Ballade autour de quelques thèmes de biologie.

Bahram Houchmandzadeh

10 octobre 2005

Table des matières

1 Introduction : de quoi on parle.

Blabla d'introduction : le thème central est le traitement moléculaire d'information par le vivant.

1.1 Les échelles de la vie.

Avant de rentrer dans le vif du sujet de décrire la vie, fixons nous les idées sur les diverses échelles de taille. La notion d'échelle est un concept flou : un objet à l'échelle du mètre par exemple peut avoir, *en gros*, entre un dixième et dix mètre.

Nous les humains vivons à l'échelle du *mètre* et c'est surtout à cette échelle que nous concevons la vie. La plupart des objets que nous construisons et manipulons, les animaux que nous mangeons etc, sont en gros de cette taille.

Zoomons mille fois ($\ell \sim 10^{-3}$m) et nous arrivons à l'échelle du mm, les plus petites graduations que nous avons sur les règles. C'est l'échelle à laquelle vivent les petits insectes, les fourmis, les puces. C'est également l'échelle de la résolution de notre oeil : nous ne distinguons pas les objets plus petits que quelques dixièmes de millimètre. Enfin, c'est à cette échelle que les phénomènes capillaires sont importants.

Zoomons encore mille fois ($\ell \sim 10^{-6}$m) , et nous arrivons à l'échelle du micron (μm). C'est l'échelle des cellules : une bactérie a environ un micron de diamètre et quelques micron de long ; la taille typique des cellules de notre corps est de l'ordre de 10 à 20 microns, même si certains neurones peuvent pousser des extensions jusqu'à un mètre. C'est également la limite de résolution de nos microscopes optiques : depuis les travaux de M. Abbe, nous savons que nos microscopes ne peuvent pas distinguer des objets plus petit que $\sim \lambda/2\text{NA}$, où λ est la longueur d'onde de la lumière ($\approx 0.5\mu$m pour le vert) et NA l'ouverture numérique de l'objectif (≈ 1). Dans le meilleurs des cas, ceci met la résolution de nos microscopes optiques à $\approx 0.2\mu$m. Un virus mesure de l'ordre de 0.05μm, et c'est pourquoi leur existence n'a été découvert qu'au début du vingtième siècle

Zoomons encore mille fois, et nous arrivons à l'échelle du nanomètre ($\ell \sim 10^{-9}$m) qui est celle des molécules de la vie. Une molécule d'ADN fait environ un nanomètre de diamètre, une protéine globulaire de 5 à 10 nm, un microtubule (des tubes creuses pour maintenir la rigidité de la cellule) une vingtaine de nm de diamètres. C'est également la limite de résolution des microscopes électroniques, qui peuvent voir, dans le meilleur des cas, jusqu'à 1/10ème de nm.

Nous sommes presque arrivé au bout de notre voyage : si nous zoomons cette fois seulement dix fois, on arrive à l'échelle d'angstroem ($\ell \sim 10^{-10}$m), qui est l'échelle des atomes : l'atome d'hydrogène a une "taille" d'environ 0.5 angstroem, celui de carbone environ 1 angstroem.

Remarquons la proximité entre les échelles des bactéries et des protéines : le long du diamètre d'une bactérie, on ne pourrai disposer qu'une centaine de protéines. En fait, il ne faut

surtout pas croire que les cellules sont des sacs d'eau avec quelques molécules qui y baignent. Les cellules sont en faite très *meublées*, et on y circule avec de la peine, poussant constamment des amas de protéines pour se faire de la place. Cela ressemblerai plutôt à une jungle tropicale. Les cellules eukaryotes (on y vient tout de suite à leur définition) sont remplies de filaments qui y tissent des filets. La taille typique des mailles de ce filet est de l'ordre de 0.2 micron.

Dans le reste de ce cours, nous passerons la majorité de notre temps à l'échelle de la biologie moléculaire (nm) et cellulaire (μm).

1.2 Les habitants du monde.

Il existe deux types d'espèces vivantes : les prokaryotes et les eukaryotes.

Les prokaryotes sont les bactéries. Elles vivent sous forme unicellulaire, même si elles savent s'organiser sous formes de communauté de citoyens libres[1]. Elles ont une membrane (une peau) rigide et ne possèdent pas de noyaux. Ce sont la forme dominante et principale de la vie et si on mesure le *succès* au nombre de copies des individus, ce sont les champions hors compétition. Certaines évaluations récentes laissent penser que même en terme de biomasse, les bactéries forment l'espèce majoritaire. Le lecteur connaît sûrement le rôle des bactéries dans le recyclage des bio-matériaux. Citons d'autres exemples : Les espèces herbivores comme la vache ou le kangourou mangent de l'herbe. Mais en faite, l'herbe est surtout constituée de cellulose, indigeste pour ces animaux. Leurs intestins cependant sont tapissés de bactéries qui, *elles*, digèrent la cellulose, le brisant en sucres élémentaires ingérabledigérable par leurs hôtes. L'intestin de tous les animaux, les humains y compris bien sûr, est en faite un haut lieu de symbiose avec les bactéries.

Toutes les molécules de la vie sont formées essentiellement de trois types d'atomes : l'oxygène, le carbone et l'azote. L'oxygène et le carbone peuvent être absorbé directement de l'atmosphère par les plantes et être incorporé dans les molécules organiques (les animaux bien sûr ne sont capable de pas grand chose et se contentent d'absorber les molécules organiques des plantes). Mais l'azote est une autre histoire : bien que 80% de l'atmosphère soit formée d'azote moléculaire N_2, les plantes n'ont jamais appris à l'utiliser directement. Certaines plantes comme le pois et le haricot établissent, au niveau de leurs racines, une relation symbiotique avec des bactéries qui elles savent fixer l'azote (l'incorporer à des molécules organiques) et le donner à l'hôte qui en échange, les nourrit. La dégradation de ces plantes enrichit le sol en azote organique, utilisable par d'autres plantes et par les animaux qui les mangent (et les animaux qui mangent ces derniers et ainsi de suite). La grande révolution agricole du dix neuvième siècle était de produire de l'azote organique (nitrate, amuium,...) et d'enrichir directement les sols en court-circuitant les bactéries.

Ces quelques exemples pour dire qu'une expédition de martiens voulant étudier la vie sur terre emporterait plutôt des bactéries que la soeur de l'agent Mulder.

Les eukaryotes sont des cellules à enveloppe extérieure souple (pour ne pas dire molle) et possèdent un noyau où l'information génétique est entreposée. Pour maintenir leur intégrité physique et se donner un peu de rigidité, elles utilisent un ensemble de tubes et de filaments

[1] On estime actuellement que 95% des bactéries vivent en biofilm.

qu'on appelle cytosquelette. Elles peuvent vivre sous forme uni-cellulaire comme la levure, le *Dictyostelium*, la *Paramécie*, le *pediastrum*, se nourrissant de bactéries ou de lumière. Elles peuvent s'organiser en ensemble de cellules peu spécialisées comme l'hydre (possédant essentiellement deux types de cellules). Elles peuvent être également sous forme de mégalopole extrêmement hierarchiséhiérarchisée possédant des centaines de types de cellules, comme le tabac, la drosophile, la grenouille, le séquoia, l'humain ou le salamandre.

Toutes les espèces vivantes utilisent les même mécanismes fondamentaux pour survivre et se dupliquer en consommant de l'énergie. En général, ces mécanismes sont plus complexes chez les eukaryotes que chez les prokaryotes mais pas fondamentalement différents. Toutes utilisent l'ATP comme pétrole et l'ADN comme banque d'information. Les machines de transcription et de translation sont extrêmement similaire. La régulation de l'expression génétique suit des principes proches dans les deux cas. Ce sont toutes ces phénomènes que l'on va visiter lors de notre tournée.

Comme le lecteur l'aura remarqué, nous avons exclu les virus de notre définition du vivant. Ce sont eux aussi des organismes qui se reproduisent mais ils ne possèdent pas la machinerie nécessaire pour cela. En faite, les virus sont des toute petites particules (de l'ordre de 10-50 nm) formées d'assemblages de protéines qui englobent des morceaux d'ADN ou d'ARN en leur sein. Pour se reproduire, ils doivent détourner la machinerie des vrais cellules (eukaryotes ou prokaryotes).

Finalement, depuis les années 80, nous connaissons un autre agent capable de se reproduire, très différent des précédents, qu'on appelle *prion*s. Ce sont des protéines qui ont été mal "repliées" et qui sont capable de rendre d'autres protéines saines à leur image. L'agent de la maladie de la vache folle est un prion.

1.3 Les échelles d'énergie.

Dans le monde à notre échelle, une carafe posée sur une table ne bouge pas. Dans le monde à l'échelle du micron et en dessous, la carafe microscopique effectue au contraire des mouvements aléatoires non stop. Le monde microscopique est étroitement lié à ces fluctuations aléatoires : rien **(** rien rien) n'est statique, tout fluctue. L'échelle d'énergie associée à ces mouvements est donné par la température T et vaut kT où k est la constante de Boltzman. Prenons par exemple une particule microscopique dans un puits de potentiel de la forme $E = (1/2)\alpha x^2$(Fig.1.1). Au lieu de rester au fond du puits à $x = 0$, la particule va constamment bouger de façon erratique, et si on pouvait mesurer son énergie potentielle moyenne (dans le temps), on trouverait qu'elle vaut $(1/2)kT$. Si on mesurait la position de la particule au cours du temps et qu'on construisait un histogramme des positions, on trouverait que ce dernier ressemble à une courbe en cloche : les positions de haute énergie (loin du centre) sont moins probables que celles de basse énergie. Plus exactement, la probabilité de se trouver à la position x est proportionelle à $\exp(-E(x)/kT)$. On prétend que cette formule figure sur la tombe de monsieur Boltzman.

Cela nous ramène directement à la stabilité des liaisons chimiques : si l'énergie libre d'une liaison chimique est de l'ordre de kT, les fluctuations thermiques vont forcément la rompre. Pourtant quelques kT suffisent pour former une liaison stable : Supposons qu'une molécule

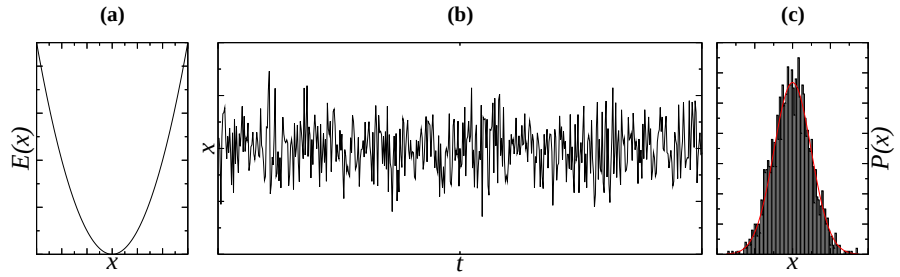

FIG. 1.1: Exemple de fluctuations thermiques. (a) l'énergie potentielle E de la particule en fonction de sa position x ; (b) à cause des fluctuations thermiques, la particule ne reste pas au fond du puits à $x = 0$, mais effectue des mouvements aléatoires ; (c) l'histogramme des positions que la particule occupe au cours du temps, ici calculé sur 2000 points. La courbe en rouge est la fonction $P(x) = \exp[-E(x)/kT]$.

TAB. 1.1: Probabilité P d'observer un état non-liée en fonction de l'énergie libre d'association $-nkT$

n	1	2	4	5	10	20
P	0.27	0.12	$1.7\,10^{-2}$	$7\,10^{-3}$	$5\,10^{-5}$	$2\,10^{-9}$

puisse exister sous forme liée avec une énergie libre de $-nkT$ et sous forme non-liée avec une énergie libre 0. Le tableau1.1 donne la probabilité de l'observer sous forme non-liée en fonction de n . Nous voyons qu'une vingtaine de kT rende une liaison presqu' eternelle !

Parlons un peu chiffre. A $T = 300$K (la température ambiante standard), $1kT = 1.38\,10^{-23} \times 300 = 4.14\,10^{-21}$ J $= 2.6\,10^{-2}$ev. 1 électron volt, qui est l'ordre de grandeur de l'énergie libre des liaisons chimiques covalentes vaut donc une quarantaine de kT (quand $T = 300$K), ce ne sont donc pas les fluctuations thermiques à température ambiante qui peuvent rompre une liaison covalente. Les liaisons faibles qui sont les interactions hydrogènes et hydrophobes ont au contraire une énergie de liaison de l'ordre de 1 à 4 kT : toutes seules elles ne sont pas suffisante, il en faut plusieurs si c'est la stabilité qui est recherchée.

Nous allons voir par la suite que beaucoup de liaisons en biologie ont des énergies libres autour d'une dizaine de kT : suffisamment stables vis à vis des fluctuations thermiques, mais facilement réarrangeables quand le besoin se fait sentir. La consommation d'une molécule d'ATP procure une vingtaine de kT et elle est largement utilisée dans ce but. Nous rencontrerons constamment des exemples par la suite.

Les biochimistes et les biologistes ont plutôt l'habitude de parler en Kilo Calorie par mol : 1Kcal/mol=1.7kT.

2 Les molécules de la vie.

La vie est essentiellement l'histoire de deux familles de molécules, toutes deux des polymères : les protéines et l'ADN. Il existe bien sûr beaucoup d'autres familles de molécules biologiquement importantes : les membranes de cellules sont formées de lipides, le monde végétal n'est rien sans la cellulose (polymère de sucre), ... Mais ce sont quand même des seconds rôles. Les molécules "stars" sont celles que nous venons de citer au début et la vie est l'histoire de transfert d'informations entre elles. Nous ne parlerons donc que de ces deux-là.

Petite précision : un polymère est une longue chaîne linéaire formée d'éléments simples appelés monomères. Le produit d'emballage classique, le polyéthylène, est, comme son nom l'indique, une chaîne formée de l'enchaînement de molécule d'éthylène : $(\text{-}CH_2 - CH_2-)_n$. La chimie humaine ne sait pas synthétiser des polymères de longueur ou d'arrangement bien contrôlé. La chimie de la vie au contraire est extrêmement précise. Le génome de bactériophage λpar exemple est un polymère d'acide nucléique formée de 48500 monomères.

2.1 les protéines

2.1.1 Que font elles ?

Toutes les fonctions essentielles de la vie sont effectuées par des protéines. Ce sont de magnifiques machines d'environ 5 à 10 nm qui chacune exécute une tâche particulière avec une précision extraordinaire. En voici une liste très non-exhaustive de leurs occupations.

- Ce sont des **catalyseurs**, des chimistes, des synthétiseurs de molécules. La chimie pratiquée par les humains est brutale, chauffant à haute température, passant les produits sur des colonnes de refroidissement et de filtration, générant une multitude de produits dérivés non souhaitables. La chimie des protéines, comparée à cela, paraît de la science fiction : une réaction se déroule à température ambiante, sans déchets (tout se recycle), rapide et extrêmement spécifique. C'est ainsi par exemple que l'énergie des photons de lumière est transformée en énergie chimique (photosynthèse) et que la terre est couverte de cellules photo-voltaïque (on les appelle *feuilles vertes*), que les molécules de glucose (un sucre élémentaire) sont cassées pour produire de l'ATP, que des antibiotiques et des hormones sont produit... Dans le paysage énergétique des réactions chimiques, les protéines connaissent les raccourcis tandis que les humains font gravir à leurs molécules monts et vallées.
- Ce sont des **capteurs**. Si nous détectons des odeurs, c'est parce que des molécules aromatiques s'accrochent à des protéines spécifiques à la surface des cellules du nez. La signalisation entre les cellules par les hormones suit le même chemin.
- Ce sont des **canaux** et des **portes**. Une cellule ne peut pas vivre isolée de l'extérieur.

Ses portes vers "dehors" sont des protéines. Ce sont par exemple des canaux ioniques qui pompent les ions dans les deux sens pour maintenir l'équilibre salin de la cellule. Dans les cellules eukaryotes, le noyau est séparé du cytoplasme par une enveloppe, et ce sont des protéines, appelées pores nucléaires qui régulent le trafic de et vers le noyau.

- Ce sont des **moteurs moléculaires**, transportant des cargaisons d'un bout de la cellule à l'autre, provoquant le battement des cils et des flagelles, la contraction des muscles.
- Ce sont des **éléments structurants** sous forme de filaments et de tubes qui procurent la rigidité à la cellule.

Cette liste peut continuer longtemps. Dans un film appelé "mariage à la grec", un immigrant installé aux Etats-Unis depuis des années aborde toujours les natifs par "donnez moi un mot, n'importe quel mot, et je vous démontre qu'il est d'origine grec". Eh bien, citez moi une fonction, n'importe qu'elle fonction, et je vous démontre qu'elle est effectuée par une protéine[1].

2.1.2 Qui sont elles ?

Les protéines sont des polymères, c'est à dire formées à partir de bloques de bases assemblées bout à bout (Fig.2.1). Les blocs de base sont des acides aminés. Bien que le nombre possible d'acides aminés soit illimité, la nature en utilise seulement une vingtaine aux noms aussi poétiques que prolyne(P), histidine(H), leucyne(L),... Par un heureux hasard, le nombre d'acides d'aminés est le même que les lettres de l'alphabet latin.

La séquence des acides aminés mis bout à bout détermine l'identité de la protéine : c'est comme la séquence des lettres qui détermine le contenu d'une page. Par exemple, Le début de la séquence de l'hémoglobine, la protéine responsable du transport d'oxygène, est `EKSAVTALWGKVNVDEVGGEALGR...` La plupart des protéines dans la nature ont une centaine à quelques milliers d'acides aminés. Une fois la chaîne protéique formée, elle se replie pour former une structure tridimensionnelle. La forme de la structure est fonction de la séquence d'acides aminés de la protéine.

Il existe une relation très étroite entre la forme tridimensionnelle d'une protéine est la fonction qu'elle exerce. En caricaturant un peu, on peut dire que la chimie des protéines peut être ramenée à la géométrie et à la science des surfaces qui s'emboîtent. Les antibiotiques par exemple sont des agents de guerre chimique entre les bactéries. Ce sont de petites molécules qui peuvent s'emboîter parfaitement dans le site actif d'une autre protéine (d'une bactérie concurrente) et inhiber la fonction de cette dernière, provoquant la mort de l'organisme. Un autre exemple est le virus d' HIV qui s'ancre sur une protéine particulière de la surface des cellules pour pouvoir ensuite pénétrer à l'intérieur. Une des molécules de la tri-thérapie épouse la forme du site actif d'ancrage, empêchant le virus d'occuper la place.

Les maladies à base de prions sont enfin un contre exemple intéressant : parfois une protéine peut mal se replier est devenir non fonctionnelle. En général, ceci ne porte pas à conséquence et la protéine inutile est dégradée au bout d'un certain temps (par d'autres protéines, bien sûr). Dans de très rare cas, une protéine mal repliée peut provoquer le mal repliement d'autres protéines de son espèce : c'est alors une sorte de réaction en chaîne et l'agent de la maladie, qui n'est ni une bactérie ni un virus mais simplement une protéine, se duplique à grande vitesse

[1] Le personnage du film a réussi a trouvé l'origine grec du mot *kimono*.

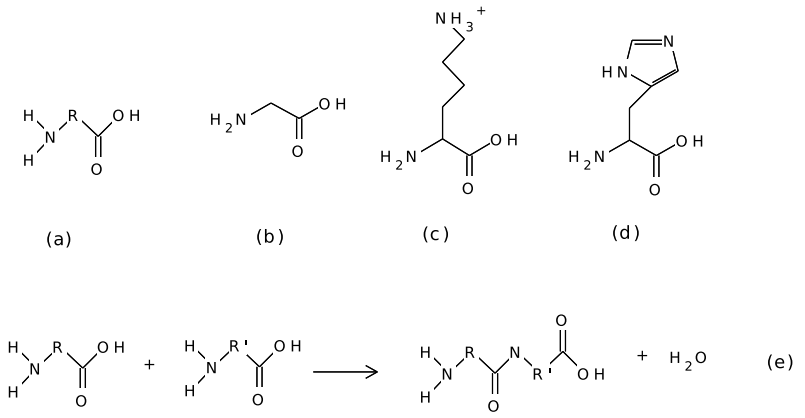

FIG. 2.1: La structure d'un acide aminé est de la forme montrée en (a) : un groupement chimique associé d'un coté à un groupe amine - qu'on appel le coté N terminal - et de l'autre coté à un groupement carboxyl -le coté C terminal. (b-d) exemples d'acides aminés : (b) glycine ; (c) lysine ; (d) histidine. Les aminoacides peuvent former des polymères (e) : un hydrogène du coté amine d'un acide aminé et l'hydroxyl du coté C terminal de l'autre acide aminé forment une molécule d'eau ; la liaison formée entre les deux acides aminés est appelée une liaison peptidique. D'autres acides aminés peuvent continuer à s'ajouter à cette chaîne par la même réaction pour former une protéine.

et provoque des dégâts dans l'ensemble de l'organisme hôte.

La connaissance de la forme des protéines est donc importante et constitue une discipline de la recherche. La méthode classique est de cristalliser la protéine et de passer ensuite le cristal dans un faisceau de rayon X pour obtenir des tâches de diffractions. La disposition de ces tâches est une signature de la structure de la protéines, très difficile à décrypter. Les premiers succès dans ce domaine ont été obtenu par Linus Pauling en 1948 pour de toutes petites protéines (quand elles comportent très peu d'acide aminé, on les appellent plutôt des polypeptides). Aujourd'hui, des instituts entiers sont consacrés à cette discipline et on construit des synchrotrons seulement dans ce but. Les structures de quelques milliers de protéines sont connues de nos jours, et elles sont entreposées dans des banques de données publiques.

Dans la vie sur terre, ne sont utilisé que la forme L des aminos acides. à compléter.

peut on déterminer la structure des protéines à partir de leur séquence ? La réponse est pour l'instant non. A compléter.

Enfin, on peut se demander pourquoi si les protéines sont des objets si magnifiques, les humains ne les synthétisent pas ? Nous avons déjà donnés la réponse : la chimie humaine est à l'âge de pierre. Quand nous avons besoin de protéines, nous les extrayons d'organismes vivants (on appelle parfois cela "manger"). Ce sont d'autres machines moléculaires[2] qui fa-

[2]Des complexes à base d'ARN et de protéine qu'on appelle ribosomes, mais on verra tout cela plus loin.

briquent des protéines. On peut voir ces dernières comme des ateliers généralistes qui construisent des protéines dont on leur a fourni les plans. Et les plans se trouvent sur une autre molécule appelé ADN. Nous verrons cela au chapitre prochain.

2.1.3 Développement : quelques exemples de protéines.

La synthèse d'antibiotique. Les antibiotiques sont des agents de guerres chimiques entre les bactéries. Une fois qu'une bactérie a colonisé une niche particulière, elle ne veut surtout pas qu'une autre espèce viennent lui prendre sa place. Pour cela, elle secrète un agent chimique dans le milieu qui constitue un poison pour une espèce de bactérie concurrente, mais pas pour elle même. Depuis les années 20, les humains ont mis à leur propre profit cette guerre. La façon classique de trouver un antibiotique est d'aller de par le monde et de trouver des nouvelles espèces de bactérie, de les cultiver et appliquer à d'autres bactéries le milieu de culture des premiers. Si ces derniers meurent, c'est que les premiers produisent des antibiotiques intéressant. On essaie alors de purifier le milieu et de trouver *la molécule* qui a eu de l'effet. Les antibiotiques sont en général des petites molécules organiques et une fois trouvé, ne sont pas très dur à synthétiser par la chimie humaine. La production de ces petites molécules est bien sûr, dans la cellules, catalysée par des protéines spécifiques. Une fois larguée dans la nature, elles interfèrent avec les fonctions de certain des espèces de protéines propre à un concurrent. Lors de la division d'*E.Coli* par exemple, une protéine spécifique doit couper des liaisons chimiques pour que les deux cellules soeurs puissent se séparer. La pénicilline inhibe la fonction de cette protéine, et les bactéries qui ne peuvent plus se séparer cessent bientôt de se diviser.

Les microtubules. Nous avons mentionnés plus haut que les cellules eukaryotes n'ont pas d'enveloppe rigide. Pour maintenir leur forme, elles ont développé un squelette à partir d'une protéine appelée tubuline. La tubuline peut être vu comme un cylindre de 4 nm de hauteur. Les tubulines s'assemblent afin de former un tube creux d'une vingtaine de nm de diamètre et dont la longueur peut atteindre plusieurs dizaines de microns. Ces tubes sont extrêmement rigide. Notons que la structure en tube creux n'est pas un hasard : à masse égale, c'est la forme qui procure la plus grande rigidité. Les cellules eukaryotes sont parcourues de part en part par les MT qui souvent, ont un de leur extrémité proche du noyau. Les MT servent également de rail aux moteurs moléculaires qui transporte toutes sorte de chose d'un bout de la cellule à l'autre. Mentionnons en passant qu'ils existe d'autres protéines qui s'assemblent sous forme de fibre et qui font partie du cytosquelette : ce sont l'actine et les filaments intermédiaires.

La Topoisomérase. L'ADN, comme on le verra, est un long filament souple et très long. Comme pour les cordes, il peut être victimes de noeud. Il existe une protéine appelé TopoII qui résout ces noeuds de la façon la plus élégante qui soit : elle coupe un brin, le fait passer de l'autre côté et le ressoude. Ce qui paraissait très étrange au début c'est que cette molécule ne peut bien sûr agir et tâter le noeud que localement, sur quelques nm. Or, pour savoir qu'un noeud est vraiment un noeud, il faut avoir une information globale sur toute la courbe.

Les anticorps. Les mammifère ont développé un système de défense immunitaire unique : ils possèdent un ensemble de protéines - appelés anticorps - capable de reconnaître *n'importe* quel forme étrangère. Par reconnaître nous voulons simplement dire que la forme de l'anticorps est *complémentaire* à la forme de l'agent étranger ; en langage de physicien, nous dirons qu'il existe une interaction attractive entre l'agent étranger et un anticorps spécifique, et ils adhèrent l'un à l'autre lors de leur rencontre. Le virus sont des morceaux d'ADN entouré d'un manteau protéique, et ce sont les protéines de ce manteau - qui ne sont pas produit par notre organisme - qui sont reconnu par les anticorps. De même, les membranes des bactéries sont recouvertes de protéines spécifiques de la bactérie en question, reconnaissable par les anticorps.

Il doit exister 10^{beaucoup}de formes étrangère. Comment on peut les reconnaître toutes est l'objet d'un paragraphe quand nous étudierons beaucoup plus en détail l'ADN.

La Clatherine. La cellule eukaryote peut absorber des molécules par endocytose : les molécules intéressantes pour la cellule sont piégées par des protéines récéptrices à la membrane de la cellules. Quand suffisement de molécules ont été capté, une invagination se produit dans la membrane, le pli se referme et une bulle contenant ces molécules rentre ainsi dans la cellules. L'invagination ne peut bien sûr se former d'elle même. Des protéines, appelées Clathrine s'ammassent, du côté du cytoplasme, sous la membrane et s'emboîtent les uns dans les autres pour former une sorte de ballon de football. La formation de cette structure "tire" la membrane vers l'intérieure et produit l'invagination.

La bacteriorhodopsine. et la détection de lumière.

2.2 Les acides nucléiques.

2.2.1 Que font ils ?

Les acides nucléiques sont les dépositaires de l'information génétique : ils contiennent le *manuel* pour construire les protéines. Le plus célèbre est l'ADN, l'acide désoxyribo-nucléique. C'est à travers cette molécule que l'information génétique est transférée entre les générations. Comme il est trop précieux pour être consulté sans cesse, des copies sont constamment effectuées sous forme d'ARN, acide ribonucléique, et c'est cette dernière qui est envoyé comme instruction à l'atelier.

En dehors de son rôle de contenairecontenair d'information génétique, l'ARN peut jouer d'autres rôle. Nous verrons plus tard que les ateliers de construction, l'endroit ou l'on fabrique les protéines, sont des énormes complexes moléculaire appelé des ribosomes. Ces machines sont en majorité fabriquée d'ARN et les protéines y sont minoritaire. L'ARN peut également jouer un rôle d'enzyme, au même titre que les protéines. Nous consacrons un chapitre plus bas aux divers aspects de l'ARN.

Comparé aux protéines, l'ADN paraît une molécule bien simple, sans structure tridimensionnelle importante. Jusqu'au milieu du vingtième siècle, on ne pensait pas qu'il puisse avoir

un rôle quelconque dans l'hérédité (pardon, le transfert de l'information génétique entre génération). C'est Avery, de l'université Rockefeller à New York qui a démontré cela en 1944[3] en injectant de l'ADN purifié d'une souche bactérienne à une autre souche, et en remarquant que cette dernière acquérait alors les propriétés de la première. La course à la compréhension de la structure de l'ADN a alors été lancé[4] et cela a donné naissance à la biologie moléculaire.

2.2.2 Qui sont ils ?

Les molécules d'ADN et d'ARN sont, comme les protéines, des polymères. Leurs monomères sont des nucléotides et sont au nombre de 4, appelés A,T,C,G dans le cas d'ADN et A,U,C,G dans le cas d'ARN. Le coeur d'un nucléotide est formé d'un sucre (un ribose) et d'un groupe phosphate, et c'est ce binôme qui est capable de polymériser. Un acide nucléique est attaché sur un carbone latérale du ribose et lui donne sont identité (figure 2.2a). La différence entre l'acide Désoxyribo-Nucléique (ADN) et l'acide Ribo-Nucléique (ARN) est, comme leur nom l'indique, dans l'oxydation de leur sucre : un groupe OH dans le ribose laisse sa place à un H dans le désoxyribose. Une coquetterie de la nature en plus utilise l'uracyle dans l'ARN au lieu du thymine dans l'ADN.

Les nucléotides sont des molécules orientées, et on appelle leur deux cotés 5' et 3', du nom de la position des atomes de carbones du sucre. La polymérisation, par formation des liaisons covalentes, se fait *toujours* par l'ajout des nucléotides au coté 3' de la chaîne (pour plus de détail, voir la figure 5.4). La lecture de l'ADN pour fabriquer des protéines se fait également dans ce sens. Les séquences sont toujours écrites donc dans ce sens : ACCGTGTT veut toujours dire 5'-ACCGTGTT-3'.

Les acides nucléiques d'un nucléotide sont capable de former des liaisons hydrogènes de façon spécifique : le A avec le T et le C avec le G. Le couple AT forme 2 liaisons hydrogènes, et le couple GC 3. Une molécule d'ADN est donc capable de former beaucoup de liaisons hydrogènes avec une autre si les bases qui se font face sont complémentaires *et* (très important ce "et") les deux brins sont *antiparallèl*es (Fig.2.2b). L'antiparallèlisme est obligatoire si l'on veut avoir formation correcte des liaisons hydrogènes. De plus, les deux brins complémentaires s'enroulent autour de l'autre pour former un double hélice dans l'espace[5]. La séquence complémentaire de 5'-ACCGTGTT-3' et donc 5'-AACACGGT-3'.

C'est très souvent sous la forme de double brins (ADNdb, ou dsDNA en anglais) que l'on trouve l'ADN dans la nature. L'ARN par contre est majoritairement trouvé sous forme de simple brin[6]. On voit tout de suite l'avantage de l'appariement, et cela n'a pas échappé aux

[3] En réalité, l'article d'Avery est resté dans l'oubli, et des chercheurs ont redécouvert l'ADN au début des années 50.

[4] Le lecteur sait probablement que la course a été gagné par Crick et Watson. L'histoire est très connu et ce dernier a même écrit un livre (la double hélice) relatant les détails de la compétition. L'histoire plaît beaucoup, probablement parce que deux amateurs, presque sans effort, ont réussi a gagné la course contre des grands comme le chimiste Pauling. Un exemple extrêmement rare dans le monde de la science.

[5] C'est quand même plus joli d'avoir une double hélice qu'une bête bande plate. Cependant, cela n'est pas sans poser quelques problèmes de topologie lors de la duplication d'ADN que la nature a résolu en élaborant des enzymes spécifiques au nom évocateurs de "topoisomerase", "gyrase", "helicase",...

[6] quelques virus transportent leur matériels génétiques via de l'ADN simple brin, et d'autres sous formes d'ARN double brins. Les prokaryotes et eukaryotes utilisent toujours l'ADN double brins comme support génétique.

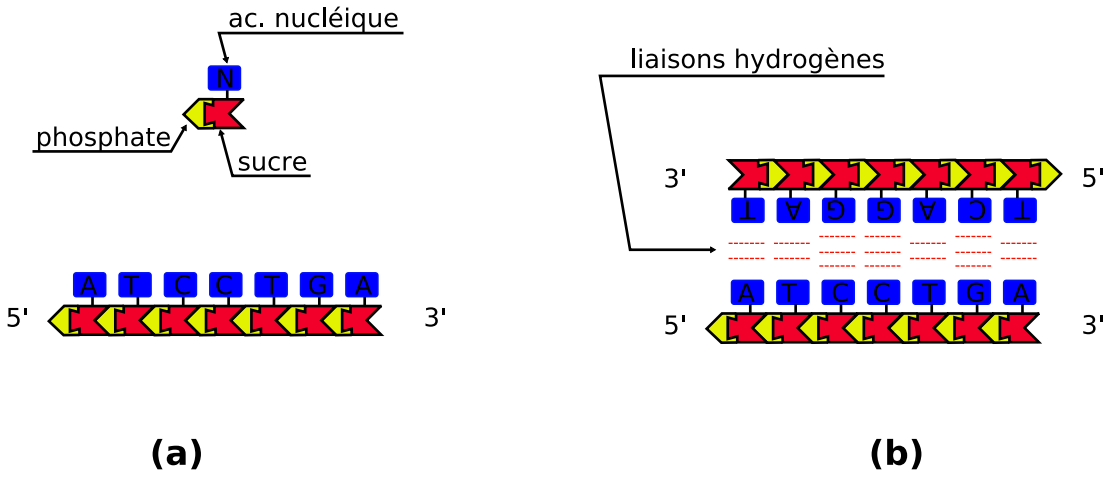

FIG. 2.2: La structure d'une molécule d'ADN. (a) le monomère est formé d'un sucre (ribose), d'un groupe phosphate et d'un acide nucléique. Le binôme sucre-phosphate est capable de polymériser (en formant des liaisons covalentes). Le monomère est *orienté*, et possède un coté 5' et un coté 3' (du nom de la position des carbones sur le sucre). Lors de la polymérisation, les nouveaux monomères sont ajoutés toujours au coté 3'. C'est toujours dans le sens $5' \rightarrow 3'$qu'on lit la séquence de l'ADN ; ici par exemple, on doit lire ATCCTGA. (b) Une molécule d'ADN peut former des liaisons hydrogènes avec une autre si les deux brins ont des séquences *complémentaires* : A en face de T et C en face de G. L'appariement est *antiparallèle*. La séquence du brin complémentaire est ici TCAGGAT. C'est souvent sous cette forme de *double brins* que l'on trouve l'ADN dans la nature.

découvreurs de la structure d'ADN : pour produire fidèlement une nouvelle molécule d'ADN double brin (qui porte l'information génétique) à partir d'une ancienne, il suffit de séparer les deux brins de l'ancien, et utiliser chaque brin comme modèle pour fabriquer un brin complémentaire. On appelle ce procédé de duplication *semiconservative*, puisque la moitié d'une nouvelle molécule d'ADN vient d'une ancienne. C'est exactement comme cela que la cellule procède, et nous verrons le détail plus loin. C'est pour cette raison que l'ADN se trouve toujours sous forme db.

Comme nous l'avons dit, les monomères d'un brin d'ADN sont reliés par des liaison covalentes, extrêmement solide donc. Les liaisons entre les deux brins complémentaires ne sont par contre que des liaisons hydrogènes, pas très solide. En augmentant la température vers 90°C[7], les deux brins se séparent facilement. On appelle cela la *fusion* de l'ADN et c'est un très joli problème de physique statistique. En abaissant la température, les deux brins peuvent s'apparier à nouveau, et on appelle cela *l'hybridation*. Ce coté "post-it" d'ADN a des avantages pour les biotechnologies, et nous en verrons des exemples plus loin.

[7] La température de fusion dépend de la séquence, et du rapport CG/AT. Le mot fusion est utilisé de façon abusif, puisque la séparation des deux brins n'est pas une transition de phase de premier ordre, contrairement à la fusion de la glace.

3 Transfert d'information entre protéines et acides nucléiques.

3.1 Schéma général du dogme central de la biologie.

Le Dogme Central de la Biologie est le suivant : ADN → ARN → proteine. Le terme Dogme a été forgé par Francis Crick au début des années 1960 parce qu' apparemment il trouvait le terme assez "chic" sans en connaître les autres connotations. Nous avons beaucoup insisté sur le fait que l'ADN contient les informations pour construire les protéines, et c'est ce que veut dire ce dogme. Voyons cela de plus près.

Comme nous l'avons dit, l'ADN est un long polymère qui contient le manuel pour construire les protéines. On peut le voir comme le *disque dur* de la cellule : sur une partie est stocké l'information pour fabriquer la protéine "clathrine", une partie *code* pour la protéine "tubuline" et ainsi de suite. De même, le disque dur d'un ordinateur contient sur une partie le programme "Microsoft Office", sur une partie le logiciel "Kaza", et sur une autre "Mortal Combat 3"[1].

L'ADN contient l'information pour construire *toutes* les protéines mais la cellule n'a pas besoin de les fabriquer toutes à un instant donné. De plus, l'ADN est très précieux et doit être transmis aux descendants sans dommage. Il doit donc être manipulé le moins possible[2]. Quand la cellule a besoin d'une protéine en particulier, elle recopie en ARN le fragment d'ADN qui code pour cette protéine. Cet ARN (qu'on appelle messager) est ensuite envoyé vers les ribosomes qui le "lisent" et construisent la protéine adéquate en suivant les instructions contenues dedans. Chaque molécule d'ARN est utilisée de nombreuses fois par les ribosomes et est ensuite dégradée.

Cela ressemble beaucoup au fonctionnement des ordinateurs : quand nous voulons exécuter un programme, le système d'exploitation copie d'abord en mémoire vive la partie du disque dur qui le contient. La copie en mémoire vive est ensuite envoyée vers le processeur pour être exécutée. Quand le programme n'est plus nécessaire, sa copie en mémoire vive est effacée.

Les deux sections suivantes sont dédiées à ces processus. Nous étudierons d'abord la lecture d'ARN par les ribosomes (traduction) et ensuite la copie d'ADN en ARN (transcription).

[1] Le disque dur de l'auteur ne contient aucun de ces logiciels.

[2] Une façon d'expliquer la biologie comme le ferait un ingénieur lamarckiste : un effet *à posteriori* justifierai le chemin que l'évolution a pris pour résoudre un problème. L'auteur de ce manuscrit se surprend lui même à donner parfois dans ce travers, comme les lignes ci-dessus le montrent.

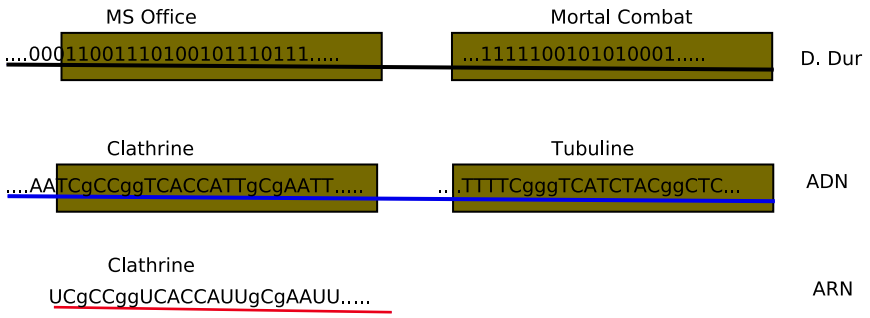

FIG. 3.1: Analogie entre disque dur et ADN : les deux contiennent des informations, le premier sous forme de bits pour des programmes, le deuxième sous forme de base pour des protéines. Quand la cellule a besoin de produire la protéine clathrine, la partie de l'ADN qui code pour cette protéine et d'abord copié sous forme d'ARN.

3.2 Développement : ARN → Protéines.

La machine moléculaire qui lit les ARN et construit des protéines est *le ribosome*. C'est un énorme complexe moléculaire, construit de plusieurs protéines *et* de l'ARN (oui oui ...)

L'ARN est une molécule très polyvalente : nous avons vu que c'est une copie temporaire d'un fragment d'ADN, on l'appelle alors ARNm (messager). Il rentre également dans l'architecture des ribosomes et on l'appelle alors ARNr. En réalité, dans le fonctionnement des ribosomes, les ARNr (et non les protéines) jouent le rôle principal.

Le fonctionnement des ribosomes est le suivant (Fig. 3.2) : un ARNm rentre dans la tête de lecture du ribosome par son côté 5'. Le ribosome lit les *trois* premières séquences, consulte un dictionnaire, trouve la signification en terme d'acide aminé de ce triplet, "pêche" l'acide aminé en question dans la solution, et le place en tête de sa ligne d'assemblage. La tête de lecture avance alors la "bande" d'ARN de trois séquences et lit le triplet suivant. En fonction de ce triplet, il choisit un autre acide aminé et le place derrière le premier sur la ligne en synthétisant la liaison peptidique entre les deux, et ainsi de suite : l'ARN est lu triplet par triplet et les acides aminés correspondants sont assemblés au fur et à mesure. Certain des triplets n'ont pas de correspondent en acide aminé, mais veulent dire STOP. Si le ribosome rencontre un tel triplet, il relâche alors la protéine dans la solution.

Le dictionnaire de correspondance entre les triplets d'acide ribo-nucléique et les acides aminés est montré dans la figure 3.3. Comme $4^3 = 64$ et que la nature n'utilise en gros que 21 acides aminés, il y a pas mal de redondance dans ce dictionnaire. En particulier, la troisième base paraît moins informative que les deux premières. Ce dictionnaire est presqu'universel, utilisé de la bactérie *E.Coli* à l'humain[3].

Le dictionnaire est en réalité contenu dans des ARN de transfert (ARNt). Voilà la troisième (et non la dernière) utilisation de l'ARN que nous rencontrons. Nous avons déjà mentionné que l'ARN est souvent rencontré sous forme de simple brin. Ceci ne veut absolument pas dire

[3]Quelques rare espèces comme la paramécie (eukaryote unicellulaire) et quelques virus utilisent une version légèrement altérée.

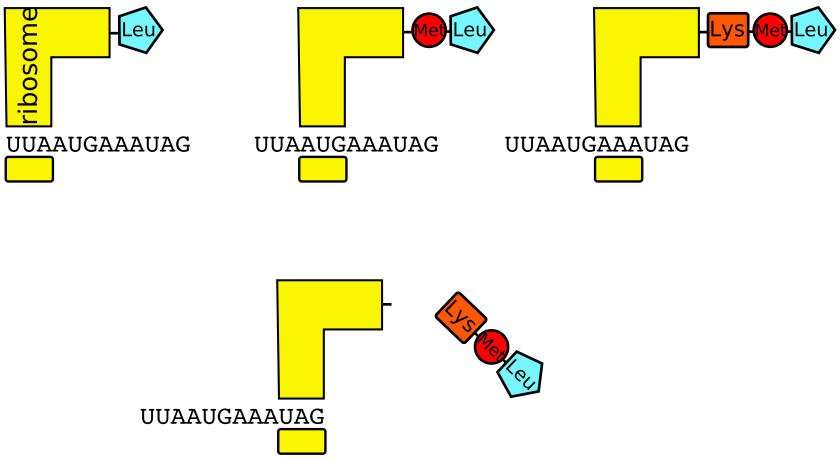

FIG. 3.2: De gauche à droite et de haut en bas : Le fonctionnement d'un ribosome. L'ARNm est lu par groupe de trois bases ; l'acide aminé correspondant à ce triplet est trouvé dans le dictionnaire du code génétique (voir fig. 3.3) et ajouté à la chaîne en cours d'assemblage ; quand la séquence STOP (ici UAG) est rencontré, la protéine est relâchée.

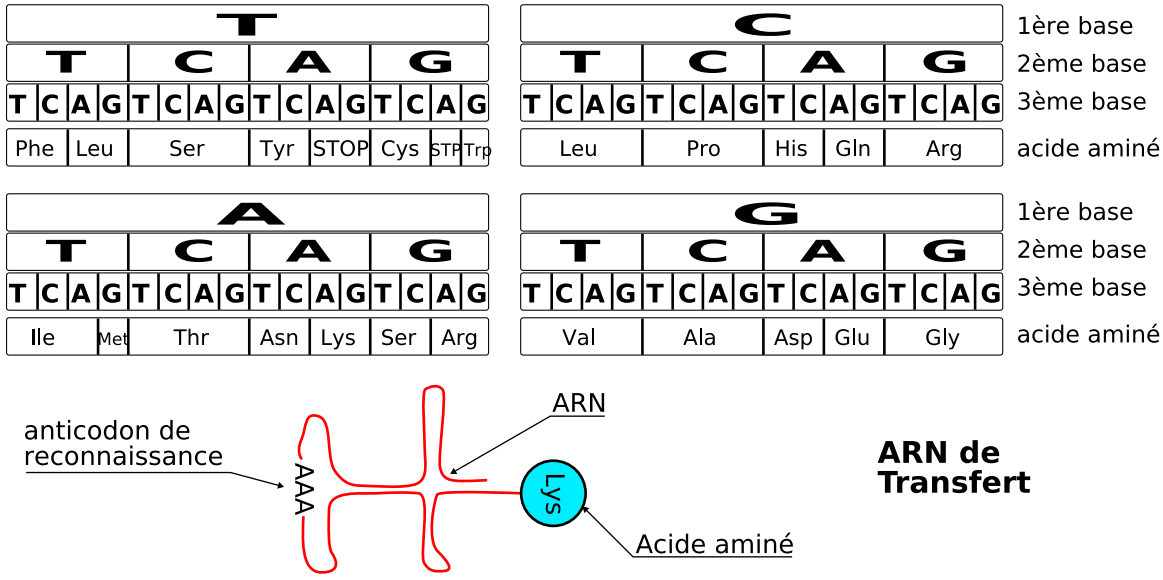

FIG. 3.3: Le code génétique ou la correspondance entre les triplets d'acides ribo-nucléiques et les acides aminés. Pour l'ARN, U remplace le T. Le dictionnaire est en réalité contenu dans des ARN de transfert : si l'anticodon établi des liaisons hydrogènes avec le triplet en cours de lecture par le ribosome, l'acide aminé associé est transféré à la chaîne peptidique en cours d'assemblage.

que ses bases ne peuvent pas former des liaisons hydrogènes. Au contraire, l'ARN se replie toujours pour former des liaisons d'hydrogènes entre ses différents fragments quand ils sont complémentaires[4]. il peut donc avoir des structures tridimensionnelles complexes et en cela il ressemble aux protéines. En particulier, il peut avoir les mêmes propriétés de reconnaissance que ces dernières. Les ARN de transfert reconnaissent chacun, sur leurs extrémité 3', un acide aminé spécifique et s'y lient. Le milieu de l'ARNt présente une boucle de reconnaissance formé de trois bases également. Si ces bases établissent des liaisons hydrogènes avec les trois bases de l'ARNm en cours de lecture par le ribosome, l'acide aminé de l'ARNt est transféré à la chaîne de protéine en cours d'assemblage.

Résumons : l'ARNm est lu par le ribosome, qui est majoritairement formé d'ARNr. Le ribosome catalyse la réaction de transfert d'un acide aminé – attaché à un ARNt – à la chaîne de protéine en cours d'assemblage. L'ARN ne joue pas vraiment les seconds rôles ! Les fonctions diverses assumées par l'ARN peuvent nous incliner à penser que la forme de vie primitive devait être essentiellement à base d'ARN. Ce modèle, appelé "le monde ARN" est des plus en vogue actuellement pour l'apparition de la vie sur terre.

3.3 Développement : ADN → ARN et ADN → ADN

Nous avons deux machines moléculaires (des protéines bien sûr) spécialisées dans la lecture et la copie d'ADN. La première est l'ARNpolymérase (ARN-Pol) qui "lit" un fragment d'ADN et synthétise un fragment d'ARN ; cet ARN est ensuite envoyé vers les ribosomes pour servir à la synthèse de protéines. La deuxième, l'ADN polymérase (ADN-Pol), lit l'ADN pour synthétiser un brin complémentaire d'ADN ; elle est utilisée lors de la duplication de l'ADN pendant la phase de duplication de la cellule. Remarquez la différence entre l'utilisation des deux enzymes : si nous avions une caméra pour filmer l'ADN à l'échelle de nanomètre, nous verrions constamment des ARN-Pol atterrir à de multiples endroits de l'ADN, avancer le long de ce brin en fabriquant une molécule d'ARN, se détacher et aller atterrir ailleurs. Pendant de rares moments (la mitose, ou la division cellulaire), le va et vient des ARN-Pol s'arrêtera soudain, et nous verrons quelques ADN-Pol atterrir à des endroits très spécifiques et commencer la duplication de l'ADN dans son entier.

Le fonctionnement des deux polymérases est très proche (voir fig. 3.4) : D'abord, les deux brins d'ADN sont écartés. Un des deux brins est choisi comme modèle ou empreinte (*template* en anglais). Le polymerase avance alors sur ce brin dans le sens $3' \rightarrow 5'$ et au fur et à mesure synthétise un brin complémentaire dans le sens $5' \rightarrow 3'$ (N'oubliez pas que la polymérisation d'ADN se fait toujours dans ce sens là). Le polymérase choisi la bonne base à inclure dans la chaîne par les liaisons hydrogènes que celle ci peut former avec le brin en cours de lecture. Le brin nouvellement synthétisé est la copie exacte du brin *non-lu* ! Le procédé ressemble à dupliquer une clef : on prend d'abord l'empreinte de la clef (ici, le brin lu est l'empreinte), et

[4]Le problème de trouver le repliement d'ARN qui possède le minimum d'énergie libre (et est donc la forme la plus stable) n'est pas simple à priori : c'est un compromis entre le maximum de liaisons hydrogènes que l'on peut former et le nombre de boucle et leur rayon de courbure qu'il faut créer pour cela. Les boucles coûtent en terme d'énergie élastique. Il existe des programmes spécialisés pour calculer la structure secondaire de l'ARN en fonction de sa séquence et des serveurs sur la toile les mettent librement à la disposition des utilisateurs.

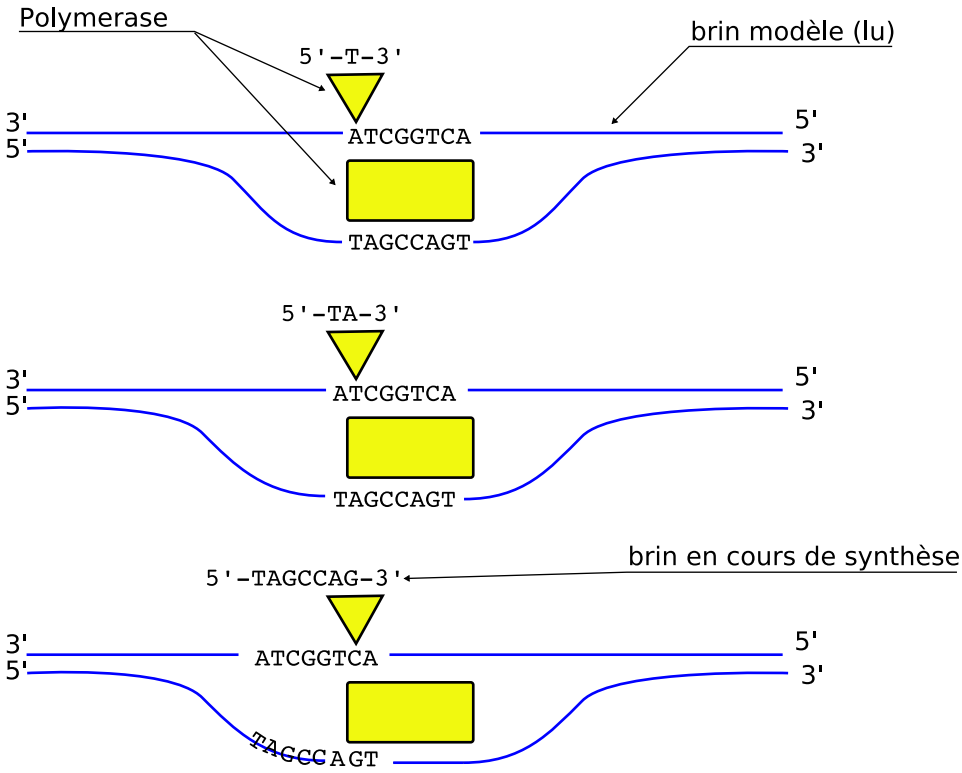

FIG. 3.4: Le fonctionnement des polymérase : les deux brins d'une molécule d'ADN sont écartés. Un des deux brins est choisi pour la lecture. La polymérase avance alors dans le sens $3' \rightarrow 5'$ sur le brin d'ADN , et synthétise un brin complémentaire dans le sens $5' \rightarrow 3'$. Le nouveau brin synthétisé est la copie exacte du brin non-lu de l'ADN original.

ensuite on prend l'empreinte de l'empreinte.

La figure 3.4 montre la synthèse d'ADN. La synthèse d'ARN suit le même chemin, à ceci près que ce sont des acides ribo-nucléique qui sont utilisés, et qu'un U remplace un T.

La synthèse d'ARN pose un problème aigu de positionnement et de début de lecture. La séquence nucléique `5'-AGC-UUC-GUA-CAG-AU...` code pour le polypeptide `Ser-Phe-Val-Gln-...`; Par contre, si l'ARN-Pol avait commencé la lecture de l'ADN une base plus loin, nous aurions eu l'ARN messager `5'-GCU-UCG-UAC-AGA-U...` qui lui, code pour la chaîne peptidique `Ala-Ser-Tyr-Arg-...`! La translation d'une base change complètement la signification. La phrase "I lfai tbea ue tchau" paraît ne pas avoir trop de sens, mais si l'on remarque que nous nous sommes simplement trompé d'une translation, on lira "Il fait beau et chaud". C'est encore plus limpide dans pour l'ARN, puisque les mots sont toujours formés de trois lettres (dans un alphabet qui en comporte quatre). Nous avons donc trois "fenêtres de lecture" (reading frame) pour une séquence donnée, et les mécanismes de lecture de l'ADN sont suffisamment élaborés pour diriger l'ARN pol vers la bonne[5]. Nous verrons ces mécanismes au chapitre 8. Le détail du fonctionnement des polymerases est donné plus loin (voir 5.1.1, 5.1.4).

La dernière subtilité avec l'ARN-Pol est l'arrêt de la transcription. Nous avons vu que dans la synthèse des protéines par les ribosomes (traduction), il existe des séquences particulières voulant dire STOP. Par contre, rien de tel n'existe pour la transcription (synthèse d'ARN à partir de l'ADN). Apparemment, l'ARN en cours de synthèse s'hybride avec lui même et forme des boucles. C'est l'existence de ces boucles qui provoque l'arrêt et le détachement de l'ARN-Pol. Cela impose aux gènes une contrainte supplémentaire vis à vis de l'évolution : non seulement les protéines qu'ils codent doivent être performants, mais en plus, la forme tridimentionnelle de l'ARNm doit être correct pour provoquer l'arrêt de la transcription au bon endroit.

Mentionnons en fin que les eukaryotes n'ont pas une, mais trois ARN-Pol, dont un seul (Pol II) est dédié à la synthèse des ARNm. Les deux autres servent à fabriquer les ARNr et t. Les prokaryotes utilisent un seul ARN-Pol pour tous leurs besoins.

[5] Pour les virus, la longueur d'ADN est un facteur limitant important. Certains ont évolué pour utiliser deux "fenêtres de lecture" : en utilisant (presque) la même séquence, ils construisent deux protéines différentes.

4 Autour d'ADN et ARN.

4.1 Qu'est ce qu'un gène ?

Nous avons donc vu que l'ADN contient des informations, des instructions pour fabriquer des protéines. Le long polymère d'ADN qui porte l'information génétique chez les êtres vivants est appelé *chromosome*. Le chromosome est comme un livre divisé en chapitre. Le chapitre 1152, de la page 1342435 à la page 1342467, contient par exemple les instructions pour fabriquer la protéine Topoisomérase. Bien sûr, les biologistes n'utilisent pas le mot chapitre, mais le mot *gène* ; de même, le mot page est remplacé par la position des bases dans la chaîne linéaire à partir d'une position (facilement repérable) prise comme origine. Il se trouve en plus que la plupart des eukaryotes ont plusieurs chromosomes, c'est à dire plusieurs livres dans leur bibliothèque : 4 pour la drosophile, 17 pour le salamandre, 23 pour l'humain[1]. Ainsi, dans un langage correct, on doit dire que le gène *rhodopsin* qui code pour la protéine Rhodopsin est situé, chez l'humain dans la région 3q21.3. La convention n'est pas unique, mais souvent le gène (la séquence d'ADN) porte le nom de la protéine pour qui il code, mais est écrit en italique. 3q21.3 veut dire qu'il se trouve sur le bras long du chromosome 3 (3q) à la position 21.3.

Tous les gènes connus de toutes les espèces (toute sorte d'information génétique) sont répertoriés dans des bases de données librement accessible sur la toile. Vous pouvez par exemple chercher dans `http ://www.ncbi.nlm.nih.gov/mapview/static/MVstart.html` la séquence, la position, la fonction, ... de tous les gènes que nous rencontrerons dans ce livre.

Pour fixer les idées, le tableau 4.1 donne le nombre de paires de base (bp, pour base pair en anglais) du génome et le nombre de gènes de divers espèces. Le nombre de chromosome n'a pas beaucoup de signification : on peut diviser un génome en beaucoup de petit livre ou en peu de grand. Des espèce très proches qui ont des tailles de génome similaires peuvent avoir un nombre de chromosomes très différent. Les bactéries en générale n'ont cependant qu'un seul chromosome qui est en plus circulaire. *E. Coli* est une (parmi des centaines) des bactéries de l'intestin humain, et c'est probablement l'organisme le plus étudié. La plupart des mécanismes fondamentaux ont d'abord été mis en évidence chez cette bactérie. La levure

[1] Une très bonne source d'information sur le génome humain est le site d'Oak Ridge national laboratory http ://www.ornl.gov/sci/techresources/Human_Genome/posters/chromosome/

	E.Coli	levure	Drosophile	humain	salamandre
# pairs de base	4.5 Mbp	15 Mbp	150 Mbp	3 Gbp	30 Gbp
# gènes	4000		13000	20000-50000 ?	

TAB. 4.1: nombre de pairs de base et de gènes de divers organismes.

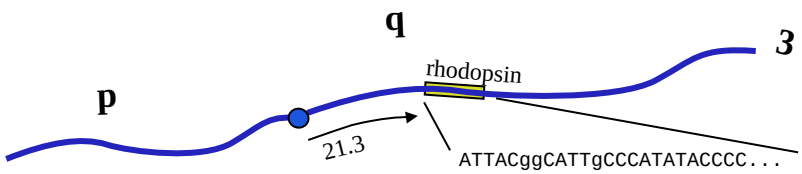

FIG. 4.1: le gène *rhodopsin* se trouve, chez l'humain, sur le bras long (q) du chromosome 3, dans la région 21.3 à partir du centromère. Pour fixer les idées, le chromosome 3 contient $2.14\,10^8$bases (0.21 Gbase) ; le gène *rhodopsin* contient 6700 bases. Le nombre "21.3" ici est une position cytologique : pendant la phase de division cellulaire, le chromosome se replie pour former un objet épais (de l'ordre de 0.5 micromètre) et donc visible au microscope optique. C'est ce qu'on appelle chromosome mitotique. On connaît des marqueurs qui se fixent différemment à différentes région du chromosome mitotique, et qui font que le chromosome apparaisse comme une succession de bandes claires et obscures. Ici, le chiffre 21 est le numéro de la bande où le gène *rhodopsin* est localisé. Il n' y a pas de proportionnalité entre le numéro de la bande et sa position en terme de pair de base. Comme la séquence de l'ensemble du chromosome 3 est connu, on sait que le gène rhodopsin s'étend environ de la base 130568400 à la base 130575100. Dernière petite subtilité : le trait bleu dans cette figure représente un polymère d'ADN *double brin*. Un gène peut se trouver sur un brin ou sur l'autre. Aucun brin ne joue le rôle de maître, portant tous les gènes.

est l'eukaryote unicellulaire responsable de lever le pain et est largement utilisé dans tous les processus de fermentation comme la fabrication de bière ou de certains type de tofu. C'est l'organisme modèle pour l'étude des eukaryotes. La drosophile est une petite mouche qui se nourrit de levure, et qu'on voit donc souvent lors de l'élaboration du vin. Elle est utilisée (avec un petit vers appelé *C. Elegans*) pour l'étude des organismes multicellulaires.

Au début des années 1960, quand la nature chimique de l'ADN et de sa relation avec la fabrication des protéines chez la bactérie avait été élucidé, on avait une vision très simple du génome : un long polymère d'ADN divisé en chapitre contigu de séquences codantes : chapitre 13 (gène *truc*), de la page 89 à la page 91 ; chapitre 14 (gène *bidule*), de la page 92 à la page 104, et ainsi de suite. Le génome des eukaryotes s'est montré cependant beaucoup plus complexe (et parfois plus absurde) et ce n'est qu'a la fin des années 70 qu'on a commencé à avoir une vision plus nette de l'organisation du génome. Il a fallut la mise au point de la plupart des outils de biologie moléculaire que nous verrons au chapitre prochain, et une lente accumulation des connaissances. En résumé : (i) les gènes des eukaryotes ne sont pas contigues, et il peut y avoir beaucoup (beaucoup) de paires de base entre la fin d'un gène et le début d'un autre dont la séquence n'a aucune signification ; (ii) à l'intérieur d'un gène même, il peut y avoir des morceaux de séquence qui ne veulent rien dire. C'est comme si dans un livre, beaucoup de pages étaient simplement un enchaînement aléatoire de lettres. Et les eukaryotes ont élaboré des mécanismes pour ne pas lire ces pages. Chez des eukaryotes tels que l'humain, 2 à 3% seulement du génome est codante, c'est à dire est formé de séquences compréhensibles par les ribosomes. C'est ce que nous verrons ci-dessous.

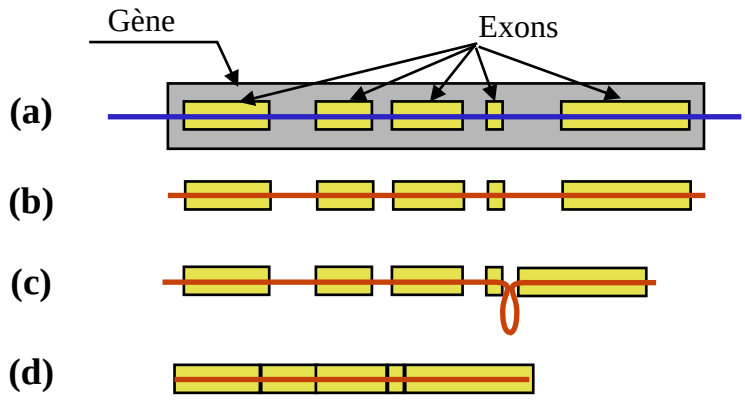

FIG. 4.2: Organisation d'un gène en introns et exons. (a) : un gène, sur-ligné ici en gris, est constitué de séquences codantes appelées exons (sur-ligné en jaune) et de séquences non-codantes appelées introns. (b-d) : fabrication d'un ARNm mature. (b) : le gène est d'abord recopié en ARN prématuré ; (c) les introns sont éliminés par épissage pour donner lieu à (d) : un ARN messager prêt à être exporté en dehors du noyau.

4.2 Les introns et les exons.

Comme nous l'avons dit plus haut, chez une bactérie, un gène est formé entièrement de séquence codante. Chez les eukaryotes, les séquences codantes, qu'on appelle *exons*, sont entrecoupé par des séquences non-codantes qu'on appelle *introns*. La copie du gène se fait en deux étapes : d'abord, l'ensemble du gène, introns et exons, est recopié par un ARN-polymerase sous forme d'ARN. Ensuite, lors d'un processus de maturation qu'on appelle *épissage*, les parties non-codantes de cet ARN sont enlevées et dégradées. L'ARN restant est un ARN messager honnête et a le droit d'être exporté vers l'extérieur du noyau et les ribosomes. Le détail de l'épissage est peu connu. Il existe des enzymes responsable de couper les morceaux en trop et de résoudre les morceaux restants. Ce rôle est parfois assumé par l'ARN lui même, qui peut avoir des propriétés catalytiques. Chez l'humain, un gène est constitué en moyenne de 4 exons et 1350 bases, ou 450 acide aminé pour le produit protéique. Certains gènes sont constitué d'une trentaine d'exons.

Mais comment un enzyme reconnaît les morceaux qui doivent faires parties de l'ARNm et ceux qui doivent être éliminés ? Cela reste encore un peu mystérieux. Il existe des ressemblances entre les séquences des introns, et probablement leur débuts et fins sont signalés par des *tags*. La grammaire de l'ADN n'est pas encore suffisamment connu pour que l'on puisse identifier avec certitude les introns. Cela reste d'ailleurs une des difficultés de trouver les gènes : quand on séquence le génome, on est en possession d'une suite linéaire de lettres A,C,T,G. Pour trouver les gènes, on cherche les longues séquences qui ont une signification. Malheureusement, on peut rarement disposer de très longues séquences codantes, et il faut ensuite décider si une courte séquence, avec un syntaxe correct, fait réellement partie d'un gène ou si il est simplement dû au hasard. Les mathématiciens tentent de mettre au point des critères probabilistes fins afin d'aider les algorithmes à mieux faire ces choix.

Cette organisation absurde du génome chez les eukaryotes peut avoir ses avantages : cela permet de générer plusieurs variantes d'une protéine à partir d'un seul gène. Supposons que nous ayons 5 exons. Dans ce cas, en assemblant les exons 1,2,3,5 ou 1,2,3,4 ou 1,2,4,5 on génère trois protéines différentes. Chez la drosophile, un seul gène code pour une douzaine de variantes de la protéine dynéine (un moteur moléculaire).

Certains ont émis l'hypothèse que les exons correspondent aux sous-domaines fonctionnelles de protéines. Les maintenir séparés par des introns permet de les échanger et partager facilement entre différents gènes. Cette affirmation cependant n'a pas de caractère très générale, et la plupart des exons sont uniques à leur gène, même si on tient compte des mutations qui se sont introduites lors de l'évolution et qui ont pu gommer les ressemblances.

4.3 l'ADN "poubelle" ?

Nous avons parlé de ce qui a à l'intérieur d'un gène. Mais qu'est ce qu'il y a entre les gènes ? Quelle est l'utilité de ces 95% de séquences non-codantes chez l'humain ? La réponse est surprenante : en grande partie, aucune pour l'organisme hôte. Le génome des eukaryotes telles que nous, le *xenopus* ou la séquoia est un écosystème à lui même, constitué de fossile de gènes qui ont perdu leurs sens, de traces de virus, des séquences parasites auto duplicateurs et des séquences parasites de parasites. L'ensemble de ces éléments constitue plus de la moitié de notre génome. On connaît peu la signification de l'autre moitié (à part les 2-3% de séquences codantes), on suspecte qu'elle a un rôle de régulation et de structuration physique[2].

4.4 Eléments mobiles et génération d'anticorps.

4.5 La physique de l'ADN

4.5.1 compactage et territoire

4.5.2 Etirement de molécule unique.

4.6 Ne négligeons pas l'ARN

[2]Quand on parle en terme aussi vague, c'est que l'on ne sait vraiment pas grand chose.

5 Les outils de la biologie moléculaire.

La majorité des expériences de biologie moléculaire utilisent les quelques outils de base que nous verrons ci-dessous. Leurs compréhension est nécessaire, et presque suffisante, pour travailler dans un laboratoire. Nous présentons d'abord ces outils (enzymes, électrophorèse,PCR) et montrons ensuite quelques unes de leur utilisation (ADN recombinant, séquençage, synthèse). L'ensemble constitue le copier-coller-dupliquer de la biologie moléculaire.

5.1 Les enzymes.

En physique, nous disposons d'instruments très sophistiqués, dont la maîtrise de certains demande un très long effort : des spectromètres très fins résolu en temps, la RMN, la diffusion de la lumière, ... La visite de n'importe quelle laboratoire de physique vous en fournira une liste non-exhaustive. La biologie moléculaire au contraire possède très peu d'instruments, rien en tout cas dont l'utilisation ne puisse être maîtrisé après un bac de lettre. Ces instruments sont des enzymes extraits du monde vivant. Ces enzymes sont de véritable magiciens nanométrique, et on ne comprend pas toujours exactement leurs fonctionnements, mais on sait s'en servir. C'est un peu comme conduire une voiture : la plupart d'entre nous ne savons pas exactement ce qu'il y a sous le capot, même si les principes générales nous sont connus. Cependant, nous savons parfaitement nous en servir pour aller d'un point à un autre.

5.1.1 Les polymerases.

Comment fabriquer une nouvelle molécule d'ADN ou d'ARN ? Watson et Crick ont mentionné dès le début, dans leur article de 1953 sur l'ADN, que la forme même de cette molécule peut être une indication pour sa synthèse : chaque brin devrait servir de *moule* pour fabriquer son complémentaire[1]. Nous connaissons aujourd'hui, grâce au travail acharné de nombreux biochimistes, les enzymes qui dirigent ces actions de synthèse.

La première enzyme de synthèse est découvert en 1955. Elle s'appelle PNPase, elle synthétise l'ARN en incluant de façon aléatoire les nucléotides présents dans la solution. Niremberg et Matthaie, dans leur course pour briser le code génétique, l'ont utilisé pour synthétiser des ARN polyU en 1961 et établir ainsi la relation entre le triplet UUU et l'acide aminé phenylalanine[2].

[1] "It has not escaped our notice that the specific pairing that we have postulated immediately suggests a possible copying mechanism for the genetic material".

[2] Ces ARN poly U était ajoutés ensuite aux extraits qui contenaient des ribosomes et des acides aminés. Après quelques heures, ils trouvaient alors dans la solution des polypeptides poly-phenylalanine. On croyait déjà savoir que le code génétique est contenu dans des triplets. D'où la déduction mentionnée ci-dessus. Ce travail

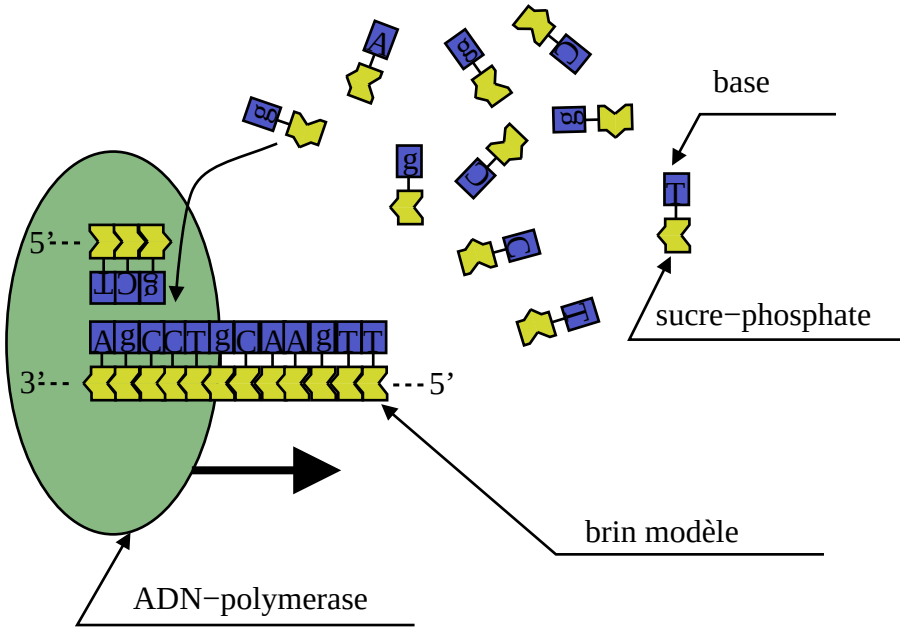

FIG. 5.1: Synthèse d'un nouveau brin d'ADN à partir d'un modèle. Une molécule d'ADN-polymerase *lit* le brin modèle à chaque étape et choisit et incorpore le nucléotide complémentaire présent en solution. La synthèse de l'ADN se fait toujours dans la direction $5' \rightarrow 3'$, l'ADN polymérase lit le brin modèle et avance donc dans la direction $3' \rightarrow 5'$.

Mais le maître absolu, l'enzyme de réplication du génome, mesdames et messieurs j'ai nommé l'ADN polymerase, a été découvert en 1958 après le travail minutieux des biochimistes Arthur Kornberg et Robert Lehman de l'Université de Stanford.

La réaction de base peut s'écrire

$$(\mathrm{ADN})_n + \mathrm{NTP} \rightarrow (\mathrm{ADN})_{n+1} + \mathrm{P_i} + \mathrm{P_i}$$

Une chaîne d'ADN de longueur *n* incorpore une nucléotide (présent en solution comme NTP un nucléotide tri-phosphate) et s'allonge d'une unité. Cette réaction est énergétiquement favorable. Le travail de l'ADN-polymerase est de choisir le bon nucléotide parmi les 4 (ATP, CTP, TTP, GTP) présent dans la solution, en lisant le brin complémentaire. Nous verrons la biochimie de ce processus plus bas. Ce qui est essentiel à retenir ici est que l'ADN-polymerase ne fonctionne qu'en présence d'un brin d'ADN utilisé comme modèle (voir fig. 5.1). L'autre élément important est que l'ADN-polymérase ne fonctionne que si il y a déjà une séquence synthétisée, et correctement hybridée au brin modèle *derrière lui* ! La longueur minimale de cette séquence est de trois à quatre bases. On pense que cet effet contribue à la grande précision de l'ADN-polymerase.

Parlons un peu de la précision. L'ADN-polymerase est là pour dupliquer le génome et le

a fait l'effet d'une bombe dans la communauté scientifique, et le reste du code a été déchiffré dans les 3 années qui ont suivi.

transmettre à la génération d'après. Il doit donc être le plus précis possible[3]. Son taux d'erreur est de l'ordre de 10^{-9}, c'est à dire que tous les milliards de base, il commet une erreur. Essayez donc de recopier 1000 livres en ne commettant qu'une seule faute pour avoir une idée de la précision de cette machine. Les considération énergétiques - un mauvais base pairing coûte deux à trois liaisons hydrogènes - donne un taux d'erreur de l'ordre de 10^{-4}. Il existe d'autres mécanismes de relecture qui permettent d'augmenter la précision. Nous exposerons le cadre théorique générale de la correction d'erreur au chapitre sur la cinétique chimique. Mais on peut constater que l'existence d'une séquence déjà hybridée derrière participe à ces mécanismes. En effet, supposons que l'ADN-polymerase inclue une mauvaise nucléotide et avance d'une séquence. La séquence qu'il a alors derrière lui n'est pas correctement hybridée et arrête le fonctionnement de l'ADN-pol. L'enzyme doit alors faire marche arrière, enlever la mauvaise base, et recommencer. Aussi surprenant que cela puisse paraître, c'est vraiment comme cela que ça se passe.

In vivo, pour dupliquer une molécule d'ADN, il faut séparer à un endroit les deux brins d'ADN et commencer à dupliquer chaque brin séparément. Le point de départ s'appelle l'origine de réplication. Nous laissons au lecteur comme exercice de voir comment on peut dupliquer un double brin d'ADN, sachant que l'ADN-pol ne se déplace que dans la direction $3' \rightarrow 5'$. Mentionnons également au passage que les deux brins étant enroulés en double hélice, l'avancé de l'ADN-pol accumule des contraintes topologiques, et ils existent des enzymes (bien sûr) pour enlever ces contraintes. *In vitro* le problème ne se pose pas, puisqu'en élevant la température (autour de 60 à 90 $^{\circ}$C) on sépare les deux brins.

L'ARN polymérase est un cousin très proche de l'ADN polymérase, à cette différence qu'il lit un brin d'ADN pour le copier en ARN. Il fut découvert en 1960 par plusieurs groupe de façon presque simultanée. Notons qu'ARN-pol est nettement moins précis que l'ADN-pol et ne possède pas plusieurs des mécanisme de correction d'erreurs de ce dernier. Enfin, le dernier enzyme de cette classe est appelé *reverse transcriptase* et qui polymérise l'ADN en lisant *l'ARN* ! Cet enzyme est propre aux retrovirus dont le matériel génétique est l'ARN et non l'ADN. La génie génétique l'utilise couramment de nos jours, comme nous le verrons par la suite.

5.1.2 Ligase

Cet enzyme, découvert en 1967, est utilisé pour "coller" deux brins d'ADN en créant un liaison covalente entre l'extrémité 3' de l'un et 5' de l'autre. Les deux morceaux doivent être immobilisés sur le brin complémentaire (voir Fig.5.2). *In vivo* cet enzyme est utilisé lors de la duplication d'ADN. Comme le chromosome est assez long, souvent sa duplication commence à plusieurs endroit et est mené en parallèle. La ligase intervient pour attacher ces copies partielles les unes aux autres.

[3]Pas tout à fait : si il était absolument précis, il n'y aurait pratiquement pas d'évolution. Ce sont ces petites erreurs, des mutations, qui s'accumulent dans le génomes et génèrent la diversité. On pense aujourd'hui que la précisons de l'ADN-polymerase elle même est un paramètre affiné au cours de l'évolution : pas trop fort, sinon la plupart des descendants seraient défectueux, mais pas trop faible pour générer de la diversité. Nous discuterons tout cela plus en détail au chapitre consacré à l'évolution.

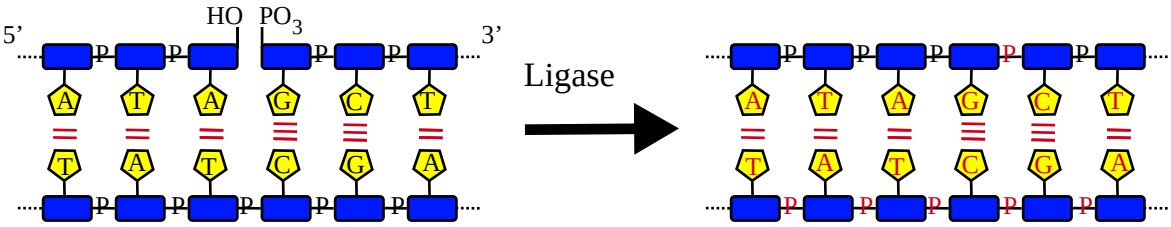

FIG. 5.2: L'action de la ligase est de former une liaison covalente phosphodiester entre deux morceaux d'ADN, immobilisés sur le brin complémentaire qui lui, est intact.

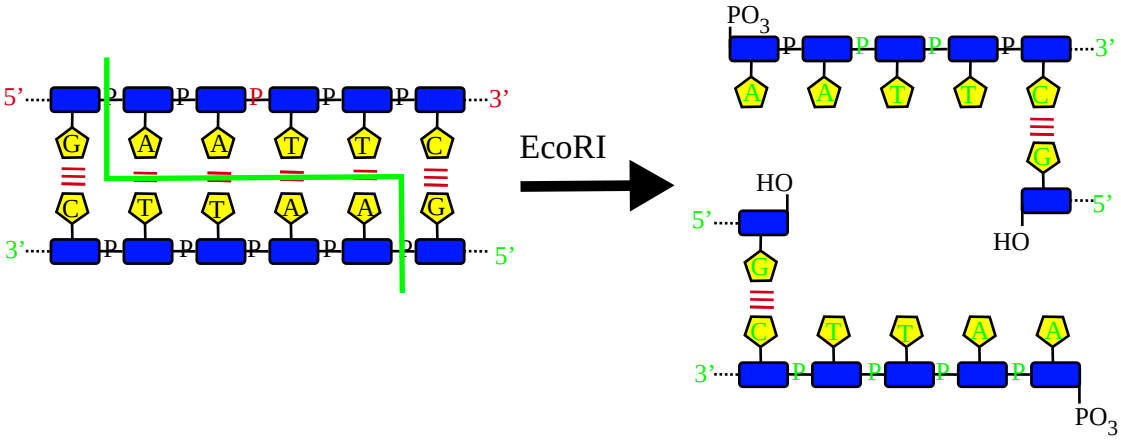

FIG. 5.3: L'action des enzymes de restriction est de couper un double brins d'ADN à un endroit spécifique. Ici, nous montrons l'action de EcoRI et son site de reconnaissance. Cet enzyme produit des bouts collants.

5.1.3 Enzymes de restrictions.

Les enzymes de restriction sont les ciseaux moléculaires, coupant l'ADN à des endroits bien précis. Nous en connaissons aujourd'hui des centaines de ces enzymes, chacun reconnaissant une séquence spécifique d'ADN pour le couper à cet endroit. La sélectivité de ces enzymes varie : certain n'ont besoin que de reconnaître une séquence de 5-6 paire de base, tandis que d'autres utilise plus d'une vingtaine. Certains produisent des bouts "collants", en laissant des petits morceaux d'ADN simple brins aux extrémités (Fig. 5.3). Nous verrons le très grand avantage de cela plus bas, dans la section consacrée à l'ADN recombinant.

5.1.4 Développement : la biochimie de ces enzymes.

Les principes généraux c'est bien joli, mais pour comprendre un peu le détail, il faut parler un minimum de la chimie. La figure 5.4 montre la structure d'un nucléotide : un sucre (ribose) et un phosphate forme le squelette principal, tandis qu'une acide nucléique (A,T,C ou G) lui donne son identité. La réaction de polymérisation se fait en ajoutant un nouveau NTP sur le carbone 3' d'une ancienne chaîne, en relâchant deux groupes phosphates (qui sont la source d'énergie qui pousse la réaction et qui est utilisée par l'ADN ou ARN polymerase). Le travail

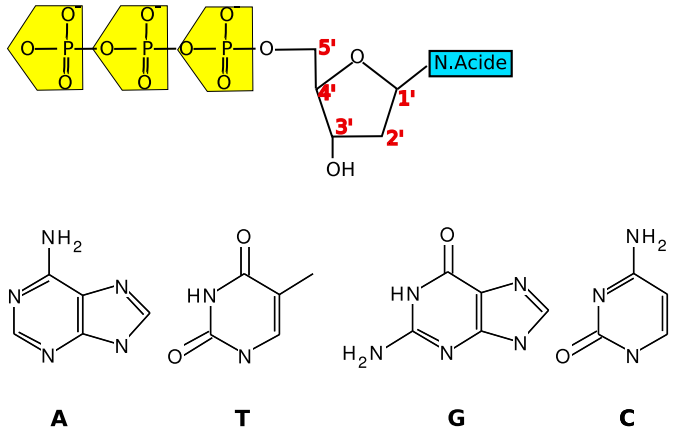

FIG. 5.4: La structure d'un nucléotide tri-phosphate : Un sucre, trois groupes phosphate et un acide nucléique. La structure des quatre acide nucléique est montré sur la deuxième ligne. Le lecteur notera la complémentarité spatiale entre AT et CG.
Les carbones du sucre sont numérotés à partir de l'oxygène. Le sucre ici est un desoxyribose qui intervient dans l'ADN. Pour l'ARN, nous avons un ribose, avec un groupe OH sur le 2'. Parfois (pour le séquençage) nous avons besoin d'une version non polymerisable de NTP ; on utilise alors du didesoxyribose, ou le groupe OH sur le 3' est remplacé par un H.

du Pol est donc de sélectionner le bon NTP dans la solution (en lisant le brin complémentaire) et le présenter sur le coté 3' de la chaîne en fabrication. L'élongation se fait *toujours* dans le sens $5' \rightarrow 3'$ (Figure 5.5).

5.2 Electrophorèse

l'électrophorèse en gel consiste à faire migrer des molécules linéaires dans un matrice poreuse : plus la chaîne est longue, plus la migration est lente et on arrive ainsi à trier les molécules par leurs taille. La technique est non destructive, et on peut récupérer les molécules après les avoir trier, si on en a besoin. Dans son principe, cette technique n'est pas loin de la chromatographie. La technique a été inventé dans les années 50 par Oliver Smithies.

Reprenons. D'abord, les molécules en question sont l'ADN ou l'ARN, qui sont bien sûr de longues chaînes polymériques. On peut également faire de l'électrophorèse avec des protéines (et on en fait beaucoup), mais il faut d'abord les dénaturer à l'aide de solvants pour leur enlever leur structures tridimensionnelles et les rendre souple. Ces molécules sont *chargées*, on peut donc les faires migrer à l'aide d'un champ électrique. La matrice poreuse est souvent faite d'un gel comme l'agarose ou la polyacrilamide : une fois la migration finie, on peut découper la bande qui nous intéresse, dissoudre le gel et récupérer notre molécule (Fig.5.6).

La charge d'une molécule d'ADN est due à son backbone sucre-phosphate, elle est donc proportionnelle à la longueur de la molécule. Les forces de frottement qui agissent sur la molécule lors de sa migration sont plus que proportionnelle à la longueur de l'ADN. Par

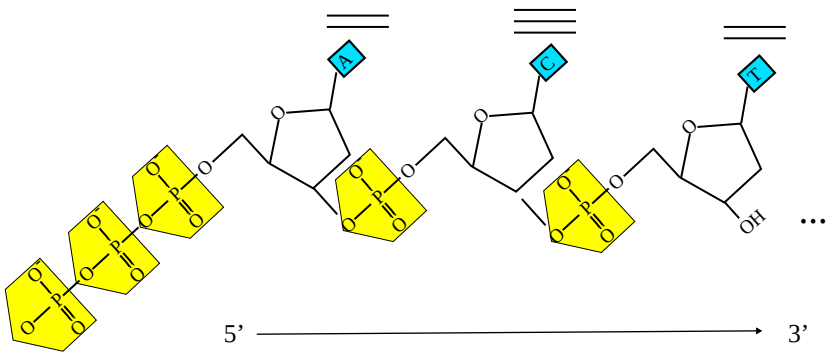

FIG. 5.5: La réaction de polymérisation se fait entre le carbone 3' d'une nucléotide déjà inséré dans la chaîne et le groupe phosphate d'un nouveau nucléotide à insérer dans la chaîne, en formant une liaison phosphodiester, et en relâchant de groupe phosphate. C'est la différence d'énergie des liaisons hydrogène avec le brin complémentaire (pas montré ici) qui guide la polymerase dans son choix du "bon" nucléotide. La réaction médiée par les enzymes de restriction est exactement l'inverse et consiste à détruire une liaison phosphodiester pour laisser un groupe phosphate d'un coté et un groupe OH de l'autre.

conséquent, plus une molécule d'ADN est longue, plus elle migre lentement.

Pour résumer, l'électrophorèse en gel est un appareil qui nous permet de séparer un mélange de molécules d'ADN en fonction de leurs tailles. Si les pores du gel sont très petites, le pouvoir de résolution peut être *d'une seule base* : on peut séparer une séquence de 45 bases d'une séquence de 46, et c'est exactement cela qu'on utilise pour séquencer l'ADN. Le problème avec cela est qu'on ne peut pas faire cela sur des longues séquences d'ADN, disons plus que 300 bases. Cela est dû aux problèmes de piégeage dont un exemple est donné dans la figure (5.6.a). En général, on n'a pas besoin d'autant de résolution et on prépare des gels avec des pores plus grandes. Savoir préparer ses gels est pour le biologiste moléculaire ce que savoir préparer ses couleurs est pour le peintre.

5.3 PCR

Les outils décrit précédemment nous permettent maintenant d'achever la dernière étapes qui nous manque : l'amplification, ou l'art de copier des millions de fois un fragment d'ADN donné. La technique, appelé PCR (Polymerase Chain Reaction) est une amplification exponentielle. Elle est d'une telle simplicité que l'on s'étonne qu'il ait fallu attendre jusque 1985 pour qu'elle soit découverte[4]. Cette technique a révolutionné la biologie moléculaire.

Supposons que nous avons un fragment d'ADN double brin (ADNdb ou dsDNA en anglais) dont la séquence est connu, ou tout au moins la séquence de ses extrémités (quelques paire de base de chaque coté suffisent). Comment en fabriquer une copie exacte *in vitro* ? Evidemment, nous devons utiliser l'ADN polymerase, mais comme le lecteur se souvient sûrement, l'ADN-

[4]et a valu à son découvreur, Mullis, non seulement le prix Nobel, mais des brevets extrêmement juteux.

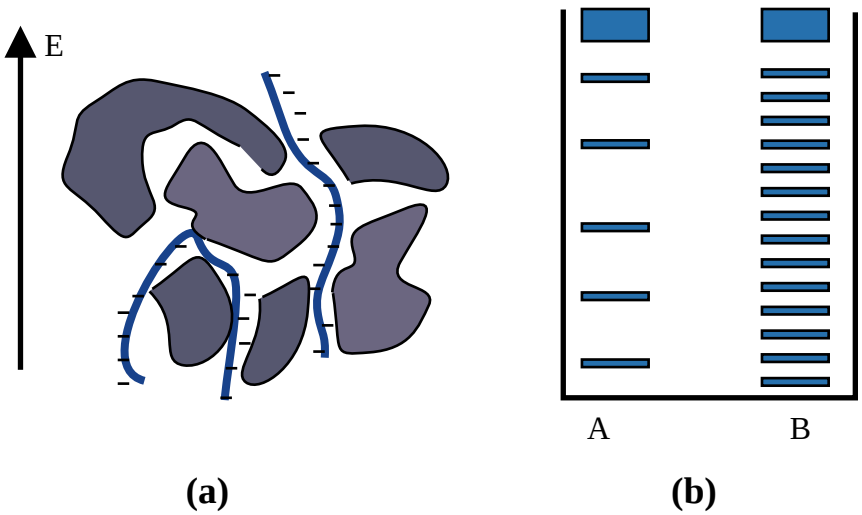

FIG. 5.6: Principe de l'électrophorèse en gel. (a) le gel est montré à l'échelle de quelques nanomètres, où les chaînes d'ADN (en bleu) migrent dans le champs. En bas à gauche, on montre le cas d'une molécule piégée : les forces s'exerçant sur ses deux bouts son à peu près égales, et elle peut rester bloquée dans cette position. (b) l'appareil de l'électrophorèse schématisé, d'une largeur d'environ 20 à 40 cm. Le haut et le bas sont branchés à des électrodes pour maintenir une différence de potentielle et créer un champs électrique à l'intérieur du gel. Au temps zéro, le mélange d'ADN (ou ARN, ou protéine) est chargée en haut (colonne A). Au bout d'un certain temps (de l'ordre de l'heure, toujours trop long de toute façon), les divers constituants migrent sous l'effet du champs, mais à des vitesses différentes, et se trouve donc à divers position du gel. Comme on ne connaît pas à priori leurs tailles, on charge toujours en parallèle (colonne B) un mélange de molécules d'ADN de longueurs connues. La colonne B constitue donc la règle qui nous permet de *lire* la taille des divers bandes de la colonne A. Toutes ces molécules sont bien sûr transparentes, et il existe divers méthode pour les visualiser : soit on les a rendu radio-actives par avance, et il suffit alors de poser une plaque photographique sur le gel, soit on doit, après la fin de la migration, utiliser des marqueurs visibles qui se lient spécifiquement à l'ADN (ou ARN, ou...). La taille des pores est un paramètre ajustable et dépend des besoins de l'utilisateur en pouvoir de résolution.

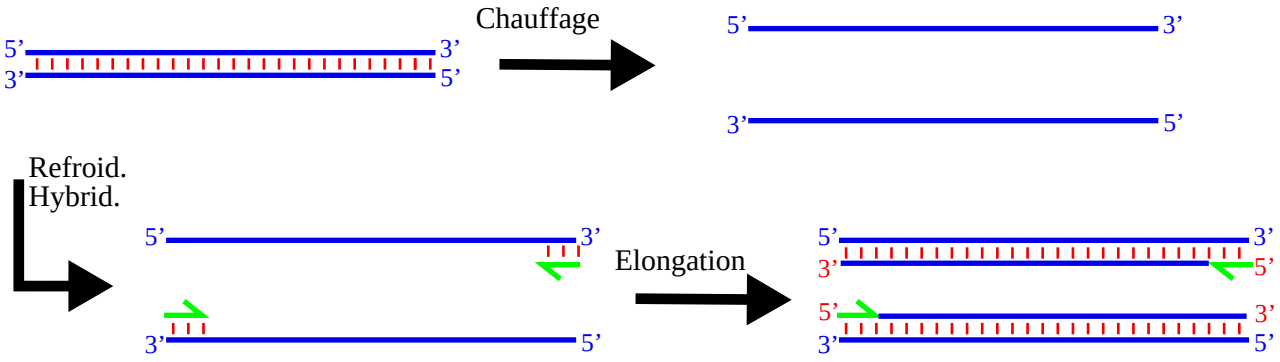

FIG. 5.7: Le principe du PCR. A basse température, les deux brins du fragment d'ADN sont liés par des liaisons hydrogène. On les sépare en les portant à haute température. En refroidissant ensuite, on permet aux amorces de s'hybrider (former des liaisons hydrogènes) avec les extrémités 3' de chaque brin. L'ADN polymerase peut alors utiliser ces amorces pour polymériser un nouveau brin, complémentaire de celui qu'il lit.

Pol ne fonctionnent que si il y a déjà quelques séquences polymérisées derrière lui. Pour cela, on utilise des *amorces* (primers en anglais), des petites séquences de 4-5 pair de base qui sont complémentaires aux extrémités 3' de chaque brins. Voilà comment on procède :

On démarre avec une solution contenant l'ADN-Pol, beaucoup de NTP (les quatre bases, évidement) et beaucoup d'amorces, dans laquelle on balance notre fragment d'ADNdb. On chauffe alors notre solution autour de 90˚C, ce qui a pour effet de séparer les deux brins. On refroidit alors la solution aux alentours de 60˚C. Les amorces peuvent alors s'hybrider avec les extrémités 3' et l'ADN-Pol peut commencer son travail en polymérisant un brin tout neuf complémentaire du brin qu'il lit. Nous sommes donc maintenant en possession de *deux* ADNdb, strictement similaire à l'original avec lequel on avait commencé (Fig. 5.7). Il suffit maintenant de chauffer à nouveaux à 90˚C et recommencer le cycle pour à la fin, avoir 4 copies et ainsi de suite. Le nombre de copie final sera 2^n, où n est le nombre de cycles . Comme les amorces sont consommées à chaque cycle, il va de soi qu'il faut en mettre plus que 2^n.

Le PCR met grandement à profit la nécessité pour l'ADN-Pol de démarrer avec des amorces déjà hybridées. Si nous avons plusieurs fragments d'ADN différent dans la solution, seul le fragment qui correspond aux bonnes amorces sera amplifié. Par exemple, pour déterminer si un patient a été infecté par un virus ou une bactérie donné (bien avant que les symptômes de la maladie apparaissent), on purifie l'ADN issue du patient, et on amplifie avec des amorces spécifiquement préparées pour le génome du pathogène. Des traces infimes de l'ADN du pathogène peuvent ainsi être détectées.

Il n'a pas échappé au lecteur que nous avons parlé des températures de 60 à 90˚C. Mais comment la protéine ADN-Pol peut rester stable à ces températures, sans parler de son fonctionnement ? En fait, cela était la clef de la technique. Dans les versions primitives du PCR, ADN-Pol était ajouté à chaque cycle, pendant la phase basse température. Mais il existe des bactéries qui ne vivent que dans des conditions extrême, proche des sources chaude au fond

des océans[5] et qui possède des versions thermostable de l'ADN polymérase. C'est la purification de ces polymerases (un des premiers et plus fameux s'appelle Taq-polymerase) qui a réellement lancé le PCR.

Le PCR a bien sûr ses limites. On ne peut amplifier des fragments de plus que quelques kilobases pour deux raisons : (i) *in vitro*, les polymerases manquent quelques mécanismes de correction d'erreur. Taq par exemple a un taux d'erreur de 10^{-4} (à comparer à 10^{-9}). (ii) Le polymerase se détache de l'ADN au bout de quelques kilobases. *In vivo* il existe des mécanismes qui le garde accroché. Beaucoup d'efforts technologiques ont permis de pousser ces limites de plus en plus loin, mais on en est là de nos jours.

Nous avons dit également plus haut que les séquences des extrémités doivent être connues. Cela est souvent le cas, puisque les fragments d'ADN ont été produit en "coupant" un long fragment par des enzymes de restriction. Les extrémités du fragment à amplifier sont donc ceux qui sont reconnu par l'enzyme de restriction utilisé.

Enfin, notons que nous pouvons amplifier non pas tout le fragment d'ADN, mais seulement une fenêtre à l'intérieur qui nous intéresse. Il suffit pour cela que l'on utilise des amorces correspondant aux extrémités de la fenêtre en question : seul la fenêtre sera amplifié exponentiellement, les autres morceaux ne seront amplifiés que linéairement. Nous laissons la démonstration de cela au lecteur.

5.4 L'ADN recombinant

Comment insérer un fragment donné d'ADN au milieu d'un autre ? Cela est maintenant le B.A. BA de la biologie moléculaire. Par exemple, pour produire en grande quantité une protéine donné, il suffit d'insérer la séquence d'ADN qui code pour cette protéine dans un plasmide[6]. Une bactérie munie de ce plasmide produira la protéine en question et deviendra ainsi une usine biologique. C'est également comme cela que l'on fabrique les organismes génétiquement modifiés, en incluant dans leur génome un gène intéressant provenant d'un autre espèce. Le maïs par exemple peut produire un insecticide dont le gène a été péché chez un algue. L'ADN recombinant est un outil inestimable pour la recherche. On peut par exemple fusionner le gène d'une protéine fluorescent avec le gène d'une protéine que nous sommes en train d'étudier et suivre en temps réel l'expression de cette dernière et sa localisation spatiale.

L'insertion d'un fragment d'ADN dans un autre s'appelle ADN recombinant. Elle a été réalisée la première fois en 1975 et a provoqué alors un vent de frayeur par le potentiel dévastateur qu'elle pouvait avoir. Les divers limites et dangers de la technique ont été depuis caractérisés et elle est couramment pratiqué dans les laboratoires de biologie moléculaire à travers le monde.

Le principe de la technique est très simple. Supposons que nous voulons inclure le gène G d'un organisme X dans un fragment d'ADN A (un plasmide ou le génome d'un virus). On

[5] Certaines ne peuvent vivre qu'a des températures entre 102 et 110 ˚C.

[6] Les plasmide sont de "petits"fragment d'ADN circulaire. En général, une bactérie possède un chromosome de quelques mégabase qui code pour l'essentiel de ses gènes, et plusieurs plasmides qui codent pour des gènes moins critiques. Les plasmides peuvent être vu comme des jetons que les bactéries sont capable de s'échanger assez facilement et propager par exemple la résistance à un antibiotique donné.

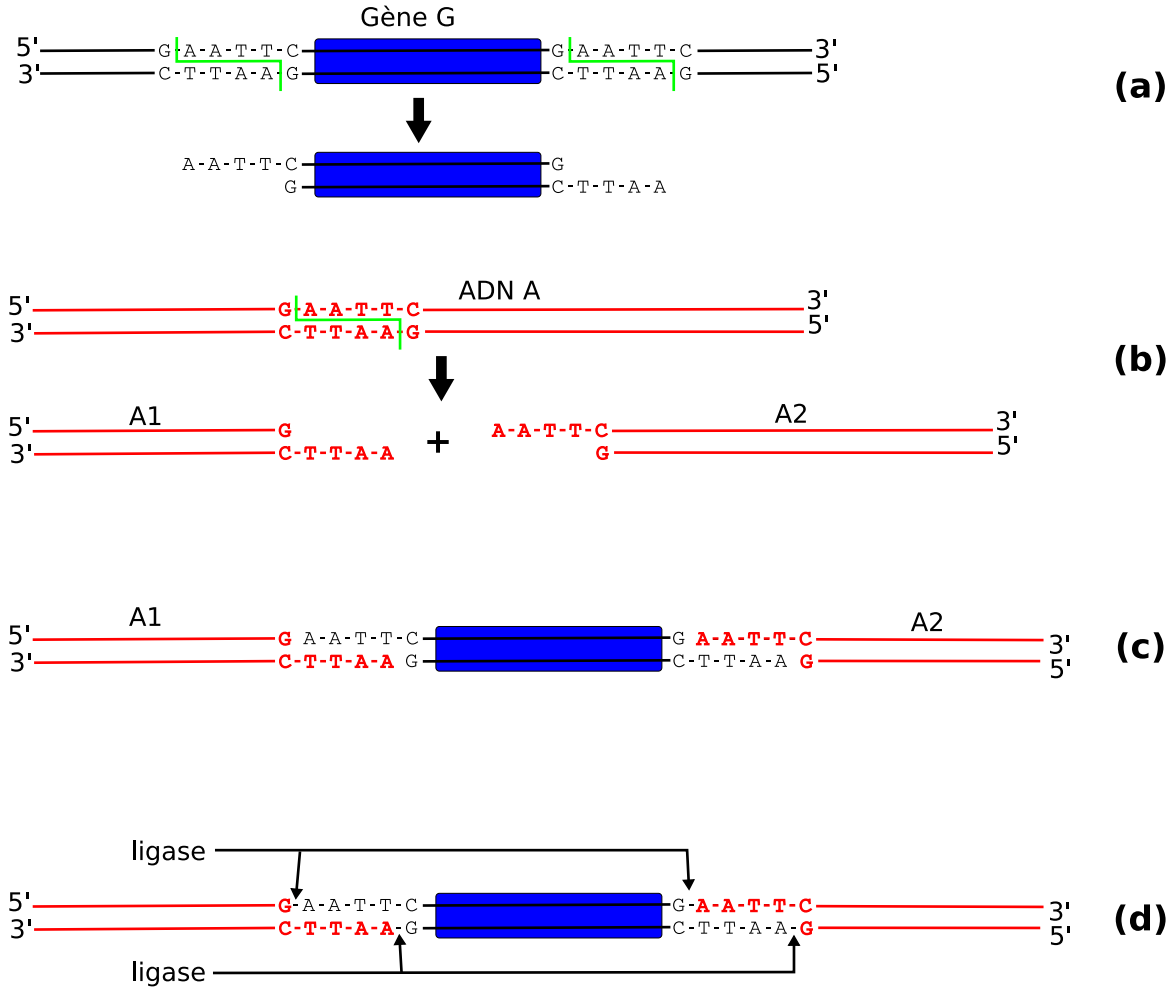

FIG. 5.8: la technique d'ADN recombinant. (a) Un enzyme de restriction est choisi pour couper une fenêtre avec des extrémités collantes autour du gène intéressant *G*. (b) L'ADN *A* dans lequel on veut effectuer l'insertion est coupé avec le même enzyme de restriction. (c) Le gène *G* peut maintenant s'hybrider avec le fragment *A*1sur la gauche et le fragment *A*2 sur la droite. (d) Finalement, la ligase soude les jonctions.

choisit d'abord un enzyme de restriction adéquat capable de couper une fenêtre incluant le gène *G*. Comme nous avons plusieurs centaines d'enzymes de restriction à notre disposition, cet étape ne pose pas de problème. Il est nécessaire que l'enzyme coupe l'ADN en laissant des "bouts collants" (voir plus haut, les enzymes de restriction). Une étape d'électrophorèse permet alors de purifier le gène *G* (éventuellement suivi d'une étape de PCR pour l'amplifier).

On utilise ensuite le *même* enzyme de restriction pour couper l'ADN *A*. On dispose dans la solution les deux fragments issue de la coupure *avec* le fragment *G*. Puisqu'ils ont des extrémités complémentaires, le gène *G* peut s'hybrider avec les deux fragments. L'utilisation de la ligase permet ensuite de souder les jonctions (Fig. 5.8). Il y a bien sûr plusieurs possibilités de recombinaison, et la bonne est sélectionnée par sa longueur lors d'une étape d'électrophorèse.

5.5 Séquençage d'ADN

Comment connaître la séquence d'un fragment d'ADN donné ? De nos jours, de grand progrès sont fait dans le domaine du séquençage et le génome de beaucoup d'organisme, y compris l'humain, la souris, la mouche drosophile, la levure, la bactérie *E.Coli*, ... est entièrement décodé. La technique utilisée de nos jours a été mis au point par Sanger en 1977 et nous verrons ci-dessous sa version moderne. La technique utilise l'ADN-polymérase, l'électrophorèse (comme filtre séparateur de longueur) et des nucléotides modifiées.

Les nucléotides dont ont parle sont des didesoxyriboses (voire Fig. 5.4) où le groupe OH sur le carbone 3' a été remplacé par un simple H. Si un tel nucléotide est inséré dans une chaîne en cours de polymérisation, l'élongation s'arrête tout de suite (le lecteur se souvient sûrement que les nouveaux nucléotides sont toujours ajoutés à l'extrémité 3' d'une chaîne en cours d'élongation). De plus, on associe de petites molécules fluorescentes à ces ddNTP, et on choisit quatre couleurs différentes pour chacune des bases : par exemple ddATP en rouge, ddTTP en bleu, ddCTP en vert et ddGTP en magenta. On mélange alors ces ddNTP fluorescents en faible proportions avec des NTP normaux dans une solution de polymérisation adéquate contenant l'ADN polymerase. On ajoute à cette solution l'ADN *simple brin* dont on veut connaître la séquence et des amorces complémentaires à son extrémité 3'. On laisse alors la réaction de polymérisation démarrer. Une fois l'amorce hybridée à l'extrémité 3', ADN-Pol prend des NTP dans la solution et les inclut à la chaîne en élongation. De temps en temps, au lieu de choisir un NTP, il choisit un ddNTP qu'il inclut dans la chaîne[7]. La réaction pour cette chaîne en particulier s'arrête alors, l'ADN-Pol se détache et va se trouver un autre ADN à allonger. Une fois que le temps s'est suffisamment écoulé, nous disposons des chaînes de toutes les longueurs entre 1 et la longueur de l'ADN à séquencer. L'astuce est que les chaînes qui se sont arrêtée après l'inclusion d'un ddATP apparaissent en rouge, celles s'étant arrêté après un ddTTP en bleu est ainsi de suite. Il suffit alors d'effectuer un électrophorèse pour séparer les longueurs et de les lire optiquement pour avoir la séquence (voir Fig. 5.9). Ces manipulations sont devenues routines et complètement automatisées grâce à des robots, et personne ne séquence l'ADN dans son laboratoire : il existe des dizaines de compagnies spécialisées dans le séquençage à qui on envoie le fragment d'ADN à séquencer par la poste

[7] La fréquence de ces évènement dépend de la proportions des ddNTP par rapport aux NTP.

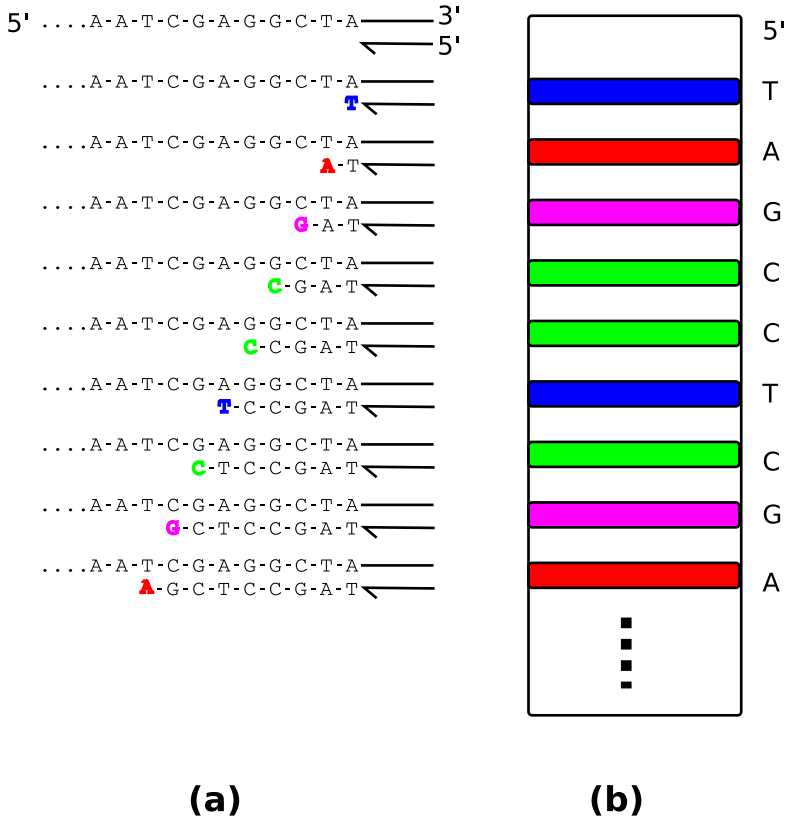

FIG. 5.9: Principe du séquençage. Dans une solution contenant l'ADN simple brin à séquencer, une amorce complémentaire à son extrémité 3', des NTP et de l'ADN polymérase, des ddNTP en faible proportion sont ajoutés. Ces ddNTP arrêtent l'élongation si ils sont inclus dans la chaîne. De plus, chaque ddNTP contient une molécule fluorescente dont la couleur est fonction de la base. Ici, ddATP est rouge, ddTTP bleu, ddCTP vert et ddGTP magenta.
(a) En haut est montré l'ADN simple brin à séquencer, hybridé à son amorce. A la suite de la réaction d'élongation, des chaînes de toutes les longueurs sont produites. Cela dépend de l'évènement aléatoire de l'inclusion d'un ddNTP au lieu d'un NTP.
(b) Ces divers fragment d'ADN peuvent être séparés par un électrophorèse en gel. Il suffit ensuite de se souvenir du code de couleur utilisé pour "lire" la séquence dans le sens $5' \rightarrow 3'$.

et qui renvoient le résultat par courrier électronique dans les deux trois jours suivants.

Réglons quelques petits détails. (i) Comment préparer des amorces adéquates puisqu'on ne connaît pas la séquence ? En réalité, le fragment à séquencer est sûrement le résultat d'une coupure par un enzyme de restriction. Comme nous savons quel enzyme nous avons utilisé, nous connaissons la séquence des extrémités de notre ADN X. (ii) Comment séparer les deux brins d'ADN et utiliser spécifiquement un seul ? En faite, dans la solution, les deux brins (séparés après portage à haute température) sont présent, mais seulement l'amorce complémentaire à l'extrémité 3' de l'un des brins est ajouté, et ce n'est donc que ce brin qui participe à la réaction d'élongation. Nous mettons ici encore à profit le besoin de l'ADN-Pol d'une amorce pour pouvoir fonctionner. D'ailleurs, pour détecter les erreurs possibles, chacun des deux brins est séquencé séparément pour vérifier que les deux séquences obtenues sont bien complémentaires.

Notons enfin que nous avons présenté la dernière version de la technique de séquençage. Il y a encore quelques année, les ddNTP n'était pas marqué par fluorescence et il fallait effectuer quatre réactions d'élongation dans quatre tubes différents (chacun contenant un seul des quatre ddNTP) et rouler ensuite quatre gel en parallèle.

5.6 Synthèse d'ADN

Souvent nous avons besoin de court (10-100 bases) fragment d'ADN simple brin de séquence bien déterminée. On peut synthétiser ces *oligonucleotides* chimiquement. L'astuce est d'utiliser des NTP dont le carbone 3' est protégé par un groupe chimique *amovible*. Quand ce "chapeau" est présent, la polymérisation d'ADN n'est pas possible.

On démarre alors avec de courts amorces d'ADN simple brin immobilisées sur des billes. On dispose de quatre réservoirs contenant chacun un des quatre NTP protégés (qu'on notera ATP*,CTP*,...). Supposons que l'on veut synthétiser la séquence 5'-AGTCG... On ouvre d'abord le robinet du réservoir contenant les ATP*, et un A* s'inclut à la suite de l'amorce. A cause du groupe protecteur, l'élongation est arrêtée à ce niveau. On lave alors la solution en enlevant les ATP* restant dans la solution. Nous avons donc pour l'instant un 5'-A*. L'étape suivante consiste à laver la solution avec un produit qui enlève le groupe protecteur. Nous avons alors un 5'-A. Nous ouvrons alors le robinet des GTP*. Comme le A n'est plus protégé, la synthèse avance d'un pas et nous obtenons 5'-AG*. Et on procède comme précédemment et effectue la synthèse pas à pas.

Comme pour le séquençage, la tâche est entièrement robotisée et il existe des dizaines de compagnies à qui on envoie par web ou courrier électronique la séquence désirée et qui nous renvoie un aliquote contenant l'oligo dans les jours suivants. Le prix de nos jours est de l'ordre de 0.5 Euros par base pour des concentrations nanomolaire.

6 Détour : ordinateur à base d'ADN ?

Nous venons de voir dans les chapitres précédents que l'ADN contient de l'information sous forme d'une chaîne linéaire de *bit* à quatre états : A,C,G,T. La vie manipule constamment cette information pour se perpétuer et nous même[1] avons détourné certains des outils du vivant à notre propre profit. Serait il possible d'aller plus loin et pousser le détournement jusqu'à effectuer des calculs mathématiques complexes à l'aide de l'ADN ? Nous verrons plus loin, au cours des chapitres sur le contrôle de la transcription, que la vie n'est qu'une grande machine de Turing (*i.e.* l'idéalisation mathématique de l'ordinateur) mais qu'elle a essentiellement implémenté des algorithmes qui servent à dupliquer l'ADN. Mais peut-on programmer l'ADN pour décomposer un grand chiffre en ses facteurs premiers ou de résoudre un problème de la théorie des graphes ?

La question n'est pas simplement d'intérêt académique. Certains calculs (notamment ceux cités ci-dessus) sont très coûteux en terme de nombres d'opérations et la seule voie de les attaquer est de les paralléliser : utiliser un grand nombre d'unité pour effectuer chacune un bout de calcul. Or, si on pouvait programmer l'ADN, nous aurons à notre disposition, dans un micro-mol, 10^{17} unités ! Les super-ordinateurs parallèles à base de silicium que nous utilisons de nos jours, utilisent au plus 10^4processeurs et coûtent une fortune.

Nous avons fait toute cette introduction publicitaire pour pouvoir répondre par oui : on peut programmer l'ADN pour résoudre par exemple un problème de la théorie des graphes, et nous en donnons un exemple ci-dessous. L'idée et l'expérience fondamentale sont dues à Adleman dans un article publié dans la revue *science* en 1995.

6.1 L'expérience d'Adleman.

Le problème résolu est celui d'un graphe hamiltonien : soit donné N noeuds liés par des liens *directionnels* (ce qui veut dire que $i \to j$ et $j \to i$ ne sont pas équivalent). Existe-t-il un chemin menant du noeud 1 au noeud N en passant par les autres noeuds *une et seulement une* fois ? La figure (6.1a) est un exemple de tels graphes. Le problème posé est classé dans la catégorie *NP* au niveau de son coût en terme de nombre d'opérations. Cela veut dire que si le nombre de noeud est N, nous ne connaissons pas d'algorithme capable de résoudre ce problème en N ou N^2ou n'importe quelle autre puissance de N opérations. Par contre, nous savons vérifier en un temps polynomial si un chemin quelconque est solution ou non[2]. Ce

[1] les humains. Il y a certes une part d'arrogance de nous distinguer de la *nature* en général. Mais ceci est un autre débat philosophique.

[2] La société mathématique internationale a lancé un défi : celui qui résoudra un problème de ce genre en un temps polynomial (ou qui démontrera que cela n'est pas possible) empochera 10^6dollars et probablement une médaille fields.

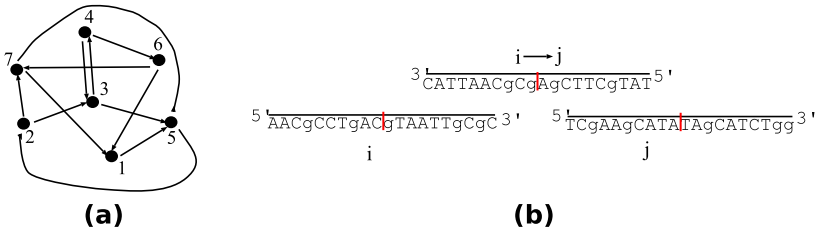

FIG. 6.1: (a) : Existe-t-il, dans ce graphe, un chemin menant de 1 à 7 en passant une et une seule fois par tous les noeuds ? La réponse est ici oui : $1 \to 5 \to 2 \to 3 \to 4 \to 6 \to 7$. (b) : Dans l'expérience d'Adleman, chaque noeud est représenté par une séquence unique de 20-mere d'ADN. Si il existe un lien du noeud i vers le noeud j, alors ce lien est également représenté par une séquence de 20-mere. La séquence du lien est choisie pour être pour moitié complémentaire à i et pour moitié à j. Notez que comme l'ADN est un polymère avec une direction (un coté 5' et un coté 3'), les liens $i \to j$ et $j \to i$ sont représentés par deux séquences distinctes.

problème peut en principe être résolu par l'algorithme suivant (qui n'est pas le plus efficace si on programmait un ordinateur classique) :

1. Générer *tous* les chemins, étant donnés les noeuds et les liens.
2. Sélectionner seulement les chemins qui commencent par 1 et finissent par 7 (dans le cas de la figure 6.1a)
3. Sélectionner ceux qui ont une longueur exactement de 6 (7-1).
4. Sélectionner ceux qui sont passer par chaque noeud au moins une fois.
5. Si à la fin de ce processus de sélection, il reste encore des chemins, la réponse est oui.

Pour implémenter cet algorithme, Adleman a utilisé des séquences de 20 bases (des 20-meres) d'ADN simple brin. Chaque noeud est représenté par une séquence unique. Si un lien de i à j existe, alors un autre 20-mere est ajouté à la sauce, qui est complémentaire pour moitié à i et pour moitié à j. L'astuce de l'expérience est d'utiliser le fait que l'ADN est un polymère dirigé, avec un coté 5' et un coté 3'. En se référant à la figure (6.1b), on se rend vite compte que le lien $i \to j$ et $j \to i$ ont des séquences différentes. La dernière astuce d'Adleman était pour les liens qui impliquent le premiers ou le derniers noeud. Le lien $1 \to j$, si il existe, est représenté par un 30-meres, dont les 20 bases coté 3' sont complémentaire à la totalité des vingts bases de 1 ; les dix autres restantes sont complémentaires à la moitié de j, comme précédemment. De même, si le lien $i \to 7$ existe, il est représenté par un 30-meres dont les 20 bases (cotés 5') sont complémentaire à la totalité de 7. Bon, on devient un peu confus là, mais il suffit pour le lecteur de dessiner quelques noeuds et liens selon les principes ci-dessus pour se rendre compte du pourquoi et comment. Passons maintenant à l'implémentation des étapes 1 à 5.

Etape 1. C'est l'étape la plus facile, et celle qui fait pratiquement toute la puissance de la méthode. On mélange tous les noeuds et tous les liens ; on laisse gentiment les morceaux complémentaires s'hybrider ; on ajoute des ligases dans le milieu pour souder les liens pendants.

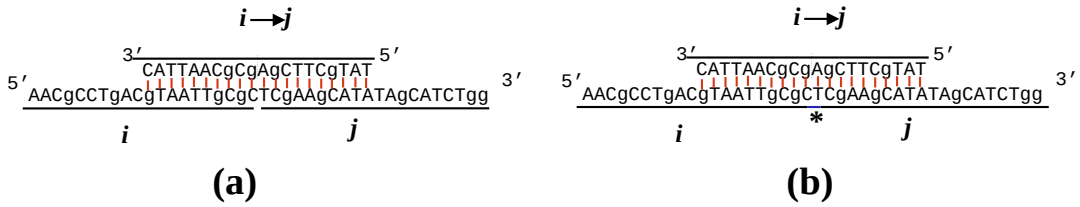

FIG. 6.2: (a) étape d'hybridation : le lien s'hybride aux deux morceaux complémentaires des noeuds *i* et *j*, en rapprochant le coté 3' de l'un au coté 5' de l'autre ; (b) la ligase soude la liaison covalente entre la dernière base de *i* et la première base de *j*. La position de la soudure est marquée par une *.

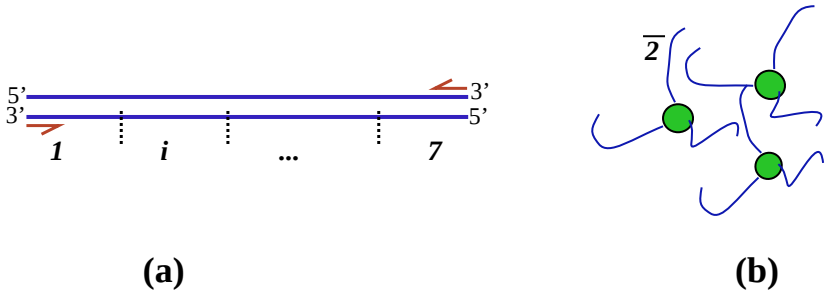

FIG. 6.3: (a) Choix des amorces de PCR pour la sélection des chemins commençant par 1 et finissant par 7 ; (b) des séquences complémentaires au noeud 2 sont greffées sur des billes : leur hybridation avec des "2" permet de sélectionner les chemins qui contiennent ces derniers.

La figure 6.2 illustre ce processus pour deux noeuds et leur liens. A la fin de cette étape, les chemins possible sont représentés par des polymère *double brins* d'ADN dans la solution. Notez également que seul les chemins qui commencent par 1 et se terminent par 7 sont entièrement double brins. Les autres auront toujours, d'un coté ou de l'autre, une séquence de 10 bases simple brin.

Etape 2. Pour sélectionner les chemins qui commencent par 1 et finissent par 7, on utilise le PCR, en donnant des amorces complémentaire au début de 1 et de $\overline{7}$(figure 6.3a). Comme nous l'avons vu au chapitre sur le PCR, seul ces chemins seront amplifiés exponentiellement.

Etape 3. La sélection des chemins qui ont exactement 6 liens, et donc 120 bases, se fait par un simple électrophorèse : nous avons vu au chapitre précédent que cet outil permet essentiellement de distinguer les longueurs. La sélection des chemins à 120 base peut être suivi d'un PCR pour une amplification supplémentaire.

Etape 4. C'est l'étape le plus fastidieux. Pour sélectionner, dans tous les chemins qui nous restent, ceux qui contiennent le noeud 2, on greffent des simples brins $\overline{2}$ sur des billes micrométriques (figure 6.3b). On ajoute l'ensemble des chemins dont on dispose ; on chauffe pour séparer les doubles brins ; on laisse refroidir pour permettre aux brins complémentaires de s'hybrider à nouveau. Certains des chemins qui contiennent le noeud 2 vont alors s'hybrider avec les brins $\overline{2}$sur les billes. On lave ensuite la solution pour enlever tous les ADN qui ne se sont pas hybridés. Finalement, on chauffe à nouveau la solution pour séparer les doubles brins et on récolte l'ADNss en solution. Ces ADN contiennent sûrement la séquence du noeud 2. On procède à un PCR pour amplifier le signal. Et on continue de la même manière pour tester la présence des noeuds 3,4,... dans les chemins restants.

Etape 5. Finalement, à la fin de tous ces processus, il suffit de rouler un gel pour voir si il reste de l'ADN dans la solution. Si oui, la réponse à la question posé est "oui, il existe des chemins allant de 1 à 7 en passant une et une seule fois par chaque noeud".

6.2 Autour du "DNA-computing".

L'expérience d'Adleman était une démonstration de principe : oui on peut programmer l'ADN pour effectuer des calculs mathématiques parallèles ! Sa force est dans sa capacité à générer en un seul coup $\sim 10^{15}$ chemins lors de l'étape 1 (cela dépend bien sûr de nombre de mol d'ADN qu'on a mis dans la solution) et de trier ce très grand nombre également en un seul coup lors des étapes 2 et 3. Le coût opérationnel de ces étapes est constante et ne dépend pas du nombre de noeuds N. L'étape 4, bien que long en temps de manip, est d'un coût N. Le coût opérationnel des algorithmes classiques connus par contre est de l'ordre de $N!$. Pour fixer les idées, $15! \approx 10^{13}$. Le DNA computing paraît donc très très rentable.

L'article d'Adleman a provoqué, après sa parution en 1994, un foisonnement de travaux théoriques proposant des algorithmes les plus divers pour résoudre des problèmes. Malheureusement, personne n'a jamais réussi à faire de l'arithmétique avec l'ADN[3], et les problèmes résolus par l'ADN sont resté cantonnés aux rayons des problèmes de la théorie des graphes. Or, ces procédés perdent beaucoup de leurs intérêt si on ne sait pas manipuler des chiffres.

Un autre défaut du calcul avec l'ADN est son "humidité" ! Ce sont des manipulations physiques qui peuvent accumuler beaucoup d'artefacts, et quand ils sont implantés, ils perdent beaucoup de leurs charmes. Adleman lui même n'a pas été très rigoureux dans son travail : pas de test par exemple pour vérifier que si il n'y a pas de chemin, ses manipulations donnent effectivement une réponse négative. D'autres groupes ont perdu quelques années dans ce genre d'artefacts.

Finalement, l'arrivée des concepts de l'informatique quantique a montré l'existence des voies autrement plus puissantes pour effectuer des calculs parallèles. L'article de principe de Shor a montré comment factoriser un nombre en ses facteurs premiers à l'aide des ordinateurs quantiques : cela a montré l'aisance de ces ordinateurs à manipuler des chiffres, et s'est attaquer de front à un problème très actuel, qui est à la base de la cryptographie à clef publique.

[3]Les algorithmes pour faire une addition sont restés à l'état embryonnaire. Quant à réalisé une multiplication ...

L'ADN-informatique est donc passée de mode. Mais nous avons trouvé utile de rappeler cet épisode pour insister sur l'information contenue et manipulable dans l'ADN. Nous verrons aux chapitres sur le contrôle de la transcription comment la nature utilise effectivement cette information pour réaliser des machines de Turing extraordinaires qu'on appelle plus communément des organismes vivants.

7 Cinétique enzymatique et correction d’erreur.

7.1 Le B.A. BA des réactions enzymatiques

7.2 Théorie d’hopfield de correction d’erreur

7.3 Régulation enzymatique.

8 Contrôle de la transcription.

Il ne suffit pas d'avoir à sa disposition des programmes informatiques, encore faut-il les exécuter dans le bon ordre pour obtenir un résultat cohérent. La même chose est valable pour les gènes : on peut avoir 30000 gènes, mais une cellule donnée ne transcrit qu'un sous ensemble réduit de ces gènes, à des amplitude différentes et à des temps différents. Comment une cellules fait pour choisir quel gène exprimer ? Cela s'appelle le contrôle de la transcription, c'est le coeur du programme de traitement d'information de la vie, et c'est ce qu'on va voir dans ce chapitre.

Il est intéressant de comparer le programme du vivant au mode de fonctionnement d'un ordinateur : les deux méthodes de traitement d'informations sont fondamentalement différentes.

L'ensemble des programmes qu'un ordinateur peut exécuter est contenu sur son disque dur[1]. Le disque dur est une longue suite d'octet. Certains de ces octets forment des programmes, et un utilisateur peut regarder directement la séquence codante en assembleur. Le disque dur contient une zone très particulière, une table de correspondance, qui contient l'adresse de *tous* les programmes - c'est un peu plus compliqué, mais cela revient au même. Supposez que vous demandez au système d'exploitation (OS, pour operating system) d'exécuter un programme particulier, "openoffice" par exemple. L'OS se réfère alors à cette table, trouve l'adresse *et* la longueur *n* de ce programme sur le disque dur, se rend à l'adresse donnée (l'octet numéro 625467834 par exemple) et charge *n* octet en mémoire vive. Ces *n* octets son ensuite envoyés vers le processeur pour être traités. L'OS lui même n'est bien sûr rien d'autre qu'un programme, une suite d'octets, à ceci prêt que c'est le premier programme qui se charge en mémoire, et cela de façon automatique[2]. Revenons au cas de notre programme "openoffice". Il est très probable qu'au cours de son exécution, il nécessite le chargement d'autres programmes auxiliaires, comme par exemple "afficher une fenêtre de dessin". La demande est transmis au processeur, qui le transmet à l'OS. A nouveau l'OS se réfère à la table de correspondance, trouve l'adresse de ce nouveau programme, s'y rend et la charge en mémoire, et ainsi de suite.

Pour le monde du vivant, la transcription, la copie d'un morceau de l'ADN en ARN, est l'équivalent du chargement en mémoire, et la traduction par les ribosomes est l'équivalent de l'exécution. Mais l'analogie s'arrête là : Il n'existe pas de table de correspondance qui contient l'adresse des gènes sur le chromosomes ; il n'existe pas *un* processeur central mais des milliers, ce sont les ARN polymerases et les ribosomes ; il n'y a pas de système d'exploitation pour gérer le choix et l'ordre d'exécution des programmes, ce sont les programmes eux même qui "décident" si ils doivent être exécuté. En langage chic, c'est un modèle de traitement

[1] A l'exemple du chromosome eukaryote, le disque dur, surtout si il est géré par MS Windows, contient également beaucoup d'informations non-codantes : des fossiles de virus, des résultats de fragmentation, ...

[2] A moins bien sûr que l'OS soit le MSWindows, auquel cas il faut parfois l'aider en appuyant plusieurs fois sur les touches Ctrl, Alt et Del en même temps.

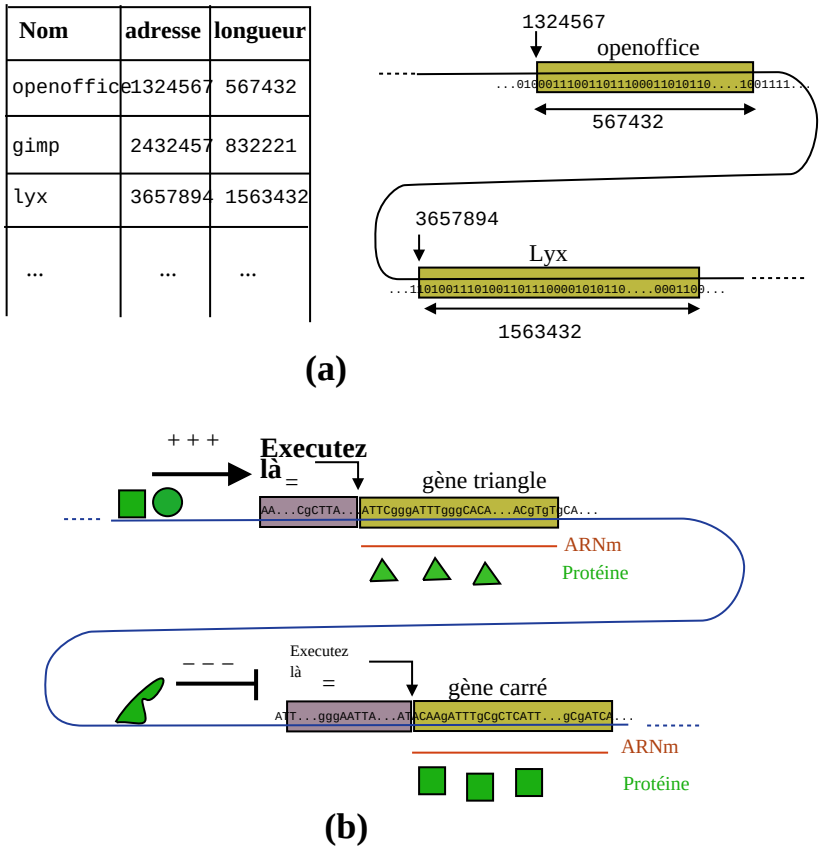

FIG. 8.1: Comparaison entre un ordinateur (a) et le vivant (b) pour le traitement de l'information. Dans les deux cas, l'information est stockée sous forme d'une chaîne linéaire. Pour un disque dur, la chaîne est formée de petits domaines magnétiques ; pour le vivant, d'un polymère d'ADN. (a) : Dans un ordinateur, l'adresse et la longueur des programmes sont contenues dans une table centrale. Pour exécuter un programme, le système d'exploitation se réfère à cette table. (b) Dans le vivant, il n'existe pas de table centrale et les gènes s'auto-régulent. Un gène est copié d'abord par ARN-polymerase en ARN messager, et cet ARNm est ensuite lu par un rybosome pour fabriquer des protéines. En amont d'un gène, il existe une séquence appelé promoteur(grisée dans la figure) qui attire et positionne l'ARN-pol, en lui indiquant l'endroit exacte où il doit commencer la transcription. L'attraction de cette séquence pour ARN-pol peut-être plus ou moins forte. D'autres protéines peuvent se fixer en amont du promoteur d'un gène et moduler son attraction. Dans le schéma ci-dessus, la protéine "carré" produit par le gène "*carré*" se fixe en amont du promoteur du gène "*triangle*" et amplifie l'attractivité du promoteur. On dit alors que "*carré*" active "*triangle*". La protéine "patatoïde" se fixe, elle, en amont du gène carré et diminue l'attractivité du promoteur de ce dernier. On dit alors que "*patatoïde*" inhibe "*carré*".

d'information distribué *et* parallèle, le Graal des informaticiens humains. Nous verrons le détail de la méthode plus bas, mais essayons d'en avoir une idée générale (fig. 8.1) :

1. l'adresse exacte d'un gène sur le chromosome n'a aucune importance, vu qu'il n'existe pas de table de correspondance. Par contre, chaque gène (une suite d'ATCG) contient juste avant sa séquence codante une séquence particulière dont la signification, en langage du vivant est : il faut transcrire à partir de la. Cette séquence est appelé le *promoteur*. C'est cette séquence qui attire l'ARN Polymerase et le positionne au bon endroit.
2. La longueur du gène n'est pas connu à l'avance. Une séquence particulière *à l'intérieur même* du gène signale à l'ARN-Pol qui est en train de transcrire qu'il faut s'arrêter.
3. Les promoteurs peuvent être plus ou moins fort. Par ce terme, on entend qu'il peuvent avoir plus ou moins d'affinité avec l'ARN-Pol : plus ils sont fort, plus ils attirent l'ARN-pol, et plus le gène qui se trouve en leurs aval va être transcrit, et plus la protéine que le gène code va être produit.
4. Attention, c'est le point crucial : la force des promoteurs n'est pas une donnée fixe, elle peut être modulée (en plus ou en moins) par d'autres protéines, produits par d'autres gènes ! Ces autres protéines se fixent sur le chromosome en amont (souvent, mais pas toujours) du promoteur. Les protéines qui régulent la transcription des autres gènes sont appelé des facteurs de transcriptions.

8.1 Le guidage primaire : les promoteurs.

Parler des σ-factor, des promoteurs, des séquences -10 et -35, de la recherche des promoteurs pour trouver les gènes...

8.2 Les contrôles actifs : les facteurs de transcription.

8.2.1 L'Operon lac.

L'operon Lac est l'exemple le plus fameux de la régulation : c'est le premier qui a été compris et qui a valut à ces découvreurs, Jacob et Monod, le prix Nobel en 1962, et c'est celui qui a servi de modèle à tous les autres.

Résumons : la bactérie *E.Coli* aime surtout le sucre glucose. Une population d'*E.Coli* croît exponentiellement dans un milieu riche en glucose. Si maintenant on enlève le glucose et qu'on le remplace par un autre sucre, le lactose, la croissance exponentielle s'arrête, puis reprend au bout d'environ 30 minutes. La bactérie produit maintenant un enzyme, β-galactosidase (β-gal pour les intimes) qui digère le sucre galactose[3]. Donc, en réponse à l'environnement, la bactérie a été capable d'activer la transcription d'un gène, "allumer" le gène comme on dit, et cela lui a pris environ 30 minutes (Fig. 8.2). Autre beauté de la régulation : il est plus économique de digérer du glucose que du galactose. En présence des deux sucres, β-gal est produit, mais en faible quantité (Tab. 8.1). Le mécanisme de la régulation de l'operon

[3] La présence et la quantité de cet enzyme peut très facilement être détecté par une coloration chimique.

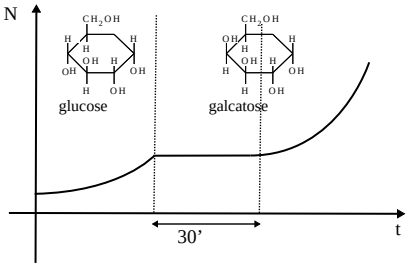

FIG. 8.2: La croissance bactérienne reprend au bout d'environ 30' quand le sucre glucose est remplacé la galactose. C'est le temps qu'il faut pour activer l'operon Lac et produire l'enzyme β-gal.

glucose	galactose	production β-gal
+	-	0
+	+	1
-	+	40

TAB. 8.1: La production de l'enzyme β-gal en fonction de la présence ou non de glucose et galactose.

lac a été décortiqué et continue a l'être de nos jours (nous n'avons pas encore tout compris). La voici dans ses grandes lignes.

Un operon est un "train" de gènes, un ensemble de gènes disposé bout à bout qui sont transcrit en même temps pour produire un ARNm géant. Le ribosome, lors de la traduction, lira cet ARNm géant et au fur et à mesure qu'il rencontrera les codons stop, relâchera des protéines finies.Dans le cas du Lac operon, il existe une protéine, le répresseur du Lac (rep pour les intimes) qui a une très grande affinité pour le promoteur du Lac. Son adhésion au promoteur empêche physiquement le facteur σ de s'ancrer et de faire son boulot.

Voilà le point crucial de la régulation : rep a encore plus d'affinité pour le sucre galactose et la présence de ce dernier *capture* le rep disponible et l'empêche de se lier au promoteur Lac. Le promoteur étant libre, il peut maintenant ancrer le σ-facteur et transcrire l'operon Lac. Nous avons donc un double contrôle *négatif* qui donne un résultat positif. Mais l'histoire n'est pas fini : le promoteur de Lac est un promoteur faible et possède une faible attraction pour le facteur σ : même libre, la transcription du lac reste faible. Par contre, une autre protéine, CAP, a une forte attraction pour une région en amont du promoteur *et* une forte attraction pour l'ARN pol. Sa présence sur l'ADN augment donc fortement l'ancrage de l'ARN-pol et donc le taux de transcription. CAP exerce donc un contrôle *positif* sur la transcription du Lac. Par contre, la présence du glucose inhibe son action[4]. Nous avons donc maintenant une explication pour le tableau 8.1. L'ensemble de ces processus est schématisé sur la figure 8.3.

Malgré son aspect simple, la régulation du Lac est d'une très grande finesse. Par exemple, le LacY code pour une protéine, Lac permease, qui est responsable de l'importation du Lac

[4]L'interaction entre CAP et glucose passe par une petite molécule d'AMP cyclique. Nous n'entrons pas plus dans le détail.

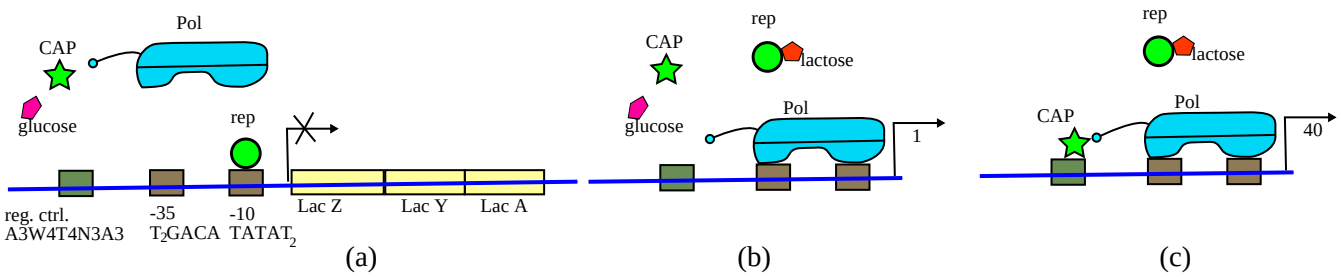

FIG. 8.3: Schéma de régulation de l'operon Lac. (a) : l'operon lac formé des gènes LacZ (fabriquant l'enzyme β-gal), LacY(lactose permease) et LacA avec les séquences compromis de son promoteur et de la région régulatrice (W signifie A ou T et N signifie n'importe quelle nucleotide). En l'absence du sucre galactose, la protéine rep adhère au promoteur et empêche physiquement le complexe σ-facteur, RNA-pol de s'ancrer. Il n'y a pas de transcription du gène. (b) En présence du lactose, rep est capturé et le promoteur devient libre : ARN-pol peut s'y accrocher faiblement et commencer la transcription. (c) en l'absence du glucose, la protéine CAP est libérée et peut s'ancrer sur la région régulatrice du gène, favorisant fortement l'ancrage de l'ARN-pol et multipliant la transcription par 40 environ.

dans la bactérie. Si le lactose est présent dans le cytoplasme, la perméase est construite et importe encore plus de Lac dans le cytoplasme, ce qui capture encore plus de rep et amplifie la transcription du β-gal. Nous référons le lecteur sur un cours plus avancé de la théorie de régulation pour toute les finesse de cet operon.

8.2.2 Le phage λ

8.2.3 Transcription chez les eukaryotes

9 Autour de la transcription.

9.1 Aperçu général des circuits génétiques

Nous avons vu qu'il existe des protéines qui régulent l'activité d'autres gènes. On les appelle des *facteur de transcriptions*. Un facteur de transcription A peut activer l'expression d'un gène B ; on note cela par $A \rightarrow B$ (comme dans le cas du CAP pour Lac). Elle peut également inhiber, réprimer l'activation du gène ; On note alors cela par $A \dashv B$ (rep inhibe Lac). Bien sûr, les gènes des facteurs de transcription sont elles mêmes régulées par d'autre facteurs de transcription. Ces gènes forment alors des *réseaux* d'interaction auto-organisé que l'on appelle également des circuits, à cause de l'analogie profonde qu'ils ont avec les circuits électroniques. Par exemple, $A \dashv B$, $B \dashv C$, $C \dashv A$ est un réseau formé de trois noeuds qui s'inhibent les uns les autres . La figure 9.1 montre quelques exemples simples et la figure 10.1(a) montre une partie d'un vrai réseau d'interaction extrêmement important pour le développement embryonnaire de la mouche drosophile. Les réseaux d'interaction sont toujours montrés sous forme de graphe.

Le schéma électronique d'un processeur (un Pentium IV par exemple) nous montre une organisation d'une effroyable complexité, et on peut même se demander comment on peut dessiner une telle chose et qui fonctionne en plus ! En faite, le processeur est fait de *modules* assez compréhensible : tel groupe de transistors et ampli-op se charge de l'horloge, tel autre groupe organise les registres, un autre se charge des communication avec la mémoire et ainsi de suite. Un électronicien apprend d'abord le fonctionnement de petits modules, et c'est ce qu'on verra à la prochaine section. Mais avant même l'apprentissage des modules, il doit comprendre le fonctionnement des diodes, transistors et autres ampli-op. C'est ce que nous allons faire ici.

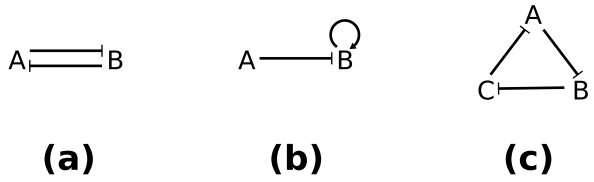

FIG. 9.1: Quelques exemples très simple de schéma d'activation. Nous verrons plus bas que (a) correspond à un bascule bistable comme dans la phageλ, (b) correspond à une mémoire et que (c) forme un oscillateur.

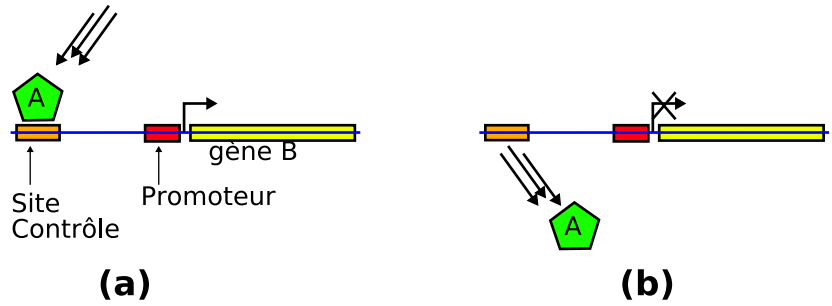

FIG. 9.2: Dans la schéma $A \rightarrow B$, l'ADN polymérase ne transcrit B que quand son site de contrôle est occupé par A (comme dans (a)). Le taux d'occupation de ce site est un compromis entre le flux d'arrivée de A, α et son taux de départ, β.

9.1.1 Activation et inhibition.

Quand $A \rightarrow B$, il va de soi que plus la concentration de la protéine A est élevée, plus le gène B est transcrit et donc plus on produira la protéine B[1]. Comme on l'a dit, le gène B n'est transcrit que quand A est présent sur sa région de contrôle. Le taux de production doit donc être proportionnel au temps de présence de A sur cette région, ou autrement dit, à la probabilité $P_+(t)$ d'occupation de ce site par A au temps t (Fig. 9.2). Soit α le flux d'arrivée de A sur le site et β le taux auquel une protéine A déjà présente sur le site le quitte. $P_+(t)$ est la probabilité pour que le site soit occupé par un A, et $P_-(t) = 1 - P_+(t)$ la probabilité pour que le site soit vide à l'instant t.

Bon, faisons un peu de calcul élémentaire. La probabilité pour que le site soit occupé à l'instant $t + dt$ est que (i) soit le site est occupé à l'instant t et qu'il le reste dans l'intervalle dt, (ii) soit qu'il était vide au temps t et qu'une molécule A arrive dessus. En langage plus formel,

$$P_+(t+dt) = (1-\beta dt)P_+(t) + \alpha dt P_-(t)$$

ou, en réarrangeant les termes,

$$\frac{dP_+}{dt} = -\beta P_+ + \alpha P_- = \alpha - (\alpha+\beta)P_+$$

L'état stationnaire $dP_+/dt = 0$ est donné par

$$P_+ = \frac{\alpha}{\alpha+\beta}$$

Le taux α doit dépendre (entre autres) de la concentration des A : plus il y a des A dans la solution, plus le flux d'arrivée est importante. Donc, $\alpha \sim A$ [2]. Par contre, β ne dépend essentiellement que de l'énergie d'interaction de A avec le site de contrôle et ne dépend à

[1] Une convention assez répandu veut que l'on note un gène en italique, et la protéine qu'il code en roman non italisé.

[2] En toute rigueur, nous devrions noter [A] la concentration de A. Mais tant que cela ne prête pas à confusion, nous omettrons les crochets.

priori pas de la concentration des A. Comme le taux de transcription de B, v_B est proportionnel elle à P_+, nous avons finalement la relation entre la production de B et la concentration de A :

$$v_B \sim \frac{A}{C_A + A} \tag{9.1}$$

où C_A est une constante dépendante des divers facteurs de proportionnalités que nous avons introduis. Ce que nous apprend cette équation est qu'à basse concentration ($A \ll C_A$), la production de B augmente linéairement avec A, tandis qu'à forte concentration ($A \gg C_A$) la production de B sature à une valeur donnée.

Bon, nous avons maintenant l'expression du taux de production de B. Quelle est maintenant sa concentration ? C'est évidement un compromis entre le taux de production et le taux de *destruction* de B. Nous n'avions jusque là pas trop parlé de la destruction des protéines. Si les protéines était éternelles, elles s'accumuleraient vite dans la cellule. De plus, la cellule n'aurait aucun moyen de contrôler le niveau des protéines. Pour conduire une voiture, il faut avoir un accélérateur *et* un frein. Plus exactement dans le cas du monde vivant, le frein est toujours actif, et on contrôle la vitesse en appuyant plus ou moins sur le champignon. Tout les protéines sont marqué par un système complexe pour être présentées à un moment de leur vie à des *protéases* et être dégradées[3]. Les acides aminés de cette dégradation sont ensuite recyclés pour la synthèse d'autres protéines. Il y a tout un pan de la régulation basé sur la destruction que nous ne traiterons pas ici. Nous supposerons simplement que chaque protéines à un temps de vie donné et noterons μ son taux de dégradation.

Revenons maintenant à notre question initiale sur la concentration de B. D'après tout ce que l'on dit, on doit avoir $dB/dt = v_B - \mu_B B$, et donc à l'état stationnaire,

$$B = \frac{K_B}{\mu_B} \frac{A}{C_A + A} \tag{9.2}$$

où K_B est une constante dénotant l'efficacité de la transcription intrinsèque de B, tenant compte de divers facteurs comme l'énergie d'interaction du promoteur avec le facteur σ. Donc, le résultat est finalement assez simple : à faible concentration de A, la concentration de B lui est proportionnelle, tandis qu'elle sature à $B_\infty = K_B/\mu_B$ à forte concentration de A.

Tout ce que nous avons dit jusque là se transcrit littéralement pour le cas de répression $A \dashv B$, sauf que cette fois, le taux de transcription est proportionnel à $P_-(t)$, la probabilité pour que le site de contrôle *ne soit pas* occupé. On obtient alors $v_B \sim 1/C_A + A$ et $B = (K_B/\mu_B)(1/C_A + A)$: à forte concentration de A, la concentration de B tombe à zéro.

Nous avons passé sous silence jusque là une approximation importante que nous ne pouvons plus taire : l'équation (9.1) dénote en toute rigueur la production d'ARNm, tandis que l'équation (9.2) donne le niveau du protéine. Nous avons négligé le fait qu'entre la transcription de l'ADN et la production de protéine, il y a la translation. Notre approximation consiste donc à supposer que la production de protéine est proportionnelle au nombre d'ARNm transcrit, et c'est un des facteurs que nous avons intégré dans le coefficient K_B. Cette approximation n'est pas toujours justifiée, mais nous l'admettrons tout au long de ce cours.

[3] Nous avons un peu exagéré le trait. Certaines protéines sont censées être présent dans la cellule à une concentration plus ou moins constante tout au long de la vie de la cellule et sont donc très stable. Mais nous référons le lecteur aux livres spécialisés pour plus de détail.

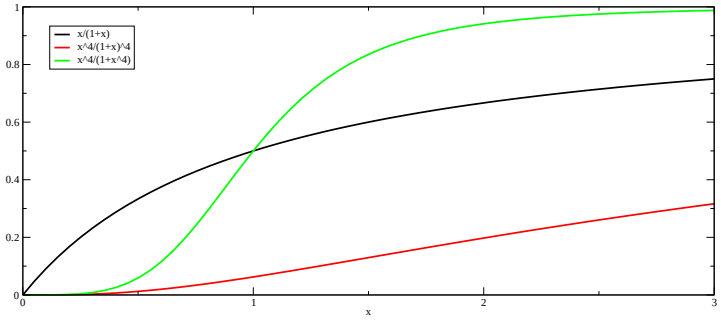

FIG. 9.3: La différence entre l'activation simple, multiple (4) et multiple et coopérative.

9.1.2 La coopérativité.

Un gène à souvent plusieurs site de contrôle où plusieurs facteurs de transcription différents doivent s'accrocher pour activer la transcription. Par exemple, le gène B peut avoir la règle suivante : si les sites de A1 et A2 sont activé et le site de A3 est vide, alors active la transcription de B. Il n'est pas difficile de suivre le raisonnement de la précédente section pour dériver l'expression du taux de production en fonction de la règle annoncer. Dans l'exemple ci-dessus, le lecteur peut démontrer que le taux de production est $v_B \sim (A_1/C_{A_1} + A_1)(A_2/C_{A_2} + A_2)(1/(C_{A_3} + A_3)$. Nous laissons au lecteur comme exercice de dériver l'expression pour le "ou" logique : active B si A1 *ou* A2 occupent leur site de contrôle.

Un cas très intéressant et très répandu est celui ou le gène possède plusieurs site de contrôle pour la même protéine A et il faut que tous ces sites soit occupés pour que l'activation s'accomplisse. D'après ce que nous venons de voir, on doit avoir

$$v_B \sim \frac{A^n}{(C_A + A)^n} \tag{9.3}$$

Cependant, l'adhésion des A aux sites de contrôle peut être *coopérative*, c'est à dire que la probabilité pour qu'un A adhère a site de contrôle est (fortement) augmentée si un site voisin est déjà occupé par un A ! Dans ce cas, l'expression du taux de production change radicalement :

$$v_B \sim \frac{A^n}{C_A^n + A^n} \tag{9.4}$$

Nous laissons au lecteur le soin de dériver cette dernière équation : cela nécessite un peu de diagonalisation de matrice et quelques approximations dues aux mots "probabilité *fortement* augmentée".

Pour voir la différence entre les expression (9.1,9.3,9.4), il suffit de jeter un coup d'oeil sur la figure 9.3 . Comme on peut le constater, dans le cas d'activation coopérative, le comportement est du genre interrupteur ON/OFF : pour $A < C$, la production est pratiquement zéro, tandis que pour $A > C_A$, la production est pratiquement à son niveau saturante. Et cela est d'autant plus vrai que le coefficient de coopérativité n, qu'on appelle également le coefficient de Hill, est élevé. Pour n suffisamment grand, on approxime l'activation par une fonction échelon.

L'avantage pour la cellule d'utiliser la coopérativité est le même qui a poussé les humains de passer de l'électronique analogique à l'électronique digital : modularité et élimination du bruit. Dans le cas des régulation avec un fort coefficient de Hill, le circuit devient quasiment booléen. Le Niveau précis d'un gène n'a plus tellement d'importance, seul compte le fait qu'il soit "haut" ou "bas". Il devient alors possible de connecter plusieurs modules, en utilisant l'output de l'un comme le input de l'autre, sans trop s'arracher les cheveux et designer tout le circuit depuis le début.

De même, Il n'est pas possible pour la cellule de maintenir strictement le niveau de A à une valeur donnée. Mais cela a peu d'importance pour la production de B si les fluctuations de A ne font pas passer sa valeur d'un côté ou de l'autre du seuil d'activation.

Nous ne rentrons pas plus dans les détails et réferrons le lecteur à un cours plus avancé sur la théorie de régulationréférons[4]. Notons simplement que les activations coopératives sont très répandu : par exemple l'operon Lac que nous avons traité plus haut.

9.2 quelques circuits simples : mémoire, oscillateur, bascule.

Voyons maintenant le fonctionnement de quelques circuits montrés sur la figure 9.1. Ces circuits sont très répandu dans la nature. Nous allons rentrer un peu dans les détails mathématiques.

9.2.1 Un bistable.

La figure 9.1.a montre le schéma de deux gènes qui s'inhibent mutuellement. Il est évident que quand le niveau de A est élevé, celui de B doit être bas et vice et versa. D'après ce que nous dit, la production des deux protéines obéit à deux équations différentielles couplées :

$$\begin{aligned} \frac{dA}{dt} &= K_A \frac{1}{C_B + B} - \mu_A A \\ \frac{dB}{dt} &= K_B \frac{1}{C_A + A} - \mu_B B \end{aligned}$$

Il y a un peu trop de paramètre là, et on peut se débarrasser de certain sans trop de perte de généralité. En effet, en prenant comme variable auxiliaire $A^* = C_A A$, $B^* = C_B B$ et $t^* = (C_A C_B / K_A)$ (en renormalisrenormalisantant les axes A,B et t), on trouve (en laissant tomber les $*$ pour ne pas alourdir la notation)

$$\frac{dA}{dt} = \frac{1}{1+B} - \mu_1 A \tag{9.5}$$

$$\frac{dB}{dt} = K \frac{1}{1+A} - \mu_2 B \tag{9.6}$$

[4]Cela est un des domaines de recherche qui avance très rapidement de nos jours.

On se rend vite compte que cela n'est pas satisfaisant. L'équation ci-dessus n'a *qu'un* point stationnaire (on dit un point de fonctionnement) dans le domaine $A,B>0$. Cela ce voit en cherchant graphiquement les points de croisement des deux courbes

$$\begin{aligned} A &= \frac{1}{\mu_1}\frac{1}{1+B} \\ A &= \frac{K}{\mu_2}\frac{1}{B}-1 \end{aligned}$$

Si on reconsidère maintenant le switch λdu bactériophage, nous nous rendons compte que le gène *c*I n'ont seulement inhibe l'activation de *cro*, mais en plus, il active *sa propre* transcription. Voilà la clef : pour qu'un bistable puisse l'être vraiment, il faut qu'au moins un des gènes ait une action d'autoactivation. Bien sûr, il faut que les autres paramètres cinétiques (les coefficient K et μ) soit judicieusement "tunés".

Une autre possibilité pour réaliser un bistable est la coopérativité. On peut montrer que si dans les équations (9.5,9.6), les termes d'inhibition sont de la forme $1/(1+B^{\alpha})$ et $1/(1+A^{\beta})$, un bistable est réalisable. Des chercheurs ont effectivement implémenté un tel circuit de façon artificielle dans la bactérie *E.Coli*. La grande difficulté était de "designer" les promoteurs pour trouver les conditions cinétiques permettant l'existence du bistable. Nous renvoyons le lecteur interessé vers des documents plus spécialisés dans le traitement des réseau de régulation.

9.2.2 Un oscillateur.

9.2.3 Une mémoire.

La cellule a parfois besoin de n'allumer un gène B que si un gène A commence à être transcrit, mais de garder B allumé une fois que A disparaît. Cela arrive fréquemment lors du développement embryonnaire, quand un facteur de transcription apparaît pendant une intervalle de temps (avant la gastrulation par exemple), mais les gènes qu'il a allumé doivent rester actif (bien après la gastrulation par exemple) pour maintenir la cellule dans un état différencié. En gros, c'est pour cela que les neurones sont des neurones, et des cellules musculaires des cellules musculaires : à un moment du développement, un facteur de transcription a allumer les gènes qui confèrent aux neurones leurs spécifités et a ensuite disparu. Dans l'individu adulte cependant, les neurones gardent leurs caractéristiques en maintenant activé ces gènes spécifiques[5]. Cela ressemble à allumer un papier : pour démarrer la combustion, nous avons besoin d'une allumette allumée, mais le papier continue ensuite à brûler même si on éteint l'allumette.

La figure 9.1.c est le schéma d'une telle mémoire : $A \to B$ et $B \to B$. Si le lecteur jette un coup d'oeil sur le réseau du développement embryonnaire (figure 10.1), il verra que l'interaction entre le gène *bcd* et *hb* est exactement de cette forme. La transcription de B s'écrit alors :

$$\frac{dB}{dt} = f(B) - \mu B + A \tag{9.7}$$

[5] De plus, les cellules filles ce ces neurones ou muscles activent les mêmes gènes que leurs parents : on appelle cela épigénétique. On en dira quelques mots plus loin.

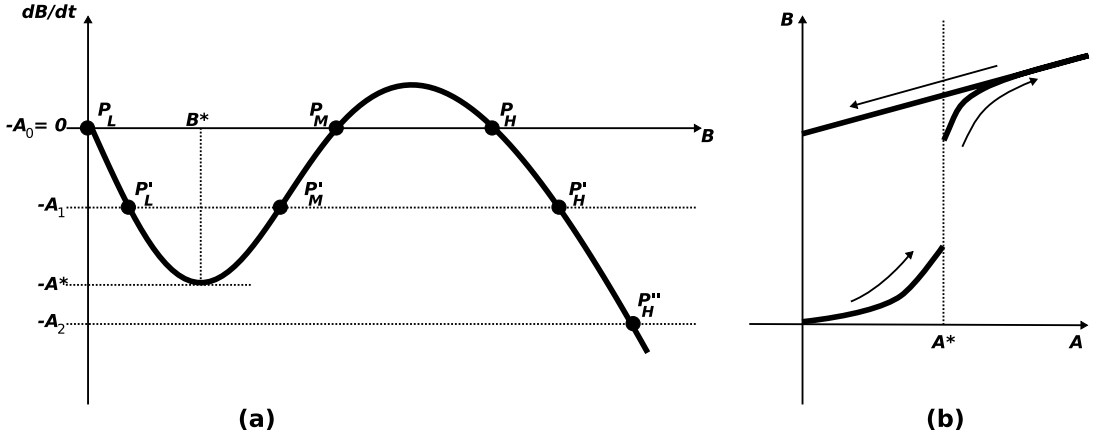

FIG. 9.4: (a)Les points stationnaires ($dB/dt = 0$) de l'équ.(9.7) sont donnée par le croisement des courbes $y = f(B) - \mu B$ et la ligne horizontal $y = -A$. (b) Comment B suit la variation (lente) de A quand celui ci croît et décroît.

où la fonction $f(B)$ contient l'auto-activation de B, le deuxième terme dénote la dégradation de B et le troisième terme l'activation de B par A. En toute rigueur, ce dernier terme aurait du s'écrire $A/(1+A)$. Nous ne perdons pas grand chose dans la généralité de la discussion en ne considérant que la partie linéaire de ce terme. La fonction $f(B)$ qui est l'activation coopérative de B par lui même doit être de la forme $B^n/(1+B^n)$, mais comme nous allons faire une analyse graphique des équations, sa forme exacte ne nous intéresse pas, nous lui demandons simplement de ressembler à un sigmoïde.

Il nous faut d'abord trouver les points stationnaires ($dB/dt = 0$) de l'équation (9.7). Pour cela, nous cherchons le croisement des courbes $y = f(B) - \mu B$ et $y = -A$. Un coup d'oeil sur la figure 9.4.a montre que pour $A < A*$, l'équation possède trois points stationnaires (P_L,P_M et P_H pour low, medium et high). On montrera plus bas que deux de ces points, P_L et P_H sont stable, tandis que le troisième, P_M est instable. Cela veut dire que si nous perturbons le système par une petite quantité δB au voisinage des points stables, le système revient à ces points, tandis que pour le voisinage du point instable, le système s'en écarte : après une petite perturbation, une bille au fond d'un bol y revient, tandis qu'une bille posée en haut d'un dôme s'en écarte. Pratiquement, si l'on résout (numériquement ou analytiquement quand on peut) l'équation (9.7) avec la condition initiale $B(t=0) = B_0$, nous verrons que $B(t) \to P_L(A)$ ou $B(t) \to P_H(A)$ selon que $B_0 < B^*$ou $B_0 > B^*$.

La situation pour $A > A^*$ est différente et nous n'avons alors qu'un seul point stationnaire P_L''(qui est toujours stable).

Supposons maintenant qu'à l'instant $t = 0$, nous n'avons pas de transcription de B (i.e. $B_0 = 0$) et que le facteur de transcription A est absent $A = 0$. La concentration de B reste alors constante à B=0 (le point P_L sur la figure). Augmentons maintenant la concentration de A à A_1. La concentration de B suit en convergeant, au bout d'un certain temps, à P_L', c'est à dire en restant toujours à un niveau *bas*. Le niveau de B reste bas tant que A n'a pas dépassé le niveau critique. Par contre, quand A dépasse $A*$, le point bas n'est plus un point stationnaire,

et $B \to P''_H$, le seul point stationnaire stable qui reste. Maintenant, si A décroît jusqu'à 0, le point B suit la valeur stationnaire haut qui est toujours stable, pour revenir à P_H (figure 9.4b) . Cela est le comportement typique d'un cycle d'hysteresis. Et nous réalisons que ce circuit effectue exactement ce qu'on lui demandé.

Jetons un coup d'oeil sur le calcul de la stabilité. Appelons $\bar{B}$ une valeur stationnaire quelconque de B. Nous supposons que $B(t=0) = \bar{B} + b_0$ où b_0 est une *petite* perturbation, et nous cherchons la réponse du système. Plus exactement, nous cherchons la fonction $b(t) = B(t) - \bar{B}$, en le supposant petit (ce qui est vrai en tout cas pour les courts temps). Un développement à l'ordre 1 en b de l'équ.(9.7) nous donne alors

$$\frac{db}{dt} = [f(\bar{B}) - \mu\bar{B} + A] + (\left.\frac{df}{dB}\right|_{B=\bar{B}} - \mu)b$$

Or, le terme entre crochet est par définition nul. En appelant α la dérivée de la fonction $f(B) - \mu B$ au point $\bar{B}$, nous avons

$$b(t) = b_0 \exp(\alpha t)$$

Une petite perturbation décroît à zéro si la dérivée au point stationnaire est négative, ce qui est le cas des points P_L et P_H. Par contre, la dérivée au point P_M est positive, la perturbation de ce point ira donc croissant.

9.3 la robustesse.

9.4 Les outils de mesure de la transcription.

9.4.1 Western Blot

9.4.2 Marquage par anticorps et GFP

9.4.3 Les puces d'ADN.

9.5 Le Cancer.

9.6 Rendre confus une image claire : les autres voies de régulation.

9.7 La révolution de l'ARN interférence.

10 Le développement embryonnaire.

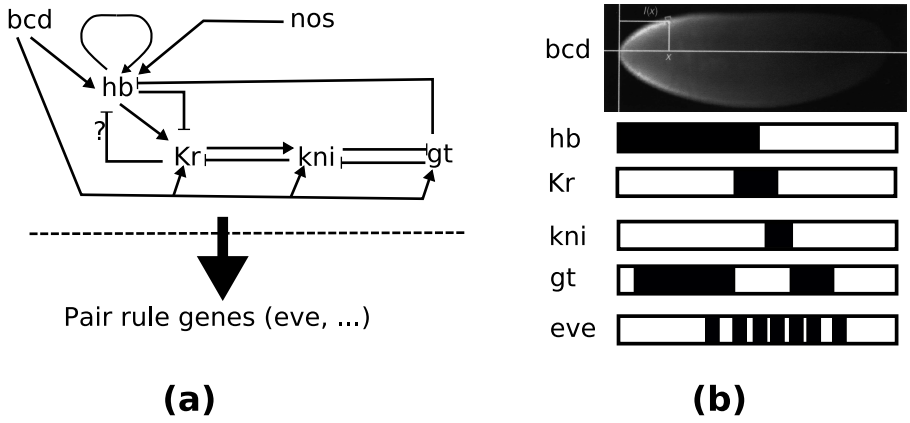

FIG. 10.1: Une partie du réseau du développement embryonnaire, responsable de la différentiation antero-postérieure.

11 L’évolution.