Chapitre 2

Le problème biologique

God is indeed a jealous God
He cannot bear to see
That we had rather not with Him
But with each other play

Poème #1719. Emily Dickinson.
The Complete Poems of Emily Dickinson.
Thomas H. Johnson (Ed.). Faber and Faber, 1975.

Le but de ce chapitre est double. Le premier est de présenter un schéma des principes de base de la mécanique de la vie ainsi que ses constituants fondamentaux. L'idée n'est pas de faire de cette première partie un cours de biologie moléculaire. Des ouvrages excellents et bien plus complets peuvent être trouvés qui remplissent cette fonction, certains de nature didactique [Creighton, 1993] [Daune, 1993] [Freifelder, 1990] [Kruh, 1987] [Lewin, 1994] [Li and Graur, 1991] [Russell, 1992] [Stryer, 1981] [Watson *et al.*, 1987] d'autres de réflexion et donc parfois plus polémiques [Danchin, 1990] [Jacob, 1970] [Jacob, 1981] [Jacob, 1987] [Kaufmann, 1993] [Lorenz, 1975] [Monod, 1970] [Nei, 1987] (le choix ici est personnel : ce sont les livres qui ont été lus ou relus durant le développement de ce travail). L'intention de cette première partie est plutôt de fournir les quelques notions qu'il faut avoir du sujet, d'abord pour situer le problème qui va nous concerner, ensuite pour introduire le vocabulaire qui va être utilisé et attirer l'attention sur les points importants, enfin surtout pour commencer à explorer certaines questions intéressantes que ce sujet soulève. Quelques-unes de ces questions sont commentées dès la fin de cette première partie. Si leur présentation paraît parfois confuse, cela est un reflet de leur complexité, qui entraîne souvent une difficulté à évaluer de façon claire leur légitimité et à en dégager les aspects essentiels. Il est néanmoins important d'essayer de les poser.

Ces questions sont reprises dans le début de la deuxième partie dont le but, le second de ce chapitre, est d'essayer de les caractériser de manière plus abstraite, puis de préciser quel problème a été choisi pour être traité dans le reste de ce travail. Les raisons de ce choix sont multiples, souvent d'ordre purement pragmatiques, et un premier commentaire sur ses avantages et limites clôt ce chapitre. Ce commentaire ainsi que les autres questions laissées de côté sont repris et approfondis plus tard, une fois qu'auront été présentées et discutées diverses façons de traiter de manière algorithmique la question qui a été sélectionnée, façons qui ont été élaborées dans le cadre de ce travail ou par d'autres auparavant.

2.1 Les mécanismes de la vie

2.1.1 Principes de base

2.1.1.1 Transmission de l'information

On appelle matériel génétique le support chimique de l'information transmise d'une génération à une autre chez les êtres vivants. Ce matériel, qui flotte librement dans la cellule (c'est le cas des organismes appelées des procaryotes) ou qui se trouve contenu à l'intérieur d'un noyau (c'est le cas des organismes appelés des eucaryotes) a une double nature : une partie représente l'information qui permet de créer les agents exécutants de la machinerie de la vie, une autre concerne les signaux qui permettent de contrôler cette machinerie. Ce contrôle est nécessaire, d'une part pour répondre rapidement aux modifications des conditions de l'environnement dans lequel se déroule la vie de l'organisme, d'autre part pour des raisons de coordination. Si le premier besoin est plus important chez les procaryotes et les eucaryotes unicellulaires car ils sont plus dépendants des variations du milieu extérieur, le second est essentiel à tous les organismes.

À l'exception de certains virus, ce matériel génétique est toujours stocké dans l'ADN et les agents exécutants de la machinerie de la vie sont essentiellement les protéines. Celles-ci sont obtenues de l'ADN de manière indirecte, leur synthèse se faisant en réalité à partir d'un autre agent moléculaire, l'ARN, dont le rôle précis dépend de son type. L'ARN peut aussi exercer une fonction de contrôle de certains processus liés à cette synthèse. Nous reviendrons là-dessus plus loin.

ADN et ARN sont communément appelés les acides nucléiques et représentent avec les protéines les deux principales macromolécules chimiques des organismes vivants, la troisième étant les polysaccharides.

Ces deux familles de macromolécules sont de longs polymères linéaires synthétisés à partir de petites molécules, appelées monomères, qui sont les nucléotides pour les acides nucléiques et les acides aminés pour les protéines.

Les nucléotides sont constitués par l'association d'une base azotée, d'un sucre qui est le ribose pour l'ARN et le désoxyribose pour l'ADN, et d'un groupement phosphate (voir figure 2.1). Chaque nucléotide est rattaché au nucléotide suivant sur la chaîne, appelée chaîne nucléique, par un lien sucre-phosphate (liaison phospho-diester). Les nucléotides sont au nombre de quatre dans l'ADN, distincts entre eux par la structure de leur base azotée. Les noms de ces bases et les symboles qui servent à les dénoter sont indiqués dans le tableau 2.1. Ces bases forment deux classes, celle des purines et celle des pyrimidines. Ce tableau montre aussi les quatres types de nucléotides présents dans l'ARN. Trois comportent les mêmes bases que ceux de l'ADN, un quatrième a pour base l'Uracile en remplacement de la Thymine mais les propriétés chimiques des deux sont très proches.

Les acides aminés quant à eux sont constitués par un groupement amino NH_2, un groupement acide carboxyl COOH, un atome d'hydrogène et une chaîne latérale représentant un radical organique, le tout lié à un atome de carbone appelé carbone α (voir figure 2.2). Le lien qui unit un acide aminé au suivant le long de la chaîne s'appelle liaison peptide et la chaîne elle-même est appelée chaîne polypeptidique. Les acides aminés sont au nombre de vingt. Ils sont caractérisés par leur chaîne latérale. Leur nom, le symbole servant à les dénoter ainsi que leurs principales propriétés physico-chimiques sont donnés dans le tableau 2.2. On emploie aussi le terme de résidu pour désigner un acide aminé dans une chaîne.

À cause des liaisons chimiques diverses qui peuvent s'établir entre les monomères, nucléotides ou acides aminés, d'une macromolécule, celle-ci adopte en outre une structure dans

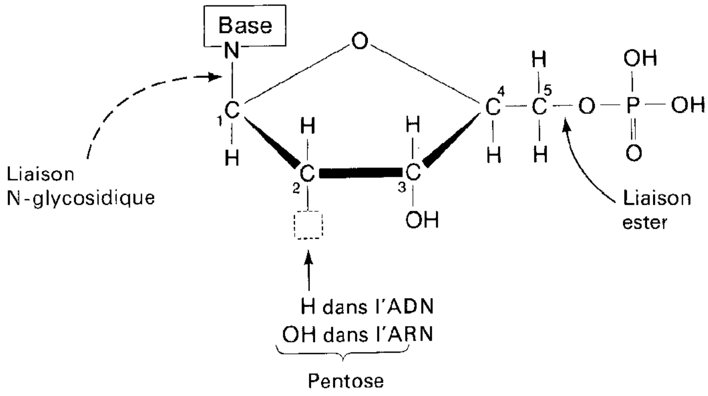

Figure 2.1: Un nucléotide.

Nom	Code
Adénosine	A
Thymine	T
Cytosine	C
Guanine	G
Uracile	U

Tableau 2.1: Les nucléotides.

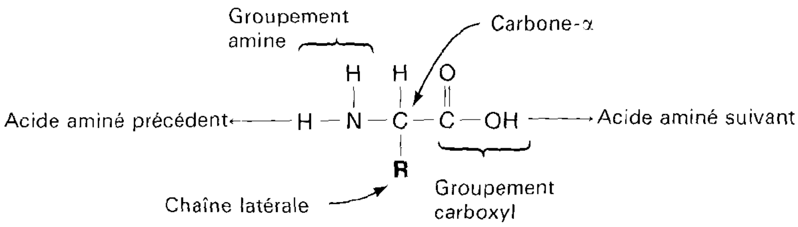

Figure 2.2: Un acide aminé.

Nom	Code à 3 lettres	Code à 1 lettre	Principales propriétés
Alanine	Ala	*A*	très petit, hydrophobe
Arginine	Arg	*R*	positif, polaire
Asparagine	Asn	*N*	petit, polaire
Acide Aspartique	Asp	*D*	petit, négatif
Cystéine	Cys	*C*	petit, polaire, hydrophobe
Acide Glutamique	Glu	*E*	polaire, négatif
Glutamine	Gln	*Q*	polaire
Glycine	Gly	*G*	très petit, hydrophobe ou hydrophile
Histidine	His	*H*	aromatique, hydrophobe, positif, polaire
Isoleucine	Ile	*I*	aliphatique, hydrophobe
Leucine	Leu	*L*	aliphatique, hydrophobe
Lysine	Lys	*K*	hydrophobe, polaire, positif
Méthionine	Met	*M*	hydrophobe
Phénylalanine	Phe	*F*	hydrophobe, aromatique
Proline	Pro	*P*	petit
Sérine	Ser	*S*	très petit, polaire
Thréonine	Thr	*T*	petit, hydrophobe, polaire
Tryptophane	Trp	*W*	hydrophobe, aromatique, polaire
Tyrosine	Tyr	*Y*	hydrophobe, aromatique, polaire
Valine	Val	*V*	petit, aliphatique, hydrophobe

Tableau 2.2: Les acides aminés.

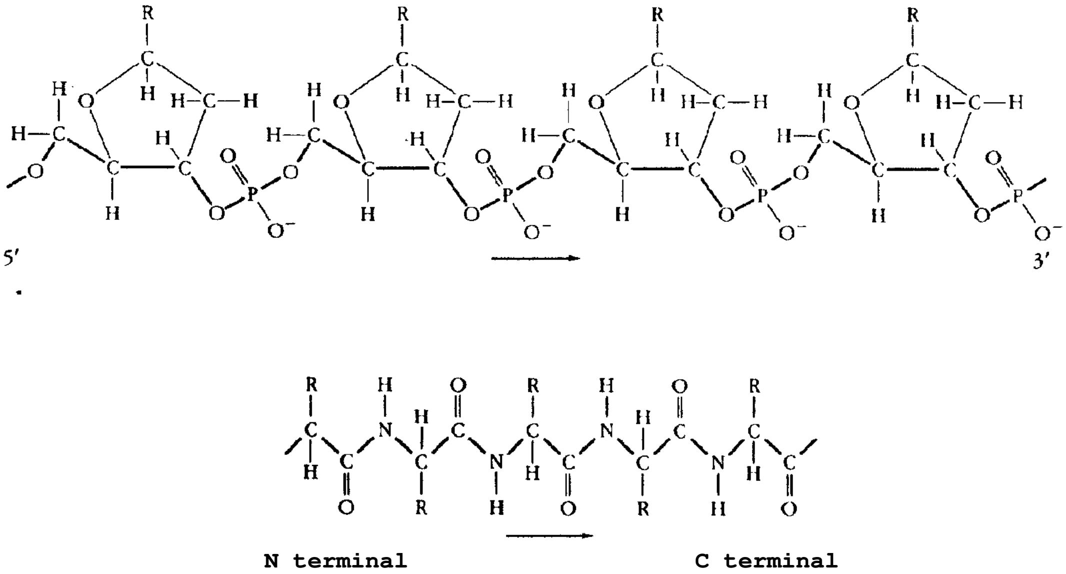

Figure 2.3: Orientation des chaînes nucléiques (figure du haut) et polypeptidiques (figure du bas).

l'espace qui lui est très spécifique. L'asymétrie des liaisons phospho-diester et peptidique implique par ailleurs que les chaînes nucléiques aussi bien que polypeptidiques possèdent une orientation. Celle-ci va de l'extrémité 5' (phosphate) à l'extrémité 3' (sucre) pour les acides nucléiques et du groupement amino dont l'amine N est libre au groupement carboxyl dont le carbone est libre pour les protéines (voir la figure 2.3). Pour l'ADN, la conformation spatiale n'est pas aussi complexe que pour les protéines et, dans une moindre mesure, pour certains ARN mais le rôle qu'elle joue est néanmoins important.

L'ADN est ainsi formé de deux chaînes qui s'enroulent l'une autour de l'autre pour former la structure bien connue en double hélice [Watson and Crick, 1953]. Les deux chaînes sont maintenues solidaires par des liaisons hydrogène entre les bases des deux brins où, pour des raisons stéréochimiques, l'Adénine (A) s'associe toujours avec la Thymine (T) et la Cytosine (C) avec la Guanine (G). Adénine et Thymine d'un côté, Cytosine et Guanine d'un autre sont appelées des bases complémentaires. La synthèse de nouvelles molécules d'ADN procède par la séparation de ces deux brins, suivie de la reconstitution nucléotide par nucléotide de deux autres brins complémentaires. Cette reconstitution donne lieu à la création de deux nouvelles molécules d'ADN, chacune possédant un brin de la molécule initiale (voir figure 2.4). Ce processus semi-conservatif est appelé la réplication.

La synthèse des molécules d'ARN est obtenue de façon similaire à partir d'une molécule d'ADN, sauf que cette fois une seule des chaînes de l'ADN sert de matrice pour la formation d'un unique brin d'ARN et que la base complémentaire de l'Adénine est l'Uracile. Cette synthèse constitue le processus de la transcription.

Trois types d'ARN sont ainsi produits, distincts par leurs structures et par le rôle qu'ils peuvent jouer dans la formation de nouvelles protéines. L'ARN messager sert de matrice dans la synthèse de ces dernières, les ARN de transfert se combinent aux acides aminés qu'ils amènent

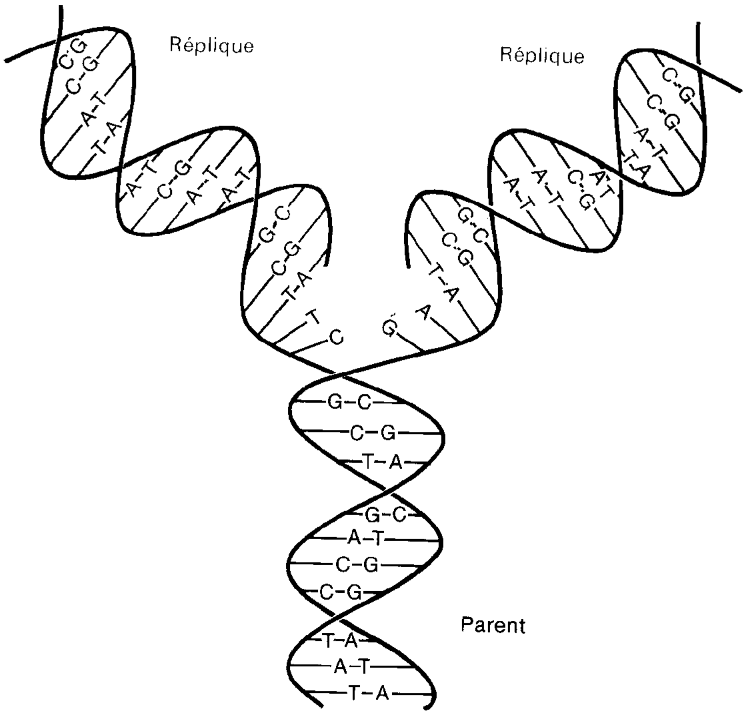

Figure 2.4: Schéma de la réplication.

1 — 2	U	C	A	G	3
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Non sens	Non sens	A
	Leu	Ser	Non sens	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Tableau 2.3: Le code génétique.

sur la matrice pour être polymérisés en protéines et finalement les ARN ribosomaux assistent l'ARN messager dans ce processus de synthèse qui est appelé la traduction.

La séquence d'ARN messager est en relation 3:1 avec la séquence d'acides aminés qui est ainsi obtenue : une suite de trois nucléotides code pour un acide aminé. On appelle un tel triplet de nucléotides un codon. Des 64 triplets possibles, 3 représentent des signaux de terminaison de la traduction et 61 codent pour 20 acides aminés. Ce code, donné dans le tableau 2.3 est donc dégénéré mais non ambigu dans le sens ARN$\rightarrow$protéines. Il est aussi le même dans pratiquement tous les organismes (des exceptions ont néanmoins été relevées chez les mitochondries, les protozoaires et les mycoplasmes).

Ce qu'il faut observer ici c'est que tout l'ADN n'est pas transcrit en ARN, et que tout l'ARN messager n'est pas traduit en protéines. Les parties de l'ADN qui sont transcrites et qui correspondent à ce qui est appelé les gènes peuvent représenter un pourcentage plus ou moins grand suivant les organismes (il est en général élevé chez les procaryotes et petit chez les eucaryotes). Certaines parties n'ont aucune fonction apparente mais d'autres régions qui ne sont pas transcrites jouent malgré tout un rôle très important dans le contrôle à l'échelle moléculaire. Ce point sera vu plus loin. Il en est de même pour certaines régions de l'ARN messager qui ne sont pas traduites en protéines. Ces régions peuvent se situer avant ou après la partie qui sera transcrite ou traduite, elles peuvent également se trouver dispersées au milieu. Cela est le cas pour les ARN messagers des eucaryotes qui présentent en alternance des zones, appelées exons, codant pour des protéines et des zones, appelées introns, dont l'origine et la fonction ne sont pas bien connues et qui sont excisées de l'ARN avant que celui-ci soit traduit (ce phénomène est appelé l'épissage, voir figure 2.5). Par contre, il peut y avoir chevauchement

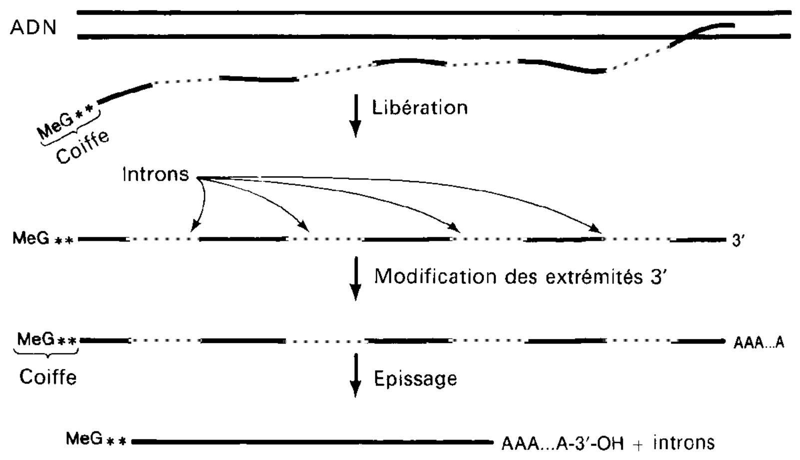

Figure 2.5: Schéma de la transcription et épissage.

de l'information chez les ARN messagers de certains procaryotes où un gène, ou une partie d'un gène, peut servir à la synthèse de plus d'une protéine. Chez ces organismes, la superposition de l'information peut aussi affecter les régions non transcrites ou non traduites qui participent néanmoins au contrôle de ces processus.

2.1.1.2 Le fonctionnement de la 'machine chimique'

C'est par le terme de 'machine chimique' que Jacques Monod [Monod, 1970] qualifie les organismes vivants. Il s'agit de machines capables de se constituer de manière autonome grâce à des "*interactions constructives internes*" dont les agents essentiels sont les protéines.

Les activités que ces dernières exercent au sein d'un organisme sont principalement de deux types, l'une est métabolique et est assurée par une classe de protéines appelées enzymes, l'autre est cybernétique et est réalisée principalement par les protéines dites régulatrices qui jouent le rôle de détecteurs de signaux chimiques et permettent à l'organisme de s'auto-contrôler. Les protéines possèdent en outre un rôle important de transport d'autres molécules, d'électrons ou de protons à l'intérieur ou vers l'extérieur des cellules. Enfin, elles peuvent avoir un rôle plus purement mécanique de maintien de la structure macroscopique. Cette dernière fonction est assurée par les protéines dites de structure qui sont des molécules très allongées. Les autres fonctions de catalyse et de régulation sont quant à elles réalisées essentiellement par les protéines dites de fonction qui sont de loin les plus nombreuses. Ce sont des macromolécules qui se replient sur elles-mêmes de façon extrêmement complexe donnant lieu à une structure généralement très compacte.

Ce qu'il est important d'observer pour l'instant est que toutes les activités chimiques de ces protéines reposent en fait sur leurs propriétés stéréochimiques, c'est-à-dire sur leurs capacités à reconnaître d'autres molécules et que cette capacité est intimement liée à la configuration que ces protéines adoptent dans l'espace.

C'est ainsi à travers une reconnaissance d'"*aires complémentaires*" [Monod, 1970] qu'une protéine se lie à une autre molécule pour former un complexe chimique. La nature de cette complémentarité n'a pas encore été déterminée. Parmi les images qui ont été utilisées, on trouve celle du modèle clé-serrure ou modèle à deux états [Monod *et al.*, 1963] [Monod *et al.*, 1965] ou celle d'un ajustement induit ou modèle séquentiel [Koshland *et al.*, 1966] (voir les figures 2.10 et 2.11). D'autres conceptions plus 'dynamiques' ont été également proposées et font intervenir soit des changements d'états de la structure de la protéine, changements considérés comme relativement rigides car limités par des contraintes stéréochimiques, soit des mouvements internes de la macromolécule pouvant être plus fluides et amples [Creighton, 1993] [Daune, 1993] [Karplus and McCammon, 1983] [Karplus *et al.*, 1987] [Gribskov, 1992].

Chez la plupart des protéines, ces zones de reconnaissance ne concernent qu'une partie relativement très petite de la structure. La localisation de ces zones dépent quant à elle du type d'interaction observé. Celle-ci peut impliquer soit deux protéines non-enzymatiques, soit une enzyme et la molécule sur laquelle elle agit appelée substrat, soit encore une protéine et une molécule d'ADN. Dans les trois cas, les aires en contact lors de la formation du complexe peuvent se situer à la surface des macromolécules. Lorsqu'il s'agit par contre d'une enzyme et de son substrat, ou d'une protéine régulatrice venant se fixer sur une molécule d'ADN, l'interaction peut également concerner des régions centrales de la protéine. Le substrat catalysé par les cytochromes P450 par exemple pénètre entièrement à l'intérieur de la protéine à travers une sorte de canal. La reconnaissance se fait alors sans doute dès l'entrée de ce canal, mais la spécificité de l'interaction est probablement réalisée par la zone de contact interne. Les contacts enzyme-substrat ne sont toutefois pas toujours de ce type, et peuvent aussi avoir lieu à la surface de l'enzyme. Cela est le cas en particulier des immunoglobulines et des récepteurs hormonaux. Un contact en surface est celui le plus souvent observé également dans le cas d'une interaction protéine-ADN. Cependant, une partie de la protéine peut parfois venir 'embrasser' la molécule d'ADN (protéine LeuZipper) ou totalement l'enserrer (protéine Topoisomérase). Ces divers types d'interactions sont illustrées dans les figures 2.6, 2.7 et 2.8.

Ces observations sont importantes et nous y reviendrons ultérieurement, ainsi que sur les structures et, surtout, sur la relation chaîne d'acides aminés, structure et fonction. Mais avant cela, nous allons parler un peu des mécanismes de contrôle d'un organisme.

2.1.1.3 Les mécanismes de contrôle

Les mécanismes de contrôle dont dispose un organisme au niveau moléculaire affectent les trois étapes décrites précédemment : réplication, transcription et traduction. Nous ne parlerons ici que de la régulation de ces deux dernières étapes.

Ces mécanismes sont extrêmement complexes et seuls sont présentés les points qui sont indispensables pour comprendre ce que nous sommes amenés à chercher lorsque nous essayons de les analyser au moyen d'algorithmes.

Comme nous l'avons déjà mentionné, le contrôle à l'intérieur d'un organisme ou d'une cellule est essentiel principalement pour deux raisons : la première est qu'il faut qu'il puisse y avoir une réaction rapide à des modifications survenues dans l'environnement externe ou interne, la seconde est que l'organisme doit pouvoir coordonner de façon efficace tous ses moyens et activités. La plupart des contrôles moléculaires font donc référence à deux types d'information : l'une concerne les signaux régulateurs et fait partie du matériel génétique, l'autre provient presque toujours du milieu extérieur et se réfère en général à la présence ou à l'absence d'une certaine substance chimique dans ce milieu [Lewin, 1994]. Cette dernière est une information qui n'est pas directement stockée dans l'ADN au contraire de l'information génétique, mais elle y est cependant préservée d'une façon indirecte puisque les sites sur l'ADN qui reconnaissent la présence de cette information externe sont conservés.

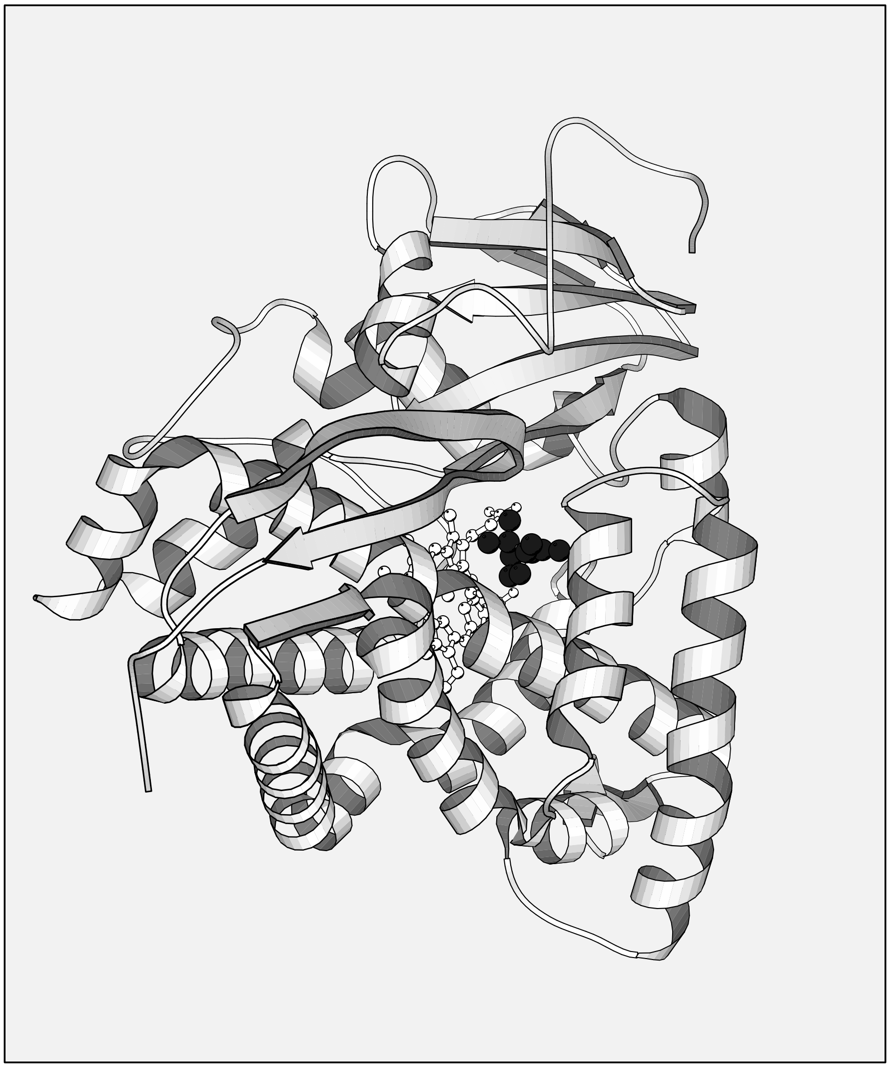

Figure 2.6: Localisation des zones de contact : pénétration du substrat (représenté en noir) à l'intérieur de la protéine (cytochrome $P450_{Cam}$).

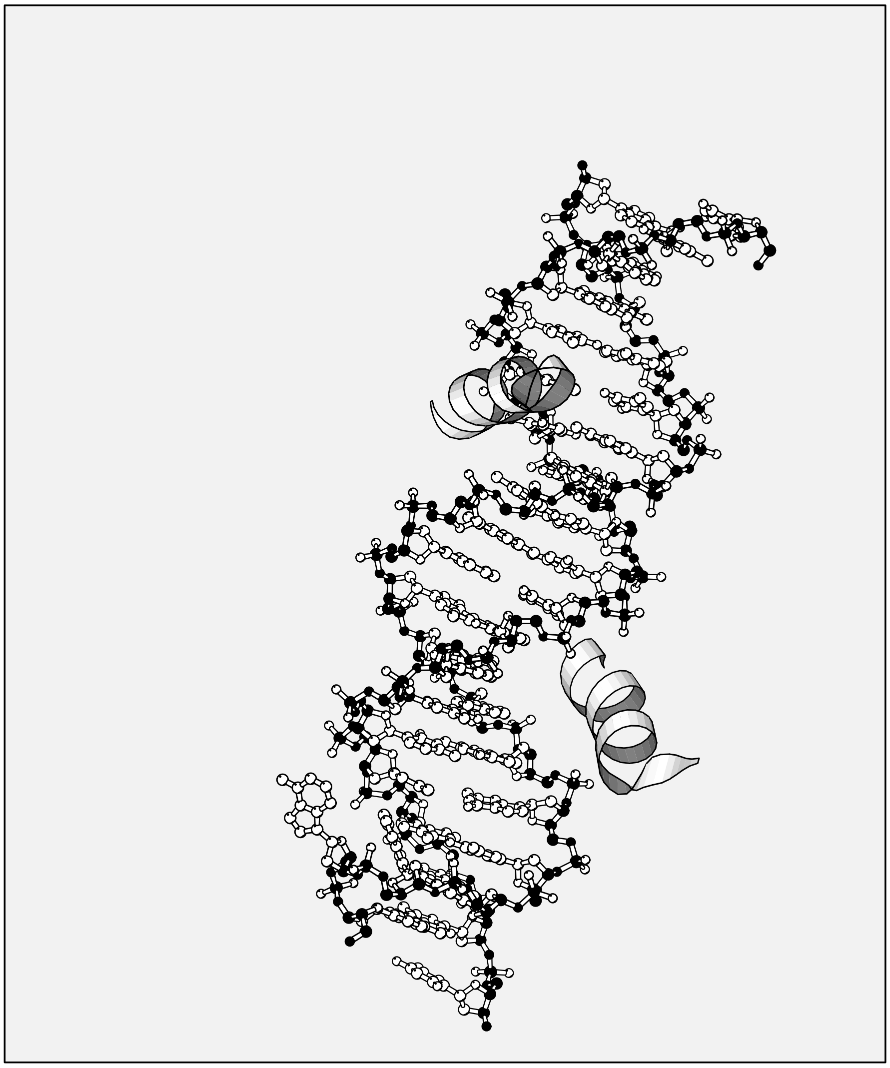

Figure 2.7: Localisation des zones de contact : la protéine LeuZipper (les deux hélices en forme de rubans) 'embrasse' la molécule d'ADN.

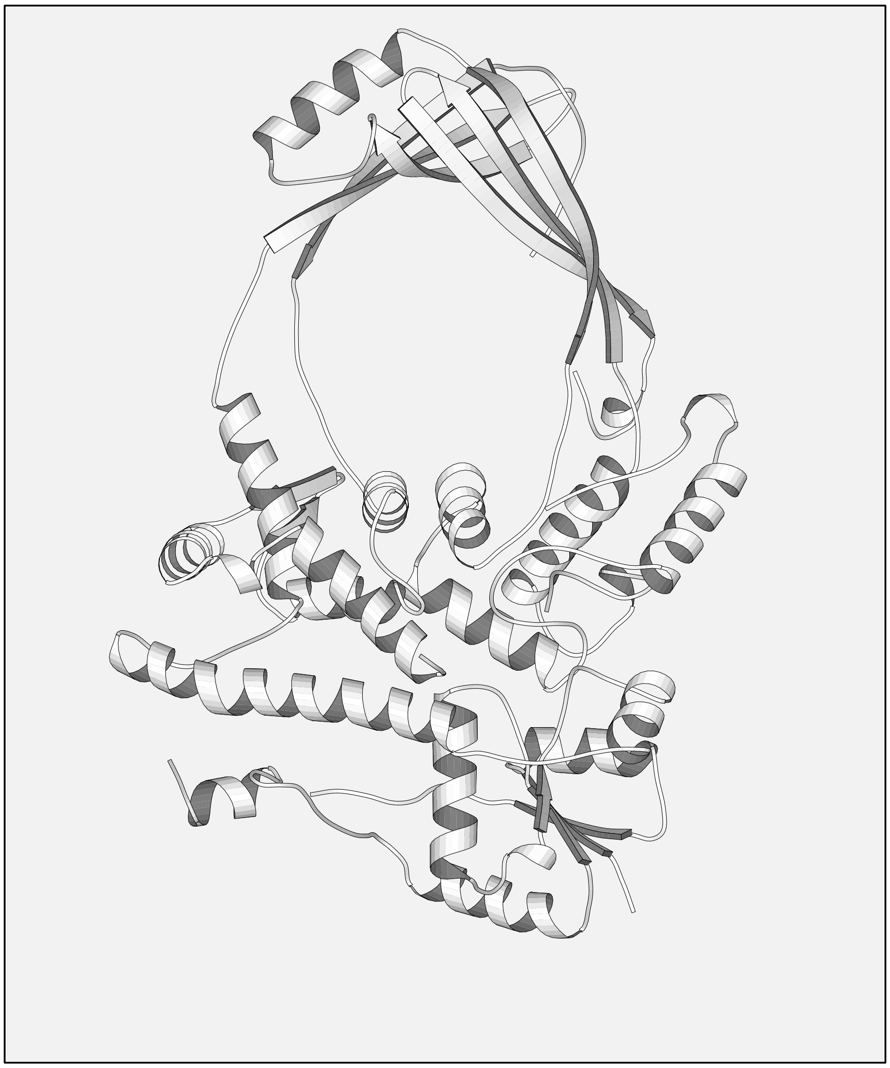

Figure 2.8: Localisation des zones de contact : la protéine Topoisomérase cristallisée sans son ADN mais dont la forme laisse supposer que la molécule d'ADN vient s'enserrer à l'intérieur de la cavité montrée dans la partie supérieure de la figure (la dimension de cette cavité (28 Angström) correspond au diamètre de la double hélice d'ADN).

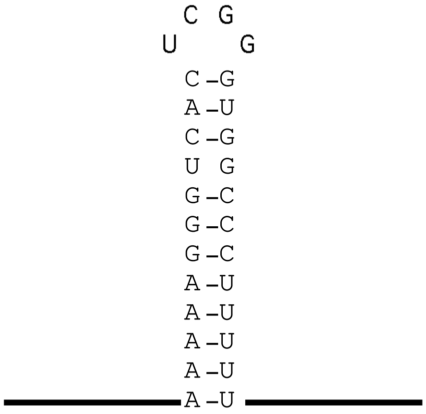

Figure 2.9: Exemple d'une structure secondaire d'ARN : terminateur Rho-indépendant du gène *acc*A de la bactérie *Escherichia coli* (entrée EMBL : EDDNAE).

L'activité de contrôle est réalisée à travers la reconnaissance par une molécule régulatrice d'un signal présent soit sur l'ADN, soit sur l'ARN messager, suivie de la formation (ou de l'empêchement de formation) d'un complexe chimique entre la molécule et la chaîne nucléique. Lorsque la molécule régulatrice est une protéine ou un ARN, la région qui reconnaît, aussi bien que celle reconnue, sont appelées sites.

La nature du site sur l'ADN ou l'ARN représentant un signal régulateur n'est pas toujours connue de manière exacte. Elle peut aussi être très complexe. Ce signal peut ainsi concerner une suite plus ou moins longue de positions contiguës sur ces macromolécules, ou une série de positions discontinues, ou encore une série de suites de plus petites tailles, le signal de chaque sous-suite pouvant moduler l'effet des signaux des suites situées en amont ou en aval sur la chaîne. Il est probable qu'il existe des corrélations à courte ou à longue distance entre les positions. Par ailleurs, d'éventuelles relations de symétrie peuvent impliquer l'existence de propriétés structurales. Ainsi, la reconnaissance du site de terminaison de la transcription de l'ARN messager dans certains organismes se fait par la propre molécule en train d'être transcrite qui adopte une structure spéciale en forme d'épingle à cheveux à l'endroit où la transcription doit prendre fin. Très brièvement, une telle structure peut être décrite comme étant formée d'une tige composée de deux segments complémentaires inversés (palindrome biologique) qui se trouvent appariés, d'une boucle de nucléotides non appariés et éventuellement d'une excroissance située quelque part le long de la tige et constituée de nucléotides également non appariés (voir figure 2.9). D'autres modifications de la conformation d'un ARN messager peuvent faire partie d'un système de contrôle de la fonction exercée par celui-ci. Ces modifications surviennent soit parce qu'une partie de la molécule a été excisée ailleurs, soit parce que l'interaction d'une molécule régulatrice (protéine par exemple) avec un site de l'ARN affecte la structure de ce dernier à un autre site.

La reconnaissance d'un signal de la part des molécules régulatrices, en particulier des protéines, est quant à elle extrèmement liée comme nous l'avons vu à leur structure tridimensionnelle. Le contrôle que les protéines exercent peut impliquer plus d'un site de cette molécule simultanément. Cela est le cas pour la régulation de début de transcription par exemple. Celle-ci peut être négative (la présence de la protéine sur la chaîne nucléique empêche le processus

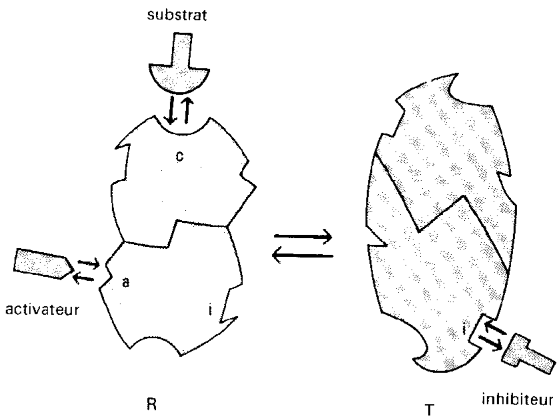

Figure 2.10: Illustration du modèle de Monod, Wyman et Changeux des enzymes allostériques.

de se déclencher) ou positive (sa présence est nécessaire pour que le processus puisse débuter). Il est important de noter ici que, dans les deux cas, l'activité de la protéine régulatrice peut être elle-même modulée par la formation ou non d'un autre complexe avec une molécule plus petite. Cela est le cas des protéines enzymatiques allostériques dont un modèle déjà mentionné a été proposé par Monod [Monod *et al.*, 1963] [Monod *et al.*, 1965] (voir la figure 2.10) et un autre l'a été par Koshland (voir la figure 2.11).

Les enzymes allostériques possèdent ainsi deux types de sites : le site actif qui se combine au substrat et qui catalyse la réaction enzymatique et les sites régulateurs qui peuvent se combiner à un effecteur, soit activateur, soit inhibiteur. Les deux types de sites sont séparés sur la molécule. Dans certaines enzymes ils appartiennent même à des chaînes polypeptidiques différentes. Il existe alors deux états de l'enzyme, un état catalytique et un état inhibé, qui se trouvent en équilibre. Dans l'état catalytique, la forme de l'enzyme lui permet de se combiner à l'activateur et au substrat (effet coopératif). Dans la forme inhibée, seul l'inhibiteur peut se combiner et l'enzyme est inactive. Le passage d'une forme à l'autre est appelée transition allostérique.

Au moins autant que pour certains ARN, la structure d'une protéine est donc essentielle pour comprendre la fonction que celle-ci exerce. La section suivante est consacrée à une rapide description de ces structures puis nous parlerons des facteurs de variabilité à l'échelle moléculaire.

2.1.2 Structures

2.1.2.1 Protéines

2.1.2.1.1 Description de la structure

Plusieurs forces de nature chimique font qu'une protéine nouvellement synthétisée se replie sur elle-même pour adopter une certaine configuration dans l'espace. La structure qui est ainsi créée, et qui est souvent très compacte, est spécifique à la protéine en question. Cela signifie que si une chaîne d'acides aminés identique était synthétisée une seconde fois, elle adopterait la même

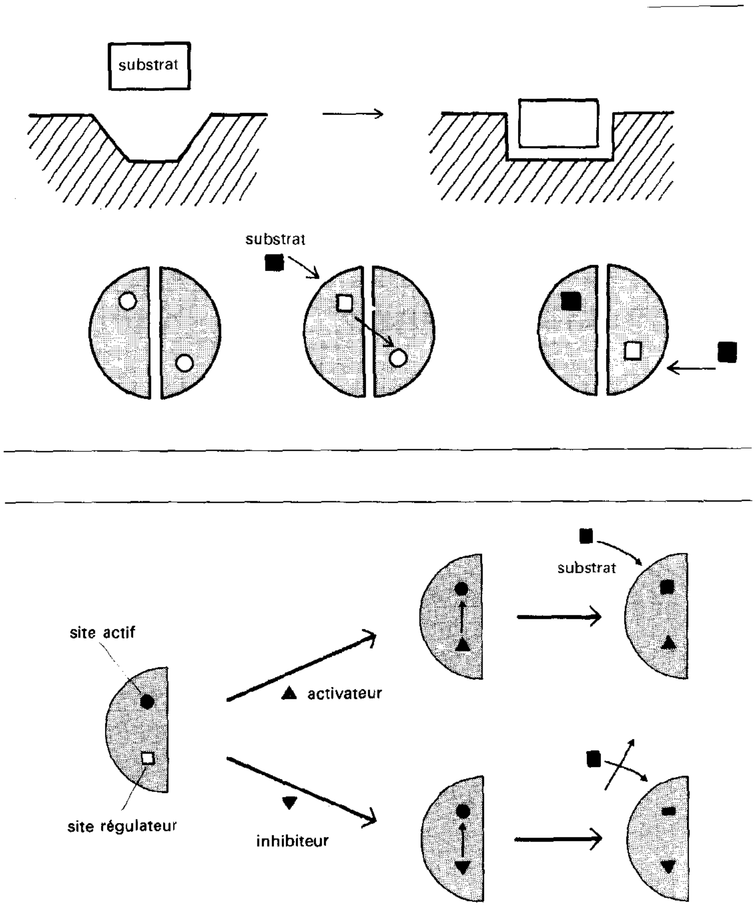

Figure 2.11: Illustration du modèle de Koshland des enzymes allostériques : Théorie de l'adaptation induite de l'enzyme par le substrat.

conformation, ou éventuellement une conformation comportant un très petit nombre d'états distincts, peu distants les uns des autres [Monod, 1970]. Sander [Kabsch and Sander, 1983] et Argos [Argos, 1987] ont montré en effet qu'un certain nombre de pentapeptides (suite de cinq acides aminés d'une chaîne) ne se replient pas toujours exactement de la même façon. La conclusion qu'Argos tire cependant de cette observation est simplement que l'environnement des pentapeptides dans l'ensemble de la protéine joue un rôle essentiel dans leur conformation finale. À ces exceptions près, la relation entre chaîne polypeptidique et structure peut ainsi être considérée comme univoque, mais nous verrons plus loin qu'elle n'est pas bi-univoque. Les lois régissant un tel repliement sont quant à elles mal connues. Il est probable que le repliement commence dès le début de la synthèse de la molécule, avant donc que celle-ci ne soit complètement formée. Une hypothèse qui a également été émise, et qui est de plus en plus acceptée, est que ce phénomène ne se réaliserait pas toujours de manière spontanée mais serait guidé par d'autres protéines appelées des chaperones. Le rôle de ces dernières ne consisterait pas à catalyser le repliement des protéines, mais plutôt à faciliter l'établissement des liens conduisant à une conformation 'correcte' des macromolécules (en général correspondant à celle d'énergie minimale) ainsi qu'à stabiliser et protéger les étapes intermédiaires du processus de repliement [Clarke, 1986] [Gething and Sambrook, 1992]. Elles réaliseraient éventuellement aussi un travail d''édition' dans la mesure où elles seraient capables de reconnaître et de dégrader les protéines mal formées [Creighton, 1993].

La structure d'une protéine peut être décrite à plusieurs niveaux de complexité spatiale.

Le plus simple se réfère à la suite des acides aminés composant la chaîne. Ce niveau est appelé la structure primaire d'une protéine.

Une protéine peut posséder par ailleurs une certaine régularité spatiale locale et ces conformations, impliquant une seule région ou un nombre réduit de régions, forment la structure secondaire d'une protéine. Les principales structures locales possibles sont les hélices α, les feuillets β et les coudes (voir figure 2.12 et 2.13).

La structure dite tertiaire concerne la conformation totale d'une protéine, résultante de toutes les interactions entre acides aminés de la chaîne, que ceux-ci soient par ailleurs impliqués ou non dans une structure secondaire (voir figure 2.14).

Un niveau intermédiaire entre la structure secondaire et la tertiaire, appelé structure super-secondaire, est parfois employé pour décrire certains assemblages de diverses structures secondaires.

Enfin, certaines protéines sont formées de l'aggrégation de chaînes polypeptidiques qui peuvent être identiques ou distinctes. On les appelle des protéines multimériques, et la structure quaternaire fait référence à l'arrangement spatial de cette aggrégation. Chaque chaîne forme une sous-unité de la structure globale, et obéit à des lois de repliement qui peuvent être indépendantes de celles gouvernant le repliement global de la protéine multimérique.

La composition en plusieurs chaînes n'est pas la seule division qu'il est possible de rencontrer dans une protéine. Même celles constituées d'une seule chaîne et qui possèdent plus de quelques centaines d'acides aminés paraissent être composées en fait de deux, ou plus de deux unités de structure appelées des domaines [Creighton, 1993]. Les domaines d'une protéine peuvent interagir entre eux mais apparemment le font de manière moins étendue que les unités de structure situées à l'intérieur d'un même domaine. L'origine de ces domaines est quant à elle mal déterminée.

2.1.2.1.2 Importance de la structure

Plusieurs observations peuvent être faites à ce point.

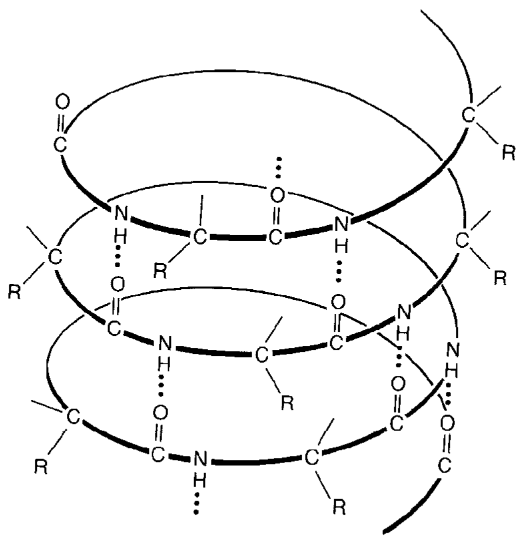

Figure 2.12: Représentation schématique de la structure secondaire d'une protéine : hélice.

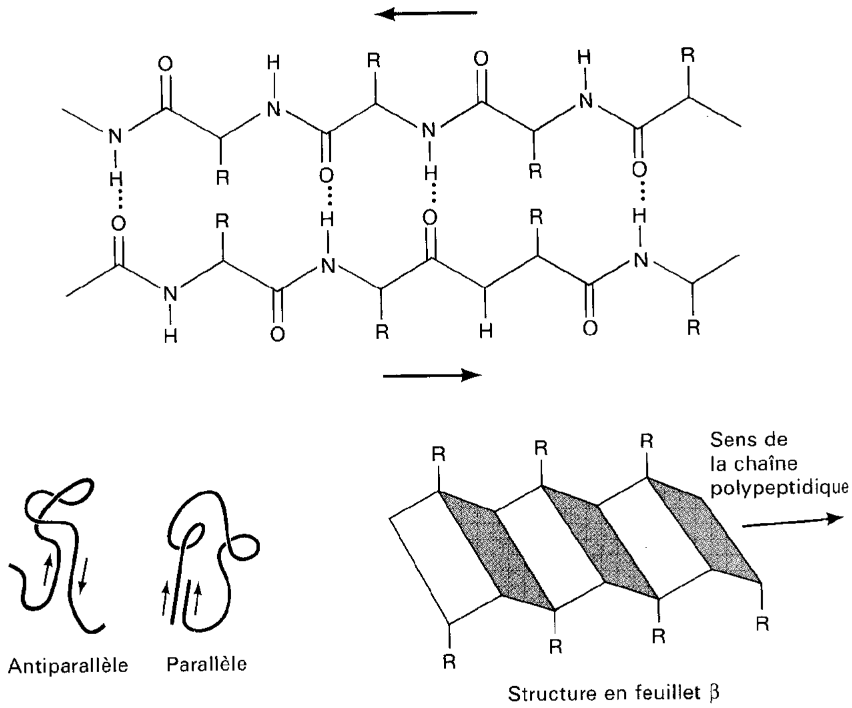

Figure 2.13: Représentation de la structure secondaire d'une protéine : feuillet.

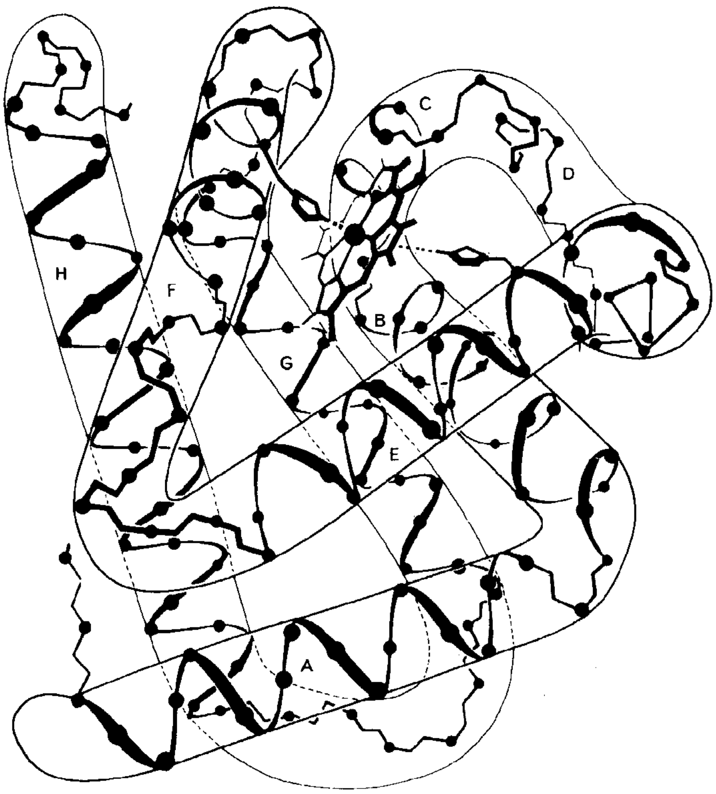

Figure 2.14: Représentation schématique de la structure tertiaire d'une protéine (myoglobine).

La première mentionnée auparavant est que la relation séquence (structure primaire) et structure (conformation spatiale globale) d'une protéine est univoque mais qu'elle n'est pas bi-univoque : deux séquences différentes peuvent ainsi adopter une même structure. Ce fait peut signifier au moins trois choses : soit tous les acides aminés d'une chaîne polypeptidique ne sont pas déterminants pour la structure que celle-ci adopte, ou bien ce qui compte n'est pas l'identité des acides aminés mais tout ou partie de leurs propriétés physico-chimiques, ou encore ce ne sont pas des acides aminés isolés mais des rapports entre acides aminés qui gouvernent la formation de la structure. Bien sûr, il peut aussi y avoir un mélange de ces trois facteurs ou l'existence d'autres encore.

La seconde observation concerne le site de la molécule qui va interagir avec une autre pour réaliser une opération catalytique ou régulatrice (dans le cas des protéines, ce site est appelé site actif). Nous avons vu que ce site n'implique le plus souvent que quelques éléments de la chaîne et que, surtout, il s'agit d'un site structural. C'est la forme qui est la plus importante. Bien sûr cette forme dépend de la composition mais nous venons de voir que la relation séquence-structure n'est pas bi-univoque. Le fait que ce site représente un segment souvent petit de la protéine, qui en outre doit posséder une structure très précise, peut impliquer néanmoins que la nature des acides aminés à ces positions ne puisse varier autant qu'à d'autres endroits sans que la structure locale change et que la fonction soit perdue. Par ailleurs, ce site peut, dans certains cas, concerner des positions qui sont proches dans l'espace mais très éloignées sur la chaîne polypeptidique. Un exemple est celui de la chymotrypsine où seuls trois acides aminés distants l'un de l'autre dans la structure primaire se rencontrent dans la structure tertiaire pour former le site de la protéine. Il faut noter que, dans ce cas, la substitution de l'un quelconque de ces trois acides aminés abolit l'activité catalytique de la molécule. Par contre, dans les protéines qui s'attachent à une chaîne nucléique, le site actif concerne des régions souvent plus étendues. Enfin, chez certaines protéines, en particulier les protéines de structure, c'est toute la structure qui est impliquée dans la fonctionnalité de la molécule.

Une dernière observation que l'on peut faire est, comme nous l'avons vu, que les sites actifs se situent tantôt à la surface de la structure protéique [Creighton, 1993], souvent dans les régions peu structurées appelées boucles [Rooman *et al.*, 1992], tantôt à l'intérieur des protéines. Or c'est dans ce noyau de la structure globale et dans les structures secondaires que la régularité des liaisons ou l'entassement des atomes laissent supposer que les exigences chimiques et stériques relatives aux acides aminés présents à ces positions sont les plus strictes. Ce sont aussi ces régions qui sont sans doute les plus déterminantes de la forme spatiale finale que va adopter la macromolécule.

Il est donc important de noter, avant de clore cette section, que les régions qui sont impliquées dans la fonction d'une protéine, et celles qui sont impliquées dans la formation de la structure, ne sont souvent pas les mêmes. Cela se trouve confirmé par le fait que l'on ait pu observer deux protéines ayant des structures différentes mais une même fonction [Bajaj and Blundell, 1984], ainsi que deux protéines ayant un même repliement global mais des fonctions différentes [Chothia, 1984] [Finkelstein *et al.*, 1993], où par un 'repliement global' on entend les mêmes structures secondaires placées dans le même ordre les unes par rapport aux autres ainsi qu'une topologie générale semblable mais pas nécessairement exactement la même structure [Chothia, 1992].

2.1.2.2 Acides nucléiques

Nous n'allons pas entrer dans le détail des structures possibles de l'ARN, ou de celles de l'ADN dont la double hélice peut également se replier dans l'espace.

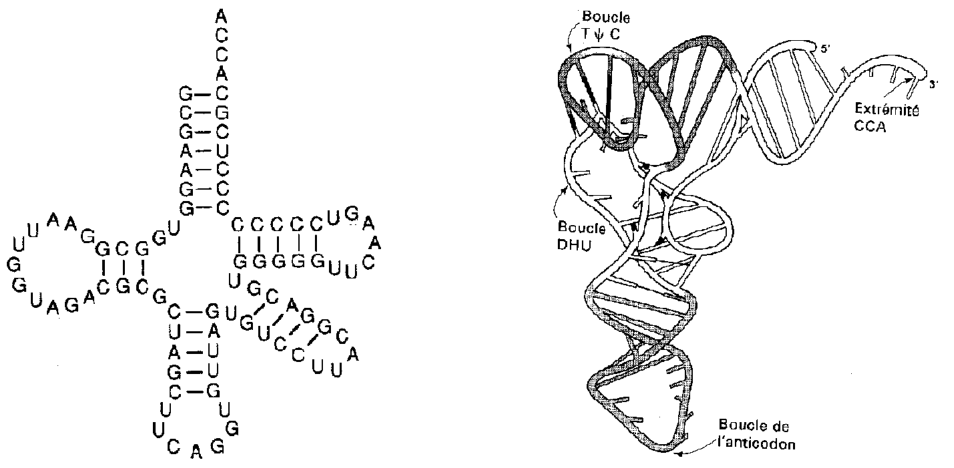

Figure 2.15: Structure en feuille de trèfle d'un ARN de transfert (représentation plate et tridimensionnelle).

Observons toutefois qu'en ce qui concerne l'ARN, les structures secondaires impliquant des nucléotides proches sur la chaîne ribonucléique sont considérées en général comme plus simples que les structures secondaires adoptées par les protéines, dans la mesure où les règles régissant leur formation sont plus faciles à déterminer. C'est en effet principalement par une combinatoire de palindromes biologiques que se créent les diverses structures qui ont pu être observées : en épingle à cheveux comme dans le cas du site de terminaison des ARN messager, en feuille de trèfles (ARN de transfert — voir la figure 2.15) etc.

Ce sont des relations plus complexes entre palindromes qui donnent alors le plus souvent naissance à des formes spatiales tridimensionnelles dans l'ARN. Un exemple est celui des pseudo-nœuds impliquant deux palindromes croisés. Ces formes sont quant à elles au moins aussi compliquées que celles observées dans la structure tertiaire des protéines.

Notons finalement qu'ici, comme dans le cas des protéines, la relation séquence nucléique et structure est univoque mais non bi-univoque. En particulier, il est facile de voir que le simple échange des positions de deux bases appariées sur une tige est déjà une source de variation qui peut ne pas affecter la structure (elle l'affectera si, pour des raisons chimiques, le changement de paire de bases complémentaires rend la structure moins stable que d'autres possibles dans la même région). En règle générale, il n'est donc pas plus aisé que pour les protéines de déterminer la structure secondaire ou tertiaire d'un ARN (ou d'un ADN) à partir de sa séquence.

2.1.3 Facteurs de variabilité

2.1.3.1 Dynamique et évolution : deux temps différents

Dans tout ce qui a été présenté jusqu'à maintenant, le facteur temps n'est jamais intervenu.

Il y a en fait deux temps en biologie, l'un qui est observable de manière directe, et l'autre qui ne l'est que de façon indirecte.

Le premier est relatif à tout changement pouvant intervenir dans une macromolécule, durant le temps de son existence. En général, les changements observables dans une même molécule sont des changements de nature spatiale. Un exemple classique est celui des protéines allostériques qui, comme nous l'avons vu, peuvent pendant le déroulement de leur activité subir une déformation plus ou moins grande de leur structure. Une molécule est ainsi considérée comme un objet dont les mouvements peuvent jouer un rôle non négligeable dans la fonction qu'elle exerce [Daune, 1993].

Le temps qui va nous intéresser est un temps qui ne concerne plus une seule entité mais un ensemble d'entités. Il ne mesure pas ce qui se passe au niveau d'un objet, ou comment cet objet bouge pendant une certaine période d'observation (qui ne peut être plus longue que le temps de vie de l'objet observé, ou que celui de l'observateur), mais, en quelque sorte, l'état relatif d'avancement des objets les uns par rapports aux autres à un moment donné où le terme avancement ne fait référence à aucune valeur qui pourrait être positive (ou négative) et où les différences portent non plus sur la forme des objets comme dans le cas précédent, mais sur leur composition chimique.

Dans le premier cas, l'objet étudié est unique et le temps représente un intervalle. Dans le second cas, le temps est figé et l'intervalle porte sur les objets. Si le passage du temps est observable malgré tout c'est que l'on imagine — ou bien alors c'est ce que l'on veut justement montrer — que ces objets ont une même origine, un ancêtre commun, dont ils sont tous issus et dont ils se sont tous écartés pas toujours de la même façon. Si l'on dit que l'observation de ce temps est indirecte, c'est que l'on ne 'voit' pas en fait les étapes intermédiaires de son action. Seuls sont connus les effets présents, ceux qui sont observables 'maintenant', tous les autres doivent en être déduits. En outre, ce qui rend la chose plus compliquée est que l'action du temps ne peut pas être mesurée par rapport à un même point référentiel, l'ancêtre commun n'étant pas connu.

Les objets en question sont bien sûr toujours les macromolécules, et ces liens, s'ils existent, sont de nature génétique. Les macromolécules subissent en effet des variations au cours de leur vie et de la vie des organismes au sein desquelles elles vivent et se transmettent, et les effets de la variation sont différents entre deux macromolécules de même nature et de même origine. Le temps qui va nous concerner ici est le temps de l'évolution dont l'échelle est bien plus longue que celle du temps de vie d'une molécule. Ce qui va nous intéresser c'est de découvrir parmi un groupe de macromolécules ce qui a changé et comment cela a changé au cours de ce temps. Les raisons de cet intérêt sont revues plus loin, mais avant nous allons examiner quels sont les facteurs qui peuvent introduire une variation au niveau moléculaire.

2.1.3.2 Facteurs de variabilité

La principale source de variation moléculaire sont les mutations. Une mutation est tout simplement un changement survenu à un instant donné au niveau de l'ADN des cellules germinales de l'organisme (et qui est donc transmissible).

Les changements qui affectent une seule paire de base sur la double hélice sont appelées des mutations ponctuelles. La mutation a lieu sur un seul brin de l'ADN. C'est au moment de la réplication, et de la formation d'une nouvelle macromolécule, que cette mutation résulte en une paire de bases différente par rapport à la macromolécule originale puisque c'est la base complémentaire de la base mutée qui vient se placer sur l'autre brin en train de se créer. La mutation peut concerner soit la substitution d'une base par une autre, soit la délétion ou l'insertion d'une base (le terme de délétion plutôt que celui de suppression est employé ici par néologisme). La substitution d'une paire de bases sur l'ADN qui est ainsi observée après réplication implique bien sûr toujours une base purine appariée à une base pyrimidine (pour que la complémentarité entre les deux brins de l'ADN soit préservée). Dans le cas où la mutation

```
séquence originale     : AAGGCAAACCTACTGGTCTTATGT

                                  *
transition (C en T)    : AAGGCAAATCTACTGGTCTTATGT

                                          *
transversion (G en C)  : AAGGCAAACCTACTGCTCTTATGT

                                    ACCTA
délétion (de ACCTA)    : AAGGCAACTGGTCTTATGT

insertion (de AAAGC)   : AAGGCAAACCTACTAAAGCGGTCTTATGT

inversion de
   5' - GCAAAC - 3'
          en
   5' - GTTTGC - 3'    : AAGGTTTGCCTACTGGTCTTATGT
```

Figure 2.16: Exemple de mutations survenant sur l'ADN : transition, transversion, délétion, insertion, inversion.

initiale transforme une purine en une purine ou une pyrimidine en une pyrimidine on parle d'une transition, sinon on parle d'une transversion (voir la figure 2.16).

Des changements portant sur un nombre de paires de bases plus important dans l'ADN peuvent être rangés en deux classes principales. D'abord il y a ceux qui modifient la quantité de matériel génétique présent dans l'ADN : ce sont la délétion et la duplication de segments entiers. La taille de ces segments peut varier considérablement, et la duplication peut produire des segments identiques qui sont contigus ou distants les uns par rapport aux autres dans l'ADN. L'ordre des gènes dans le segment dupliqué peut aussi se trouver inversé par rapport à celui du segment original. Les deux autres types de changements possibles n'affectent pas la quantité de matériel génétique, mais son ordre relatif : il s'agit de l'inversion et de la translocation (aussi appelée transposition). L'inversion intervient lorsqu'un segment d'un brin de la double hélice de l'ADN est excisé puis réintégré de manière inversée ailleurs (voir la figure 2.16 — [Li and Graur, 1991]) et la translocation est un déplacement d'un segment d'ADN intra ou inter-molécules.

Ce qu'il est important d'observer dans tous les cas c'est l'effet que ces mutations peuvent avoir pour la vie de l'organisme. Cet effet est bien sûr différent suivant l'endroit où a lieu la mutation, et son ampleur n'est pas non plus toujours la même. Considérons l'exemple des mutations ponctuelles survenant dans une région de l'ADN où se situe un gène. À cause de la dégénérescence du code, la substitution d'une paire de base sur l'ADN donne ainsi lieu à un changement de codon sur l'ARN qui peut, sur le produit final qui est la protéine :

- résulter en une substitution d'un acide aminé par un autre,
- introduire un des trois codons de fin de traduction à un endroit où la traduction aurait du se poursuivre,

- coder pour le même acide aminé.

Dans ce dernier cas, on parle d'une mutation silencieuse, et de mutation neutre dans le premier cas si l'acide aminé qui se substitue à l'autre possède des propriétés physico-chimiques similaires et que la fonction de la molécule n'est ainsi pas perdue. Toute autre mutation est en général nocive et résulte en une protéine non ou partiellement fonctionnelle. Cela n'est pas toujours le cas pour les délétions ou les insertions si elles sont suivies d'autres mutations sur la même molécule d'ADN qui viennent en partie compenser l'effet de la première. Comme l'ARN est traduit en protéines à travers une relation 3:1, une seule délétion ou insertion à un endroit quelconque de l'ADN va affecter ce que l'on appelle le cadre de lecture de la traduction à partir de cet endroit. Toutefois, si une telle mutation est suivie quelques nucléotides plus loin par une opération inverse (une délétion suivie d'une insertion ou vice-versa) alors ce cadre ne sera affecté qu'entre ces deux positions et la protéine peut demeurer en partie ou totalement fonctionnelle. Le même cas peut se produire pour des délétions ou insertions survenant à la fin d'un gène ou par multiple de trois. Ces dernières mutations résultent en effet en une simple délétion ou insertion d'un (ou de plusieurs) acides aminés sur la molécule sans que les autres résidus soient modifiés.

Si les mutations surviennent dans les régions régulatrices de la transcription, ou affectent les régions régulatrices de la traduction, c'est alors l'expression des gènes qui peut se trouver modifiée. Par exemple, une translocation peut amener un gène qui se trouve dans une région non exprimée (parce que le site portant le signal de début de traduction est absent ou a été trop muté) vers une région où la transcription peut avoir lieu.

Parmi tous ces changements possibles au niveau du matériel génétique, il y en a ainsi qui ne nuisent pas à la vie de l'organisme dans lequel ils sont survenus et qui peuvent donc être préservés par transmission au cours du temps. La suite de ces changements préservés est ce que l'on appelle l'évolution. Toute une controverse existe autour de la détermination des forces qui dirigent cette évolution. La première est l'application au niveau moléculaire de la théorie néo-darwinienne de la sélection naturelle qui affirme que tout changement qui est préservé l'est parce qu'il permet à l'organisme où ce changement s'est produit, et à ses descendants à qui il le transmet, de mieux s'adapter à leur environnement et leur confère ainsi un avantage reproductif. La précision la plus importante portée à cette théorie est celle de la théorie neutraliste de Kimura [Kimura, 1983] qui affirme que la plupart des changements préservés le sont, non pas parce qu'ils confèrent nécessairement un avantage à l'organisme où ce changement a eu lieu, mais parce qu'ils ne lui sont pas immédiatement nuisibles. Si par ailleurs l'ensemble des organismes où un tel changement est survenu est à un moment donné écarté du reste de la population parmi laquelle ils vivaient, par isolement géographique ou écologique, le trait nouveau qu'a pu leur conférer le changement survenu par une mutation au hasard peut alors se trouver renforcé sans qu'il soit pour autant plus ou moins favorable à l'organisme. La théorie neutraliste n'est ainsi pas une réfutation de la sélection naturelle, mais veut modérer l'importance que celle-ci joue dans le processus évolutif au niveau moléculaire. Cette importance est aussi remise en cause quant au rôle joué par la sélection naturelle comme 'créatrice d'ordre' dans le monde vivant. La théorie proposée par Kaufmann en particulier [Kaufmann, 1993] est que la sélection naturelle ne peut pas expliquer tout l'ordre que l'on voit dans les organismes vivants, et qu'elle n'est qu'une force qui vient se greffer sur un ordre qui apparaît de manière spontanée dans tout organisme suffisamment complexe, et que c'est cet ordre-là qui en fait "*autorise, permet et limite*" l'efficacité de la sélection naturelle. Cette théorie correspond en fait à des idées exprimées par Darwin mais qui n'ont pas été reprises par ses disciples, en particulier Spencer.

Ces questions sont absolument passionnantes, mais malheureusement hors du cadre de ce travail. La chose importante à noter cependant, même si cela est exprimé de manière un peu simpliste dans un premier abord, est que l'essentiel ne peut changer pour le 'bon fonctionnement' d'un organisme, c'est-à-dire que certaines choses doivent demeurer les mêmes (ou presque). La section suivante développe cette observation.

2.1.4 Conservation

Une macromolécule peut être considérée sous au moins trois aspects différents : de quoi elle est faite (sa séquence) et quelle forme elle a, à quoi elle 'sert' et quelle est son origine. Ces trois aspects sont liés, mais la relation entre eux n'est pas simple. Nous avons déjà commenté cela à propos des rapports entre séquence, structure et fonction. La relation entre ces trois éléments et l'évolution est encore plus difficile à établir. Si l'on peut affirmer par exemple de deux protéines ayant même séquence, et donc même structure et même fonction, qu'elles sont issues d'un ancêtre commun — l'expression employée dans ce cas est qu'elles sont homologues — la chose est beaucoup moins claire lorsque la similarité ne joue que sur la structure ou la fonction, ou que la similarité de séquence est faible. L'histoire est d'autant plus compliquée que si une ressemblance entre deux structures macromoléculaires peut être expliquée par le fait que celles-ci ont divergé d'une même molécule primitive, elle peut aussi être le résultat d'une convergence au niveau de la structure globale ou locale (qui se produit pour des raisons stéréochimiques).

Ce qui permet d'analyser ces macromolécules malgré tout est ce que nous avons mentionné à la fin de la section précédente : tout ne change pas. En fait, on peut même dire que beaucoup de choses changent assez peu, et surtout, que ce qui est préservé est le plus souvent lié à certaines propriétés importantes de la molécule, ADN, ARN ou protéines. Ceci est particulièrement vrai pour ce qui concerne les sites actifs et les substrats porteurs d'un signal. Il a été observé qu'à ces endroits-là, le taux de mutation est souvent bien moindre que sur le reste de la molécule — on dit que ces régions sont 'mieux conservées'. Cette conservation est donc presque toujours locale plutôt que globale, au moins par rapport aux séquences, et concerne des positions diverses sur la macromolécule qui ne sont pas nécessairement contiguës ni même proches. Comme c'est principalement la forme qui est déterminante par rapport à la fonction ou à la présence d'un signal, et que, d'autre part, plusieurs séquences peuvent acquérir la même conformation spatiale, cette conservation est également beaucoup plus fortement observée au niveau de la structure qu'à celui de la séquence [Creighton, 1993] [Chothia, 1984] [Thornton *et al.*, 1991]. Dans ce cas, elle n'est pas seulement localement mais aussi très souvent globalement vérifiée. On peut exprimer cette différence entre séquence et structure d'une autre façon en disant que la meilleure conservation de certaines régions fonctionnelles est certes relative à l'identité des monomères qui s'y trouvent, mais concerne surtout les propriétés que possèdent ces monomères et les relations qui peuvent s'établir entre eux. La structure étant un reflet direct de ces propriétés et relations, il est donc normal que sa forme soit mieux préservée.

L'étude de ces macromolécules est cependant importante aux deux niveaux de description, séquence et structure. La principale raison en est sans doute qu'aucun des deux niveaux n'apporte à lui seul toute l'information que l'on peut vouloir extraire de ces molécules (même une analyse des deux ensemble n'est souvent pas suffisante).

Celle fournie par l'analyse des structures a l'avantage d'être plus fiable dans la mesure où elle est presque palpable (on peut la 'voir'), et parce que les structures changent moins. Par ailleurs, c'est à travers les structures qu'il est possible d'envisager une étude de la dynamique des molécules qui n'est pas visible au niveau des séquences et qui peut jouer un rôle important pour la fonction que ces molécules exercent au sein des organismes. Enfin, les structures permettent une analyse plus fine des facteurs stériques et physiques contribuant à l'activité d'un site.

L'intérêt que l'on continue à porter aux séquences est quant à lui le plus souvent expliqué par le fait que l'on a pu en déterminer beaucoup plus que de structures (le rapport est environ de 20:1 et à tendance à augmenter). C'est une raison suffisante pour justifier leur analyse, mais elle n'est peut-être pas la plus essentielle. Le fait que les séquences soient non seulement moins bien conservées que les structures, mais aussi conservées localement plutôt que globalement, peut en particulier représenter à la fois une difficulté supplémentaire dans leur analyse et une source de richesse dans l'information qu'elles fournissent.

Cela peut être vrai d'au moins deux façons différentes. Considérons par exemple le cas de protéines qui possèdent la même fonction et qui ont par ailleurs la même conformation globale. Le fait justement que ces molécules soient trop bien conservées structuralement peut rendre difficile l'identification des parties nécessaires à leur activité. Si le taux de conservation au niveau des séquences est plus variable suivant la position, ce qui est presque toujours le cas, c'est une analyse de celles-ci qui peut alors permettre de localiser ces sites, soit parce qu'en général ils sont mieux conservés qu'ailleurs comme on l'a vu, soit parce qu'ils mutent de façon différente [Johnson and Overington, 1993]. Observons qu'il est important dans ce cas de travailler avec des ensembles de macromolécules dont l'homologie est due non à une duplication (paralogie) mais plutôt à un événement de spéciation (orthologie), c'est-à-dire à un événement ayant provoqué l'apparition d'espèces différentes à partir d'un ancêtre commun par suite d'une dérive géographique ou écologique d'une partie de la population. Deux macromolécules paralogues peuvent en effet avoir des fonctions et des modes de régulation divers malgré le fait qu'elles puissent par ailleurs être très proches, en particulier au niveau de leurs sites. La seconde raison pour laquelle la moindre conservation des séquences est une source potentielle d'information est que l'on peut avoir au moins autant à apprendre des différences mutationnelles mêmes subtiles pouvant survenir aux endroits importants pour le fonctionnement d'une macromolécule que de la ressemblance que ces sites peuvent présenter. Par exemple, un signal sur une molécule n'est presque jamais une simple affaire de tout ou rien. Il peut aussi être plus ou moins fort, et ainsi que l'a suggéré Lewin [Lewin, 1994], la force de ce signal peut être modulée non seulement par la présence d'autres signaux aux alentours, mais aussi justement par des types et des taux différents de mutation. Des études dans ce sens ont déjà été réalisées [Carafa *et al.*, 1990] mais il reste encore beaucoup à faire pour établir les conditions permettant éventuellement de traiter ce problème par des approches algorithmiques combinatoires.

Un des problèmes les plus fascinants consiste à établir les conditions dans lesquelles les deux approches peuvent être combinées, d'abord pour que ces approches puissent s'éclairer et s'affiner mutuellement, ensuite pour essayer, sinon de résoudre l'énigme du rapport séquence-structure, au moins de mieux comprendre comment "*l'architecture dans l'espace a pu se trouver engendrée par la simplicité d'une combinaison linéaire*" [Jacob, 1970] et pouvoir ainsi mieux apprécier toute la richesse que cette simplicité cache. Cette richesse est telle qu'apparemment, la seule loi qui semble gouverner la relation existant entre toutes les séquences produisant une même structure est celle du hasard. Mais il est difficile de distinguer le hasard d'une complexité trop grande pour qu'aucun ordre puisse y être immédiatement visible. Ce qui ne veut pas dire que cet ordre n'existe pas.

Le problème séquence-structure n'est pas directement traité dans ce travail mais tout ce qui est présenté ici tente d'aller dans cette direction, en particulier ce qui va nous concerner maintenant est la recherche d'un certain ordre dans les macromolécules, ce qui va se traduire initialement de manière plus concrète par une recherche de ressemblances. La question séquence-structure quant à elle est reprise bien plus tard.

2.2 Recherche d'un ordre

2.2.1 Motivation et critères

L'un des désirs les plus profonds de l'esprit humain est sans doute que les choses dans l'univers puissent présenter un certain ordre.

Toute tentative d'appréhension ou d'exploration de cet univers est donc principalement une recherche d'ordre et finit tôt ou tard par faire appel à l'analogie pour essayer d'attribuer une place et une fonction aux objets de l'univers. Comparer des objets, mesurer leur ressemblance, les classer, distinguer ce qui leur est essentiel de ce qui leur est purement accidentel (bien que le purement accidentel puisse aussi avoir son intérêt) sont ainsi des activités que nous sommes amenés à faire en permanence.

Nous avons vu que l'univers qui nous intéresse ici est celui de la biologie moléculaire et les objets sont les macromolécules. Déterminer ce qu'il est essentiel de caractériser sur ces objets, parfois même réussir à savoir s'il y a quelque chose à caractériser, peut être une tâche complexe. L'idée principale de l'analogie, ce qui motive l'importance qu'on lui accorde dans toute tentative de distinguer une structure dans l'univers, c'est que justement à travers elle "*l'invisible devient visible*" [Jacob, 1970], ou le devient en partie, parce qu'elle permet de repérer ce qui ne change pas, ou change peu, ou encore change d'une même façon et à un même rythme. L'hypothèse est faite que ce sont ces facteurs-là qui donnent un ordre à l'univers. La recherche d'un certain ordre va donc commencer par une recherche de facteurs qui se ressemblent, ceux qui changent peu les uns par rapport aux autres dans le temps ou dans l'espace, c'est-à-dire parmi un ensemble d'objets qui sont par ailleurs différents.

Ce premier critère étant fixé, tout n'est pas résolu pour autant car la difficulté de cette recherche va dépendre bien sûr aussi de la nature de ce qui change peu. Un objet peut être vu comme un ensemble d'objets 'concrets' plus élémentaires arrangés entre eux d'une certaine manière (les nucléotides et les acides aminés, ou à un niveau encore plus élémentaire, les atomes composant ces monomères), ou comme un ensemble de propriétés (en général physico-chimiques), ou encore comme un ensemble de relations. Il peut aussi être vu sous une forme statique ou dynamique. A priori, toute régularité observée à l'un quelconque de ces niveaux (ou à plusieurs d'entre eux simultanément) est potentiellement intéressante. La première difficulté ici consiste à établir la fréquence qu'un phénomène doit présenter pour qu'on le qualifie de régulier : est-ce une fréquence très forte ou très faible, ou simplement une fréquence différente de la normale, mais alors comment mesurer le 'normal'? La seconde difficulté est qu'il peut y avoir un très grand nombre de types de régularités. Quelques exemples ont été mentionnés dans la première partie de ce chapitre. Si l'on ne veut pas avoir à entrer dans des considérations de valeur qui peuvent être subjectives, il faudrait pouvoir établir des définitions précises de toutes les régularités qui semblent pertinentes ou en effectuer un classement. C'est une tâche qu'il serait très intéressant d'essayer de réaliser mais qui est impossible à envisager complètement. En général, des choix sont donc faits qui dépendent de ce qui est observé, qui lui même dépend des hypothèses que les personnes se font sur le fonctionnement des objets étudiés (c'est-à-dire dépend en fin de compte de leurs expectatives). Enfin, de manière plus prosaïque, ces choix dépendent aussi de ce qu'il est possible de chercher avec les instruments dont nous disposons en un temps qui soit raisonnable.

La façon de procéder est donc celle indiquée par François Jacob [Jacob, 1970] : "*le plus simple doit éclairer le plus complexe*". C'est aussi l'idée présentée par Gombrich [Gombrich, 1987] à propos de l'exécution d'un tableau mais qui s'applique tout aussi bien à la recherche qui nous intéresse : "*À défaut d'un point de départ, d'un schéma initial, nous ne pourrions jamais parvenir à contrôler le flux mouvant de l'expérience. Nous serions incapables de classer, de mettre* *nos impressions en bon ordre, si nous ne disposions pas de catégories. Assez paradoxalement, on s'est aperçu que la nature de ces premières catégories n'avait pas une grande importance, il nous est toujours possible de l'adapter à nos besoins, ... pourvu que le dessin initial ne contienne aucune indication fausse*".

Une manière de simplifier toute recherche consiste donc à imposer des limites à l'univers que l'on cherche à explorer, soit en restreignant son champ, soit en établissant des contraintes à la méthode servant à effectuer cette exploration. Les contraintes peuvent être diverses, mais l'important est qu'elles soient toujours clairement définies de telle sorte que leur introduction ne rende pas l'interprétation des résultats obtenus impossible à réaliser, soit parce qu'elles déforment la réalité, soit parce qu'elles la brouillent encore plus.

Nous allons voir dans la section suivante les principales contraintes et simplifications que nous avons choisies d'établir dans le cadre de ce travail.

2.2.2 Le cadre spécifique de notre recherche

La première contrainte que nous allons imposer au modèle adopté pour analyser et comparer les macromolécules, ADN et ARN ou protéines, est que celui-ci tienne compte de l'ordre linéaire qui existe naturellement dans ces molécules. Cela signifie que nous allons travailler avec des chaînes de caractères. Celles-ci peuvent donc représenter directement la structure primaire, c'est-à-dire la séquence, des macromolécules où, comme on l'a vu, à chaque monomère correspond un symbole sur un alphabet de taille 4 ou 20. Elles peuvent aussi, dans le cas des protéines, être obtenues de la structure par un codage des coordonnées internes de la chaîne des carbones α. Il n'est pas important pour l'instant de comprendre comment cela est fait, simplement de noter que, dans tous les cas, ce que l'on obtient est un ensemble de chaînes sur un certain alphabet plus ou moins grand, et que ce sont ces chaînes que l'on compare. Le terme générique de chaîne est adopté ici pour tenir compte du fait que les objets désignés représentent tantôt des séquences, tantôt des structures dans l'espace codées en une suite linéaire de caractères. Ce terme a été préféré à celui de texte dont nous avons voulu éviter la connotation linguistique.

La seconde contrainte va porter alors sur la façon de comparer ces chaînes. Il y en a deux principales, la première consiste à les comparer globalement, la seconde à réaliser des comparaisons locales. Dans ce second cas, il s'agit seulement de repérer certains objets qui sont communs à ces chaînes. Nous présentons ici les deux approches, globale et locale, bien que notre choix se soit porté sur la locale. Les objets en question représentent toujours pour nous des suites contiguës de symboles de l'alphabet présents dans ces chaînes. Nous utilisons le terme de mot pour désigner ces suites de symboles. Dans le cas des séquences, un mot dénote bien un motif lexical et correspond ainsi à l'usage courant du terme, dans celui des structures un mot correspond en fait à un motif structural, c'est-à-dire à une forme géométrique sur laquelle joue une contrainte de séquentialité. Ce que l'on cherche dans les deux situations ce sont des mots qui se ressemblent. La raison de ce choix est double. Le premier est que ces mots sont intéressants en eux-mêmes. L'activité d'une macromolécule étant souvent liée à certaines régions uniquement, ce sont celles-ci qu'il faut essayer d'identifier. La seconde est plus philosophique et correspond à une vision de la nature des objets comparés que nous pouvons illustrer avec l'image du bijou dans son écrin. Supposez que l'on vous présente deux boîtes avec chacune dix petites pierres semi-précieuses et que l'on vous demande d'estimer la valeur relative des boîtes. Vous n'allez sans doute pas vous préoccuper de la taille, forme ou couleur de la boîte elle-même. Celle-ci peut bien sûr avoir un effet subjectif sur votre estimation de la valeur des pierres (elles peuvent par exemple en rehausser l'éclat), mais ce sont sur ces dernières que vous allez concentrer votre attention et que vous allez donc comparer. Cette image du ou des bijoux dans leur écrin n'a pas été choisie au hasard, elle a en effet déjà été utilisée par rapport aux macromolécules biologiques, plus spécialement les protéines [Ptitsyn and Volkenstein, 1986], l'idée étant vraiment que seulement certaines parties de la molécule sont importantes pour expliquer sa fonction (les bijoux) et que tout le reste de la structure (la boîte) ne joue qu'un rôle de protection de ces parties qui sont essentielles. Si cette hypothèse est admise, il paraît de bon sens de mesurer la ressemblance entre objets par une comparaison de ce qui leur est essentiel uniquement, pas du reste. Repérer des mots similaires d'abord est donc une façon de comparer des chaînes, cela peut en constituer en tout cas une première étape importante. Est-ce vraiment une bonne stratégie? Plus généralement, est-ce que l'image elle-même du bijou et de la boîte est correcte? Nous laissons une discussion autour de ces questions pour plus tard. Bonne ou mauvaise, cette méthode est celle que nous adoptons. Nous parlerons aussi des méthodes de comparaisons globales portant sur les chaînes entières, mais en ce qui concerne ce travail ce que nous comparons sont des mots. Notons aussi par ailleurs que, bien que nous allons nous intéresser principalement à identifier les traits communs à un ensemble de chaînes, la recherche de mots peut être faite sur une seule chaîne. Ce que nous cherchons alors sont des mots répétés et des difficultés algorithmiques supplémentaires sont introduites suivant que l'on exige que ceux-ci soient contigus ou non, ou vérifient certaines propriétés de symétrie, directes ou miroir (c'est le cas des palindromes). Comme pour les mots communs à plusieurs chaînes, les mots répétés qui peuvent nous concerner sont bien sûr non seulement ceux exactement identiques, mais aussi ceux qui sont simplement similaires entre eux.

Une troisième contrainte (en fait il s'agit d'une famille de contraintes diverses) va jouer alors précisément sur les critères permettant de mesurer le degré de ressemblance de deux ou plusieurs chaînes, ou de considérer que deux ou plusieurs mots sont similaires.

Dans le cas d'une recherche de mots, une quatrième et dernière contrainte porte sur la fréquence avec laquelle un mot ou ses similaires doit apparaître dans une ou plusieurs chaînes pour qu'on le qualifie d'intéressant. Suivant la définition de ressemblance qui est établie entre les mots, il existe des critères statistiques qui disent quelles sont les fréquences d'apparition de ces mots qui sont 'anormales' parce qu'elles sont plus grandes ou plus petites que les fréquences avec lesquelles on peut espérer trouver un mot pris au hasard dans une ou plusieurs chaînes. Dans ce travail ces critères ne sont pas utilisés directement dans la recherche des mots similaires mais plutôt a posteriori. Initialement, nous établissons simplement que les mots similaires qui vont nous intéresser sont ceux qui apparaissent au moins (ou au plus) un certain nombre de fois. Si l'intérêt que peut présenter un mot apparaissant très souvent est assez intuitif à comprendre, celle d'un mot apparaissant peu souvent peut le sembler moins et demande donc une explication. En fait cet intérêt devient plus clair dès que l'on couple à la fréquence du mot sa localisation. L'hypothèse a ainsi été faite que certains signaux indiquant la jonction exon-intron dans l'ARN messager peuvent être constitués par des suites de nucléotides qui apparaissent avec beaucoup moins de fréquence dans la région des exons proche de la jonction que dans l'intron avoisinnant.

Enfin, si nous présentons des approches variées quant à la méthode employée pour effectuer ces comparaisons une fois fixés les critères donnés ci-dessus, celle que nous avons adoptée est une approche combinatoire. Cela veut dire que nous cherchons, de manière exacte, tous les mots qui vérifient la définition de ressemblance que nous avons choisie et qui apparaissent au moins (au plus) le nombre de fois que nous avons établi dans les chaînes à comparer.

2.2.3 Un premier commentaire concernant notre approche

Avant de clore ce chapitre entièrement dédié à la biologie, nous voudrions faire un premier commentaire concernant l'approche que nous avons choisie.

La première observation porte sur le type de régularité recherché — des mots communs à un ensemble de chaînes, ou qui se trouvent répétés dans une seule chaîne. Nous ne voulons pas nous attarder pour l'instant sur les limitations que peut présenter ce choix mais il est important de noter que ce qui va nous intéresser ce sont rarement des mots pris individuellement, mais plutôt des ensembles de mots qu'unissent certaines propriétés. Par exemple, nous allons vouloir identifier tous les ensembles de mots tels que, à l'intérieur d'un même ensemble, n'importe quelle paire se ressemble suffisamment. Peu importe pour l'instant comment est définie cette 'ressemblance suffisante', ce qu'il faut observer c'est que l'objet qui nous intéresse — l'ensemble — représente souvent un niveau de description plus riche que celui des éléments qu'il contient. La caractérisation de ces ensembles va dépendre bien sûr des critères choisis pour grouper des mots, c'est-à-dire de la définition de ressemblance adoptée.

Une seconde observation concerne précisément ces définitions. Dans le cas des structures, il est relativement aisé et immédiat de déterminer les critères qui permettent de mesurer la ressemblance entre objets car ceux-ci sont des formes géométriques, et il est très intuitif de considérer deux d'entre elles comme étant similaires si elles peuvent être approximativement superposées, après éventuellement une translation et une rotation de l'une par rapport à l'autre. La définition de ressemblance est ici assez peu sujette à polémique parce que les objets en question ont une claire identité. Mais dans le cas des séquences, cela n'est plus vrai. Même à leur niveau le plus élémentaire, celui des monomères, les objets que l'on compare sont des entités à multiples facettes, composés d'un ensemble de propriétés physico-chimiques et de relations qui dépendent non seulement de l'élément concerné mais aussi des rapports qu'il entretient avec les autres éléments de la molécule et de sa position le long de la chaîne. Les critères permettant de déterminer la ressemblance éventuelle de ces objets sont donc beaucoup plus fortement liés à ce que l'on essaye d'identifier, toutes les facettes n'ayant pas la même importance suivant ce que l'on recherche. Un point de vue est ainsi introduit dès le premier abord, dont il est souvent difficile de juger la portée et les conséquences. Cela n'est pas toujours mauvais et peut même être utilisé à bon escient, mais il faut savoir que ces choix qui sont faits a priori vont se trouver nécessairement reflétés dans les résultats obtenus. Nous allons présenter dans ce travail une façon possible d'essayer de se libérer de ce point de vue.

Une troisième observation concerne le fait que, contrairement aux définitions, la recherche de mots communs est elle, par contre, complètement détachée de toute considération biologique. Comme nous l'avons vu, une fois défini un critère de ressemblance entre mots, tous les objets qui se 'ressemblent' suivant cette définition sont localisés quel que puisse être par ailleurs leur intérêt biologique. Une question que l'on peut se poser alors concerne par exemple la pertinence de l'information qu'ils représentent. Est-ce que ces mots trouvés permettent de bien discriminer soit un lien évolutif ou structural, soit une fonction ou un signal biologique éventuellement présent sur les séquences et que l'on espérait identifier par comparaison de celles-ci? Les méthodes introduites dans ce travail ne font qu'aborder cette question, qui n'est d'ailleurs pas une question à laquelle il est facile de répondre quel que soit le moyen utilisé. C'est pourtant un problème très important.

Ce que l'on peut dire pour l'instant c'est que les choix qui ont été faits ont l'avantage d'être clairs et simples. Cette simplicité justement permet finalement une assez grande souplesse dans l'exploration de la notion de ressemblance. Est-ce que cette simplicité préserve la 'véracité du monde observé' de telle sorte que le dessin initial qu'elle permet d'en fournir "*ne contienne aucune indication fausse*" pour reprendre l'observation de Gombrich [Gombrich, 1987]? C'est une interrogation sur laquelle nous reviendrons à maintes reprises au cours de ce travail. Pour l'instant, c'est de la notion de ressemblance dont nous allons parler dans le chapitre suivant.