des exons des lncARN et de la localisation de structures conservées au sein des exons. Les observations selon lesquelles les exons épissés alternativement sont localisés avec des promoteurs et/ou préférentiellement localisés dans les nucléosomes suggèrent que les exons des lncARN peuvent non seulement être l'unité modulaire de structure-fonction, mais aussi de la régulation épigénétique à base d'histones. Analyser les relations structure-fonction dans les lncARN est un grand défi qui sera rendu plus facile si les séquences d'ARN régulatrices sont vraiment modulaires – auquel cas, une fois reconnues, ces modules peuvent former la base d'un nouveau Rfam, comme la base de données Pfam a été si utile dans l'identification des domaines orthologues pour les protéines.

Il y a beaucoup de mystères dans la biologie des lncARN, deux en particulier. Le premier est l'expression de 3'UTR séparés de leurs séquences codant pour des protéines normalement associées, avec des preuves génétiques que ces lncARN 3'UTR peuvent transmettre des informations en *trans*. Les 3'UTR sont bien établis pour contrôler la traduction de l'ARNm et la demi-vie grâce à des protéines agissant en *cis*. On ignore pourquoi ils auraient dû évoluer vers des fonctions en *trans*, mais le fait que les séquences 3'UTR se soient considérablement développées au cours de l'évolution des vertébrés les rend plus hautement conservés que les séquences

codantes pour les protéines associées. Le deuxième mystère est l'observation que près de 14 000 éléments ultraconservés de 100 paires de bases ou plus sont exprimés en lncARN. Ceux-ci évoluent rapidement dans l'évolution des tétrapodes, puis se figent dans les amniotes, étant presque identiques chez tous les mammifères. Ces séquences non codantes sont transcrites, mais ne peuvent être expliquées, ni en termes évolutifs ni en termes moléculaires. Certains de ces UTR ont été mutés chez la souris, mais ne montrent aucun phénotype manifeste du développement, ce qui suggère qu'ils pourraient avoir un autre rôle (très important) chez les oiseaux et les mammifères, probablement dans la parentalité et l'apprentissage.

En tout cas, à cause de ces molécules ne codant pour rien, nous réalisons que nous ne comprenons que très partiellement la biologie humaine ou son évolution. La nouvelle frontière consiste à comprendre le rôle de l'édition de l'ARN avec, au moins, ces 140 modifications différentes de l'ARN, à inventer l'« épitranscriptome ». La modification de l'adénosine en inosine se développe massivement avec la cognition, surtout chez les primates, successivement avec les trois vagues de colonisation de la lignée des primates par les éléments mobiles Alu. Les enzymes APOBEC, qui catalysent l'édition C>U, sont spécifiques aux vertébrés et se développent chez les mammifères, en particulier chez les primates où une famille (impliquée dans la régulation de la mobilisation du rétrotransposon) montre une forte sélection positive dans l'évolution humaine. La perte ou mutation d'enzymes qui catalysent les modifications de l'ARN conduit à diverses maladies, notamment la déficience intellectuelle et le cancer. Il existe également des preuves émergentes de l'héritage épigénétique transgénérationnel médié par l'ARN, il se peut donc que l'ARN ait également été coopté pour permettre la plasticité entre les générations. La présence de transcriptases inverses, en particulier dans le cerveau, implique également une interaction beaucoup plus dynamique entre l'ADN et l'ARN, à la fois en temps réel et lors de l'évolution.

En conclusion, ce qui a été rejeté comme indésirable parce qu'il n'était pas compris détient probablement la clé pour comprendre le développement humain et l'intelligence humaine. L'ARN n'est pas simplement un intermédiaire passif et éphémère entre le gène et la protéine, mais le moteur de calcul de la biologie cellulaire, du développement, du cerveau et probablement de l'évolution elle-même.

John S. MATTICK
Institut Garvan de recherche médicale
et Université de Nouvelle-Galles du Sud,
Sydney, Australie

Avant-propos

98 % du génome humain ne code pas pour des protéines et représente ce que l'on appelle communément la matière noire, ou la face cachée, du génome. Il est maintenant admis que cette face cachée est exprimée en ARN non codants et leur nombre ne fait qu'augmenter avec l'évolution de la sensibilité des technologies de séquençage. Si le nombre de gènes codants pour des protéines reste stable depuis quelques années aux alentours de 20 000 chez l'homme, le nombre de gènes non codants atteint plusieurs dizaines de milliers de représentants. Plusieurs exemples, caractérisés par différents laboratoires dans le monde, s'avèrent être fondamentaux dans la régulation de l'expression des gènes et ont des impacts majeurs sur le développement cellulaire et la progression de la tumeur cancéreuse. Une large communauté internationale de scientifiques s'est constituée et l'intérêt que suscitent ces longs ARN non codants ne fait qu'augmenter dans l'idée de comprendre les régulations épigénétiques du génome.

Dans cet ouvrage, je dresse un état des lieux non exhaustif des connaissances que nous avons aujourd'hui des longs ARN non codants. Dans le chapitre 1, je replace leurs découvertes dans l'histoire de la biologie moléculaire et je discute, dans le chapitre 2, les nomenclatures qui sont proposées aujourd'hui pour les regrouper en familles. Malgré tous les efforts qui sont faits depuis une dizaine d'années, il reste encore impossible de prédire la fonction d'un ARN non codant en ne connaissant que sa séquence primaire. Toutefois, un certain nombre d'entre eux, dans différents organismes modèles, ont été caractérisés moléculairement, biochimiquement et génétiquement et ont révélé des fonctions cruciales dans des mécanismes essentiels de régulation de l'expression des gènes, de maintenance et d'intégrité du génome. Dans le chapitre 3, je m'attache à décrire ces quelques exemples pour susciter la volonté de se confronter à leurs littératures respectives, chaque mois plus dense.

Enfin, dans les deux derniers chapitres, je décris un peu plus précisément les rôles de certains longs ARN non codants lors du développement de la cellule et de l'organisme, mais aussi dans le contrôle de la progression tumorale, ces deux aspects ayant souvent beaucoup de points communs. J'espère donner au lecteur les clés pour comprendre ces longs ARN non codants dans ces deux axes de recherche.

Le cancer est aussi l'occasion de souligner que la grande singularité de ces longs ARN non codants, vis-à-vis de chaque tissu ou type cellulaire, en fait d'excellents candidats comme biomarqueurs, pour le diagnostic ou le pronostic. Les applications sont énormes et certaines études s'attachent désormais à les exploiter comme cible thérapeutique.

Avec cet ouvrage, j'espère que chacun y trouvera sa grille de lecture, lui permettant, au besoin, d'explorer l'immense littérature que je ne cite pas intégralement, et donner l'envie aux étudiants, aux chercheurs d'apporter leur regard pour éventuellement illuminer cette face cachée des génomes dans leurs recherches présentes ou futures.

On trouvera aussi en fin d'ouvrage un glossaire et une liste d'abréviations qui rendent la lecture du livre plus accessible aux étudiants et chercheurs qui veulent entrer dans l'immense quantité de données générées par ce nouveau continent.