Université de Montréal

Propriétés relationnelles tridimensionnelles des structures d'acides nucléiques

par
Abdelmjid Ftouhi
Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en informatique

Septembre, 1997

Université de Montréal

Faculté des études supérieurs

Ce mémoire intitulé:

"Propriétés relationnelles tridimensionnelles des structures d'acides nucléiques"

présenté par:
Abdelmjid Ftouhi

a été évalué par un jury composé des personnes suivantes:

Jean Meunier (président rapporteur)

François Major (directeur de recherche)

Gena Hahn (membre du jury)

Mémoire accepté le: 7 décembre 1997

"Il peut arriver que de petites différences dans les conditions initiales en engendrent de très grandes dans les phénomènes finaux. Une petite erreur sur les premières produirait une erreur énorme sur les dernières. La prédiction devient impossible et nous avons le phénomène fortuit... Une cause très petite qui nous échappe, détermine un effet considérable que nous ne pouvons pas ne pas voir, et alors nous disons que cet effet est dû au hasard"

Jules Henri Poincaré

Remerciements

Je tiens à remercier tout particulièrement mon directeur de recherches, François Major, pour ses encouragements et ses conseils précieux qui ont été un enrichissement dans l'évolution de mon travail. Le support financier et la disponibilité qu'il m'a accordés ont été un grand stimulant pour mener à terme ce mémoire. Je remercie Sergey Steinberg, attaché de recherche au laboratoire de Robert Cedergren au département de biochimie à l'université de Montréal, pour ses discussions sur le sujet.

Je profite de cette occasion pour exprimer ma reconnaissance la plus vive à ma chère aimée Lise pour l'encouragement et le soutien moral qu'elle m'a offerts durant toute cette période d'étude.

Sommaire

L'analyse des structures nucléiques d'une manière générale devient de plus en plus importante face à la croissance des structures moléculaires tridimensionnelles trouvées. Le coeur de ce mémoire est une présentation des concepts de biochimie et de mathématiques qui formeront un programme d'analyse des structures tridimensionnelles de molécules d'acides nucléiques. Le phénomène d'empilement des bases azotées dans une structure d'acide nucléique est encore mal défini. Les deux sortes de bases, purines et pyrimidines, tendent à être empilées sous forme de molécules planaires parallèles. Ceci a été observé dans plusieurs centaines de structures cristallines par rayon-X contenant des bases d'acides nucléiques. Ces bases sont souvent empilées une sur l'autre partiellement. Apparemment ces empilements sont dûs à des interactions hydrophobiques et électrostatiques.

Dans le premier chapitre, on parlera de ces composés chimiques connus sous le nom d'*acides nucléiques*, existant sous forme de deux types ADN et ARN. Les structures de l'ADN et de l'ARN sont largement décrites par des empilements de bases d'acides nucléiques. En contraste avec les liens hydrogènes responsables des appariements de bases, les interactions de l'empilement de bases sont beaucoup plus importantes dans la stabilisation des structures hélicoïdales présentes au sein des acides nucléiques. Ainsi, on discutera de l'importance de l'analyse des empilements de bases pour comprendre les liens qui peuvent exister entre la structure d'une molécule et sa fonction biologique dans la nature. On montrera alors l'utilité de l'empilement de bases pour la modélisation dans le but de prédire des structures moléculaires tridimensionnelles. Cela a été intégré, par exemple, au système de modélisation MC-SYM [26].

Jusqu'à maintenant, pour localiser et identifier les interactions d'empilement de

bases dans une structure moléculaire d'ADN ou ARN, on se base sur la visualisation ou des méthodes décrivant le profil d'énergie potentielle de ces interactions. Ces deux procédures présentent beaucoup de faiblesses. Pour la première, elle reste ardue et demande plus de temps; de plus, elle ne permet pas de comprendre le comportement de l'empilement ni de l'évaluer. Aussi les avis peuvent différer d'un visualisateur à un autre dans certains cas limites. Pour la deuxième méthode, premièrement les contributions relatives des forces stabilisatrices de l'empilement des bases sont mal connues et deuxièmement les composantes hydrostatiques et hydrophobiques de l'empilement de bases sont mal décrites par les forces d'énergie potentielle existantes.

Dans le deuxième chapitre, nous présenterons une approche automatisée basée uniquement sur des notions de géométrie, dans le but de chercher l'empilement base-base dans une molécule d'acide nucléique. Nous étudierons ainsi avec cette approche les empilements présents au sein de l'ARNt^{Phe} et d'une molécule connue sous le nom de *hammerhead*. De plus, plusieurs paramètres propres à un empilement de deux bases seront analysés. L'empilement de bases sera intégré au système de modélisation MC-SYM, formant ainsi avec un algorithme de recherche d'appariement de bases dans une molécule d'acide nucléique le programme NANOTE, qui servira d'un guide dans les arbres de recherche pour la prédiction d'une structure tridimensionnelle pour une certaine molécule donnée.

Dans le troisième chapitre, on discutera d'un exemple de modélisation à l'aide de MC-SYM pour la prédiction de la structure tridimensionnelle d'une molécule appelée *leadzyme* [25]. La modélisation de cette molécule nécessitait un choix parmi différentes hypothèses structurales, pour aboutir à un modèle moléculaire représentant les contraintes connues. Un modèle mathématique basé sur la logique floue sera présenté pour sélectionner les hypothèses structurales suivant leur possibilité de con-

tenir les conformations actives du leadzyme [27]. Le modèle est appliqué à une relation séquence-structure de la molécule du leadzyme. L'hypothèse structurale la plus plausible sélectionnée par le modèle de logique floue a été retenue pour développer d'autres expériences biochimiques, déduisant de nouvelles données et contraintes. Ce procédé a permis de construire un modèle final qui explique la capacité catalytique du leadzyme.

Table des matières

Liste des figures

Introduction

Comprendre comment les structures tridimensionnelles des macromolécules, protéines et acides nucléiques, affectent leurs activités biologiques est un problème qui stimule l'intérêt de plusieurs chercheurs. Ceux qui sont à la recherche de nouveaux traitements, pour les maladies virales par exemple, souhaitent identifier des composés capables de reconnaître spécialement les séquences d'acides nucléiques uniques à chaque génome viral, de manière à servir comme drogue spécifique. Dans ce but, connaître les structures tridimensionnelles des molécules impliquées constitue un atout.

Un très grand nombre de macromolécules provenant de plusieurs organismes différents, ont été identifiées en déterminant leur séquence d'acides aminés, pour une protéine, ou de nucléotides, pour un acide nucléique. Cependant, la méthode classique, la cristallographie à rayons X, ne permet pas dans certains cas de déterminer la structure tridimensionnelle d'une macromolécule; elle est coûteuse et nécessite beaucoup de temps. Par le fait même, l'écart entre le nombre de séquences identifiées et le nombre de structures connues est énorme. Ce qui explique le besoin de développer des approches théoriques pour prédire la structure d'une macromolécule à partir de sa séquence. Théoriquement, il est suffisant de connaître uniquement la séquence d'une macromolécule pour aboutir à sa structure [2]. Mais le phénomène de repliement de séquence dans l'espace est encore bien peu compris, ce qui rend le problème de prédiction très complexe. Il est nécessaire alors, dans certains cas, d'avoir des informations additionnelles telles que les variations possibles des configurations moléculaires tridimensionnelles locales ou globales. Ainsi, à cet égard beaucoup d'efforts sont consacrés à la recherche de méthodes théoriques pour analyser la banque de données des

structures d'acides nucléiques résolues.

L'analyse des structures d'acides nucléiques est non seulement importante pour étudier les propriétés structurales de ces dernières, mais elle joue un rôle dominant pour avancer les théories sur l'évolution moléculaire et la modélisation afin de prédire des structures tridimensionnelles. Un des principes de l'évolution moléculaire est que les biopolymères subissant des mutations génétiques mais conservant la même fonction biologique préservent la même structure. L'analyse des structures pourra être utile pour comprendre le lien entre la structure d'un biopolymère et la fonction biologique qui le caractérise. Aussi, l'identification de certaines conformations structurales semblables dans des biopolymères différents pourra mener à déterminer leur fonction biologique. Les systèmes de modélisation génèrent un nombre énorme de conformations possibles pour une molécule donnée. Pour procéder à une classification et à un raffinement des modèles trouvés, les méthodes d'analyse de structures deviennent de plus en plus importantes pour de tels besoins.

Les principales caractéristiques structurales des acides nucléiques sont les appariements et les empilements des bases azotées. Nous présenterons un outil pour identifier l'empilement et les appariements dans une structure d'acide nucléique d'une manière automatique. Cela va permettre de générer une banque de données regroupant toute l'information pertinente sans avoir besoin de visualiser systématiquement une structure tridimensionnelle. La visualisation s'avérant une procédure ardue, très lente et non systématique, aussi les avis peuvent différer d'un visualisateur à un autre.

Chapitre 1

Les acides nucléiques

1.1 La nature des acides nucléiques

Il est approprié de commencer ce chapitre avec une description des acides nucléiques. Ils sont des constituants essentiels à toute cellule vivante. Ces remarquables molécules sont la mémoire de la vie. En effet, elles conservent en elles-mêmes et transmettent toute information nécessaire au développement de tout organisme, à l'échelle intracellulaire comme à l'échelle de l'organisme complet. Que ce soit la nature d'une protéine, la fonction qu'elle occupe, le moment où elle sera synthétisée, le temps de la dégrader, tout est enregistré sur ce ruban moléculaire. L'avènement d'un changement à une partie de la mémoire est fondamental dans le processus d'évolution des formes vivantes. La nature des acides nucléiques et leur capacité à se dupliquer justifie leur fonction.

Nous en apprenons de plus en plus sur les structures et les fonctions des acides nucléiques, par le biais de nouvelles techniques microscopiques permettant la visualisation à l'échelle atomique. De plus, la manipulation de la mémoire génétique par les techniques de biologie moléculaire nous a donné depuis déjà plus d'une décennie, la capacité de diriger le développement futur de la vie. En agriculture, notamment, les exemples sont nombreux: plant à croissance plus rapide, résistance améliorée aux

insectes parasites, etc.

Au début de ce chapitre, nous introduirons brièvement comment les acides nucléiques préservent et transmettent les informations génétiques.

1.1.1 Les deux types d'acide nucléique : ADN et ARN

Il y a deux types d'acide nucléique, *acide ribonucléique (***ARN***)* et *acide désoxyribonucléique* (**ADN**). Chacun est un polymère constitué d'unités *monomères* liées par des liaisons covalentes. Les unités monomères sont montrées à la figure 1.1:

Dans chaque cas, un monomère contient un sucre à cinq carbones (ribose dans l'ARN, 2'-désoxyribose dans l'ADN) liant un groupement phosphate à son cinquième carbone (**C5'**). Les résidus de sucres et de phosphates forment le squelette d'une molécule d'acide nucléique, parfois contenant plusieurs milliards d'unités construites l'une sur l'autre. Par lui-même, le squelette est une structure en répétition, incapable d'encoder de l'information. L'importance des acides nucléiques dans l'emmagasinement de l'information et la transmission provient du fait qu'ils sont hétéropolymères. Chaque monomère dans la chaîne porte un groupement basique appelé *base azotée*, toujours attaché au carbone 1' (**C1'**) du sucre (voir figure 1.1). Il y a deux types de ces substances basiques, les *purines* et les *pyrimidines*.

La structure des bases azotées majeures trouvées dans les acides nucléiques est représentée à la figure 1.2. Notez que l'ADN admet les purines *adénine* (**A**) et *guanine* (**G**) et les pyrimidines *cytosine* (**C**) et *thymine* (**T**). Dans l'ARN, les bases sont les mêmes excepté que l'*uracile* (**U**) remplace la thymine. Ainsi, l'ADN et l'ARN contiennent chacun quatre bases. Chacun peut être vu comme un polymère formé de quatre monomères. Les monomères sont des molécules de riboses ou de désoxyriboses avec les bases purines ou pyrimidines attachées à leur C1'. Dans les purines, l'at-

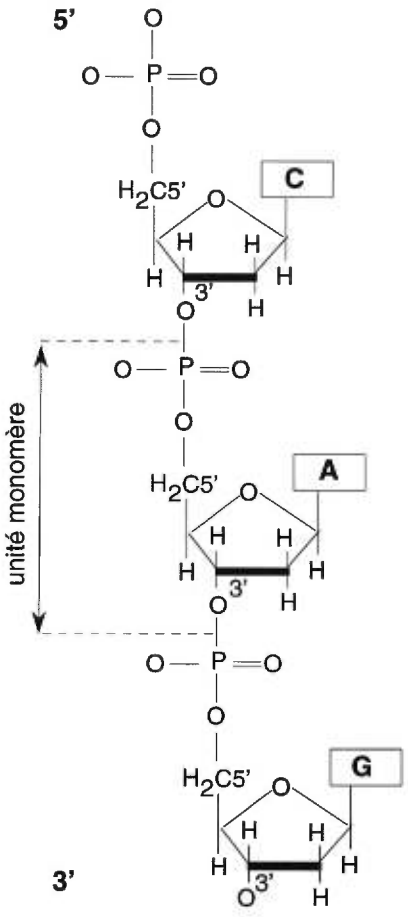


Figure 1.1: La figure montre l'unité monomère répétée de l'ADN. Les phosphates qui lient entre eux les désoxyriboses (ou ribose dans le cas de l'ARN) de la chaîne principale relient le carbone 3' d'un sucre au carbone 5' du suivant. Ceci confère une *polarité* à la chaîne et force un sens de lecture à la machinerie chargée de décoder l'information. On lit alors la séquence présentée: 5' C-A-G 3'.

tachement se fait au niveau de l'azote 9 **(N9)** , dans les pyrimidines au niveau du N1. Ces monomères sont appelés des *nucléotides.* Ainsi, tous les acides nucléiques peuvent être vus comme des polymères de nucléotides, ils sont aussi appelés par le terme générique *polynucléotides.*

L' examen de la figure 1.1 révélera deux importantes caractéristiques de tous les polynucléotides :

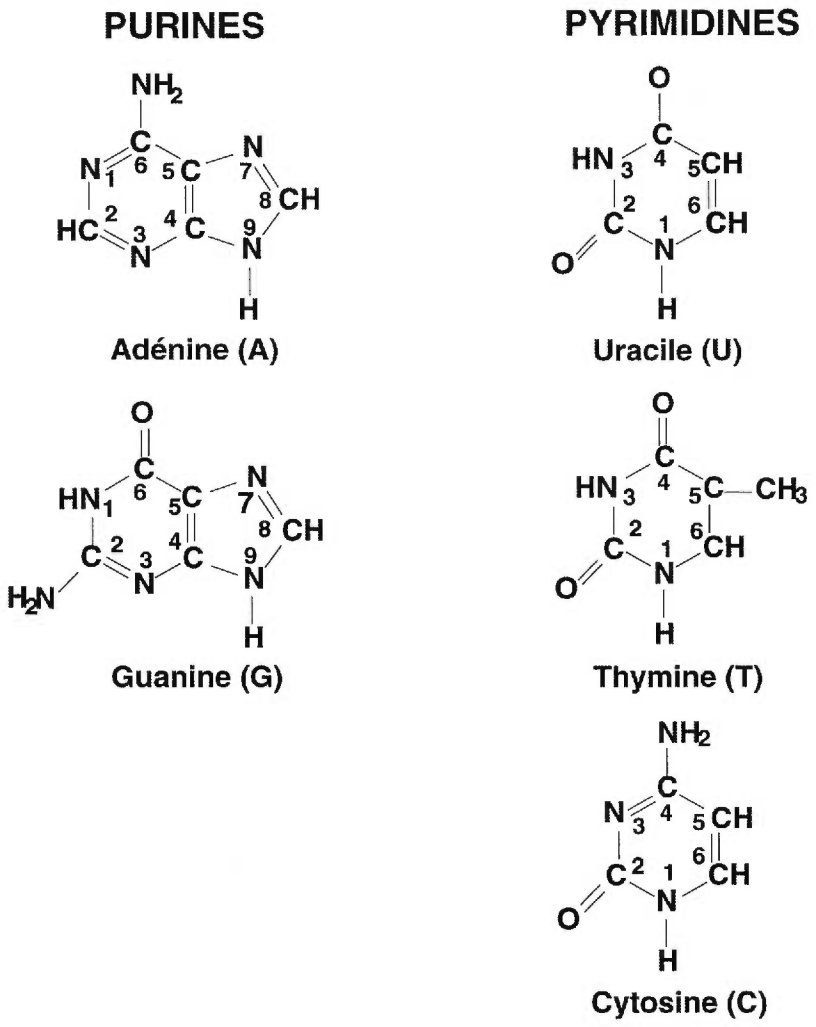


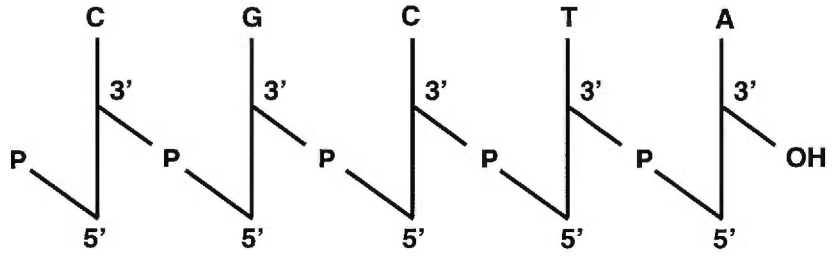
Figure 1.2: Les structures chimiques des bases azotées principales dans les acides nucléiques. L'atome N9 des purines et l'atome N1 des pyrimidines sont liés au carbone C1' du ribose ou du désoxyribose dans les nucléotides.

1. la chaîne polynucléotidique a une direction; la liaison phosphodiester entre les monomères s'établit toujours entre le groupement phosphate en C5' du premier sucre et le groupement OH en C3' du sucre suivant. Ainsi, les deux extrémités de la chaîne polynucléotidique linéaire sont identifiées.

2. la chaîne polynucléotidique est unique par la séquence de ses bases, ie. la séquence des nucléotides. Cette séquence est appelée *structure primaire* de

cet acide nucléique particulier. C'est dans la structure primaire de l'ADN que l'information génétique est emmagasinée. Un *gène* n'est rien d'autre qu'une séquence d'ADN particulière, encodant l'information dans un langage de quatre lettres dans lequel chaque lettre représente une des bases azotées.

Si on veut décrire une chaîne polynucléotidique en particulier (ADN ou ARN), c'est extrêmement difficile de dessiner la molécule comme dans la figure 1.1. Pour cela, d'autres nomenclatures plus compactes ont été adoptées. Nous pouvons faire l'abréviation d'une petite molécule d'ADN comme suit :



Cette notation montre :

1. la séquence des nucléotides est représentée par les lettres (C, G, C, T, A);
2. toutes les liaisons phosphodiesters sont $3' \rightarrow 5'$.

Toutes les liaisons phosphodiesters étant systématiquement $3' \rightarrow 5'$, une notation plus compacte est possible pour la même molécule:

$$pCpGpCpTpA.$$

Finalement, si on est concerné seulement par la séquence des bases dans la molécule, comme c'est généralement le cas, nous pouvons la noter plus compactement comme:

$$CGCTA,$$

remarquez que les séquences sont toujours écrites, par convention, avec l'extrémité 5' à la gauche et l'extrémité 3' à la droite.

Une des plus importantes découvertes dans l'histoire de la biologie moléculaire est la découverte de la structure de l'ADN en 1953. On commençait à s'intéresser aux acides nucléiques depuis déjà quelques années, quand Watson et Crick proposent un modèle de la structure tridimensionnelle de l'ADN, modèle qui donne un support chimique réel aux hypothèses génétiques. De ce tremplin, la génétique a pris son envol; et la biologie moléculaire est née de ces événements.

Watson et Crick ont d'abord observé que l'ADN formait une hélice et que celle-ci était constituée de deux chaînes (*brins*) d'ADN (double hélice). Ces deux brins sont stabilisés par des liaisons hydrogènes (*ponts hydrogènes* ou *appariements*) établies entre les bases azotées: le A avec le T et le C avec le G. Les paires de bases (**pb**) sont empilées les unes sur les autres, leurs plans sont presque perpendiculaires à l'axe de l'hélice. L'empilement comme montré dans la figure 1.3, permet une sorte d'interaction chimique (interaction de van der Waals) entre les bases. Le squelette phosphates-désoxyriboses hydrophile occupe les côtés extérieurs de l'hélice prenant contact avec l'environnement aqueux. Chaque paire de bases fait une rotation de 36° par rapport à la suivante pour organiser 10 pb dans chaque tour d'hélice.

En développant leur modèle moléculaire de la structure de l'ADN en double brins, Watson et Crick ont été convaincus que les deux brins d'une hélice doivent se replier dans des directions opposées. Le modèle de Watson-Crick n'a pas seulement expliqué la structure de l'ADN. Il a eu des implications qui sont au coeur de la biologie. La base A est toujours pairée avec la base T, et la base G est toujours pairée avec la base C; ainsi les deux brins sont complémentaires. Les brins peuvent être séparés et un nouvel ADN est synthétisé le long de chacun d'eux, suivant la même règle de complémentarité, et deux copies exactes des doubles brins d'ADN seront obtenus. Ceci est précisement la propriété que le matériel génétique doit avoir: lorsque la

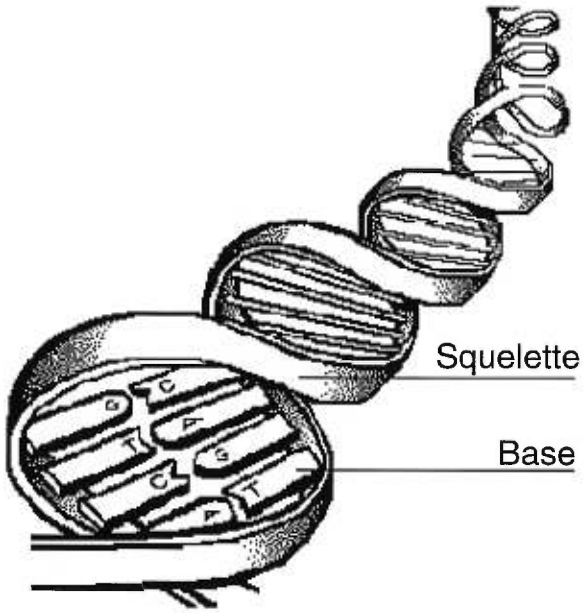


Figure 1.3: La structure de l'ADN est illustrée par la double hélice. Chaque brin spiral, comprenant le squelette phosphates-désoxyriboses et les bases qui lui sont attachées, est lié à un brin complémentaire par des liaisons covalentes entre les paires de bases, adénine (A) avec thymine (T) et guanine (G) avec cytosine (C). Cette structure a été décrite la première fois par Watson et Crick en 1953.

cellule se divise, deux copies complètes de l'information génétique portée dans la cellule originale doivent être produites (voir figure 1.4).

Les molécules d'ADN se présentent sous un grand nombre de formes ou configurations topologiques. Elles peuvent être repliées ou non sur elles-mêmes ou former un cercle, forme *relaxée*.

Les molécules d'ARN trouvées dans la nature sont en général plus petites que les molécules d'ADN et dans la plupart des cas, elles existent sous forme d'un simple brin. Malgré d'être produites en simple brin, elles peuvent se replier en région double brins, dûe à la complémentarité intrinsèque. L'existence de ces régions donne aux molécules d'ARN des structures 3D complexes comme montré dans la figure 1.5.

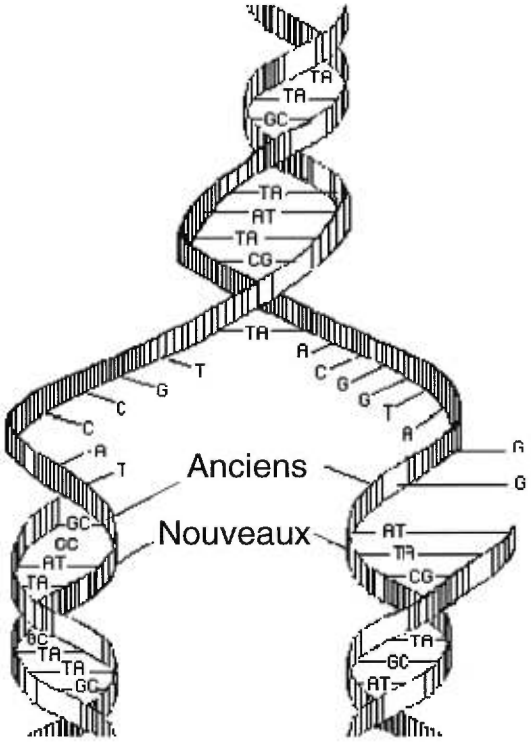


Figure 1.4: Une représentation simplifiée de la séparation de la molécule d'ADN pour former deux copies identiques. Les deux brins de la double hélice sont séparées par des enzymes.

La molécule représentée dans la figure 1.5 est un membre d'une classe parmi les plus importantes des acides ribonucléiques, les *ARNs de transfert*, ou *ARNts*. Ces derniers jouent un rôle essentiel dans l'assemblage des protéines en apportant les acides aminés dans le ribosome pendant la synthèse des protéines.

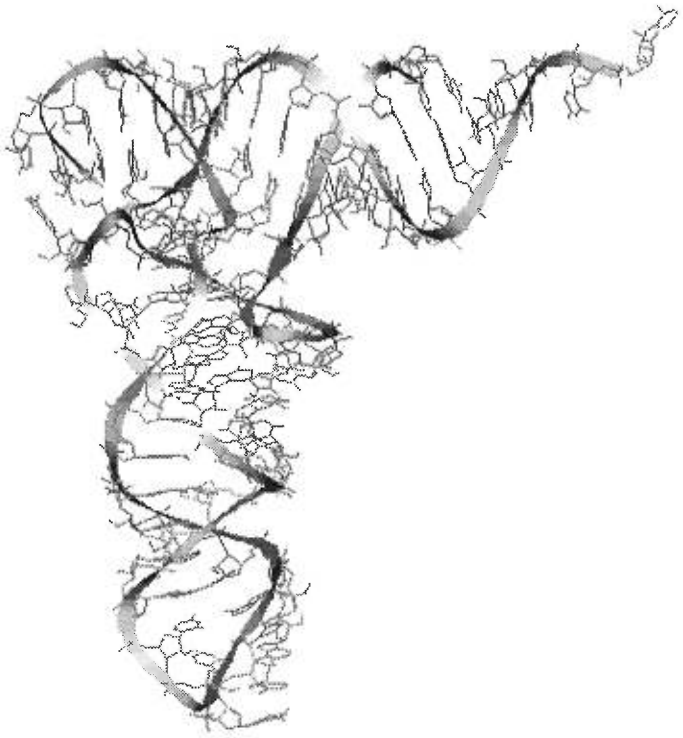


Figure 1.5: Structure tridimensionnelle l'ARN de transfert du Phélylalanine chez la levure (ARNt^{Phe}). Le ruban montre les courbures subies par le squelette phosphates-riboses de la molécule.

1.2 Analyse des structures tridimensionnelles des acides nucléiques

Nous assistons au développement de plusieurs méthodes mathématiques pour analyser les structures tridimensionnelles des acides nucléiques et incorporer les caractéristiques de leurs séquences dans les simulations par ordinateurs [9, 3, 28, 14, 10]. Pour mieux s'attaquer au sujet, il est nécessaire de comprendre la nature des changements locaux et globaux des structure impliquées dans la fonction biologique de ces molécules. Ils se manifestent des mouvements subtiles (torsion, translation et courbure) au niveau des paires de bases. Deux interactions principales caractérisent la structure tridimensionnelle des acides nucléiques: les appariements et l'empilement

des bases azotées. La capacité des nucléotides à coder de l'information découle de leur capacité à former des paires de bases complémentaires de façon stable et spécifique. Cette stabilité et cette spécificité sont inhérentes à la géométrie des bases ainsi qu'à la possibilité de former des appariements, c'est-à-dire des *liaisons hydrogènes.* Ces liaisons s'établissent entre les atomes capables de partager un proton (un atome d'hydrogène). En ce qui concerne la deuxième caractéristique, l'empilement des bases azotées décrit largement les structures tridimensionnelles de l'ADN et de l'ARN. L'empilement entre deux bases est une sorte d'interactions verticales entre elles. Ces interactions sont très importantes dans la stabilité des hélices présentes dans les structures d'acides nucléiques. Dans ce travail, on étudiera l'empilement des bases azotées d'une manière géométrique. L'approche mathématique de la détection des empilements d'une manière automatique en détail sera traitée dans le deuxième chapitre. De plus, on analysera les empilements se trouvant dans deux molécules: l'ARNt^{Phe} [38] et le ribosyme "*hammerhead*" [36].

L'empilement de bases a été étudié dans différents systèmes, variant des bases isolées aux longs duplexes d'ADN. Plusieurs études expérimentales ont été développées et divers résultats ont été obtenus pour des systèmes simples formés seulement de deux nucléotides [11]. Les calculs semi-empiriques, les mécaniques moléculaires, les simulations des dynamiques moléculaires et les méthodes de perturbations de l'énergie libre ont été utilisés pour comprendre l'empilement des bases [1, 17, 23, 31, 6]. La plupart de ces études théoriques [1, 17, 23, 6] se sont intéressées aux interactions base-base en ignorant la dynamique du squelette ou l'influence des interactions du solvant. Le mécanisme de l'empilement des bases n'est pas encore complètement compris, et différentes forces comme les effets hydrophobiques ou les interactions électrostatiques directes existant entre les bases ont été suggérées comme étant l'élément dominant

contribuant à la stabilisation des bases empilées [30].

L'empilement des bases est un facteur primaire pour stabiliser la structure des acides nucléiques. Jusqu'à maintenant, les méthodes pour localiser les interactions d'empilement dans l'ADN et l'ARN sont encore rares. La méthode développée dans ce travail est une méthode géométrique localisant l'empilement dans une structure d'acide nucléique avec un poids symbolisant l'importance de l'interaction. Les avantages de la méthode (une heuristique pour analyse rapide) peuvent se résumer par deux principales raisons: les forces impliquées dans la stabilisation des bases empilées sont jusqu'à maintenant inconnues, de même les paramètres des effets hydrophobiques ne sont pas encore bien définis pour les fonctions connues d'énergie potentielles. Ainsi, la méthode ne prend pas en considération les fonctions d'énergie empiriques ni les données tirées de certains paramètres associés à des empilements localisés dans des structures d'acides nucléiques. Elle ne favorise pas non plus la position relative des bases dans la détermination des empilements présents au sein d'une structure d'acide nucléique.

Ultérieurement aux études de Watson et Crick, des études cristallographiques plus précises ont confirmé leur modèle [57]. Elles ont aussi montré que l'ADN est très flexible et capable de subir des déviations significatives à partir de cette structure canonique [41]. On a montré aussi que les acides nucléiques adoptent des conformations qui ne sont pas similaires au modèle original de Watson et Crick, comme mentionné dans [55]. Dans les doubles hélices, les liens hydrogènes entre les paires de bases servent à maintenir les deux brins ensemble. En général, les différents motifs de liens hydrogènes restent constants dans les hélices régulières des acides nucléiques et respectent un appariement standard appelé *appariement Watson-Crick* (voir la figure 1.7 pour un exemple d'appariement Watson-Crick, ce dernier est caractérisé par

les liens d'hydrogène partagés entre les atomes donneurs et accepteurs de chacune des bases, voir la table 1.1). De plus, les petites fluctuations des six angles entre les atomes du squelette phosphate-sucre local (i.e. $\alpha, \beta, \gamma, \delta, \epsilon$ et ζ sont appelés les angles *déhydraux* ou *de torsion*, voir la figure 1.6) peuvent facilement permettent les changements dans la géométrie des paires de bases [45].

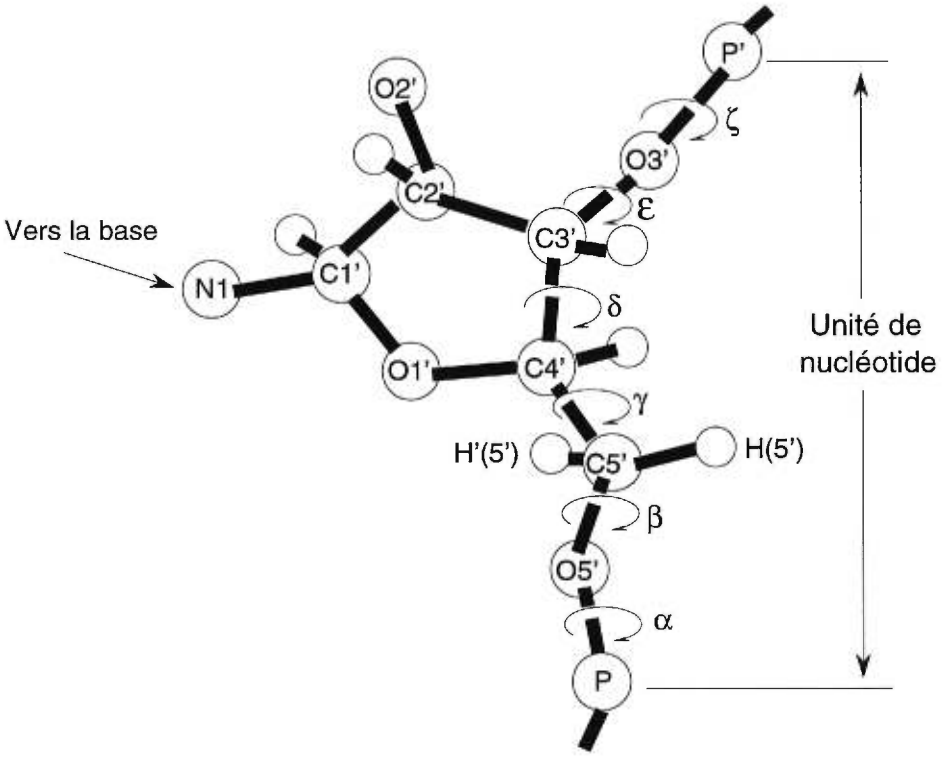


Figure 1.6: La conformation d'un nucléotide est déterminée par les six angles de torsion indiqués, $\alpha, \beta, \gamma, \delta, \epsilon$ et ζ.

Puisque les paires de bases seules ont une tendance naturelle à former des hélices en solution [48, 46], ceci montre que les résidus de sucres et phosphates jouent un rôle mineur dans la structure des acides nucléiques; alors que les interactions hydrophobiques entre les bases empilées jouent un rôle dominant dans la formation de cette structure hélicoïdale particulière. Malgré l'importance de l'empilement des bases, il

Atome\ Base	**Adénine**	**Cytosine**	**Guanine**	**Thymine**	**Uracile**
Donneur	N6	N4	N1, N2	N3	N3
Accepteur	N1, N3, N7	O2, N3	N3, N7, O4	O2, O4	O2, O4

Table 1.1: Les atomes donneurs et accepteurs des bases adénine, cytosine, guanine, thymine et uracile. On remarque que la thymine et l'uracile ont le même atome donneur et les mêmes atomes accepteurs.

reste que les méthodes pour identifier les bases empilées d'une manière automatique pour une structure moléculaire ne sont pas encore nombreuses [13, 32]. En ce qui concerne l'empilement des bases, la visualisation est un moyen d'identification laborieux et subjectif. Il est donc nécessaire d'analyser d'une manière automatique les modèles dans l'espace pour pouvoir étudier l'empilement entre deux bases quelconques. L'empilement peut être quantifié en utilisant une fonction d'énergie, une approche qui a été écartée en s'appuyant sur ce qui suit. L'empilement est un phénomène dû à deux types d'interactions, hydrophobiques et électrostatiques. Cependant les forces dénergie potentielle impliquées dans ces interactions ne sont pas encore comprises pour permettre une évaluation fiable dans des cas d'empilement (consulter [44]).

Pour les raisons mentionnées ci-dessus, on a développé une approche pour analyser l'empilement de bases en définissant ce dernier comme étant successivement deux interactions entre deux bases planaires dans l'espace. Chaque base exerce alors orthogonalement une interaction sur le plan supportant l'autre base. Cette définition s'avère rigoureuse par rapport à celle utilisée dans d'autres études. Notre approche utilise des critères géométriques pour identifier l'empilement. Ces critères formalisent ce dernier, ils sont facilement implémentables et rigoureux pour une analyse flexible de plusieurs cas d'empilement. Ils offrent aussi la possibilité d'étudier différents paramètres géométriques impliqués dans l'empilement des bases, ignorés jusqu'à maintenant. Notre algorithme a été implémenté en langage de programma-

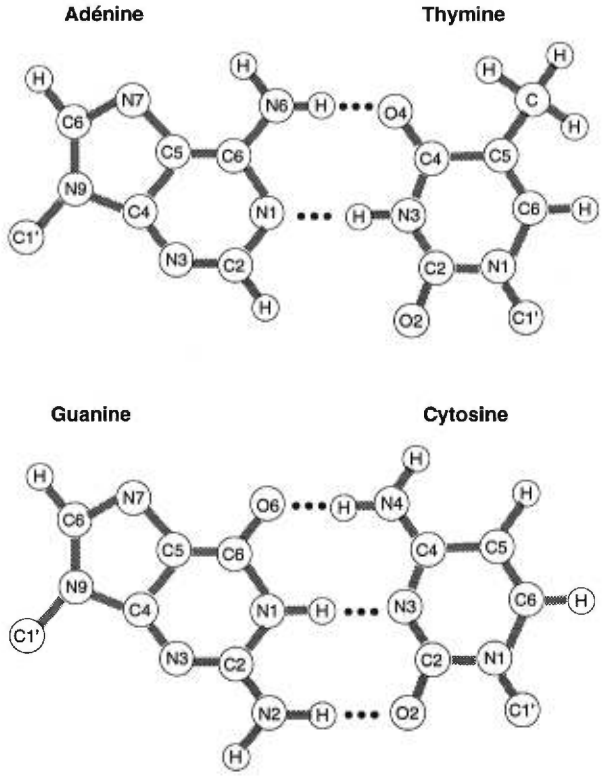


Figure 1.7: Appariement Watson-Crick des paires de bases adénine-thymine et guanine-cytosine. Dans le cas de la paire de base adénine-thymine, les atomes de l'adénine N6 donneur et N1 accepteur forment respectivement des liens hydrogènes avec les atomes de la thymine O4 accepteur et N3 donneur. Pour la paire de base guanine-cytosine, les atomes O6 accepteur, N1 donneur et N2 donneur de la guanine présentent successivement des liens hydrogènes avec les atomes N4 donneur, N3 accepteur et O2 accepteur de la cytosine.

tion C++, à l'aide de la librairie LEDA (Library of Efficient Data and Algorithms: librairie d'objets et fonctions qui s'attachent au langage C++) [29]. Les ordinateurs utilisés pour la compilation du programme sont SGI UNIX (IRIX 5.3). Le format des fichiers de données soumis au programme pour fin d'analyse des empilements qu'ils contiennent, est celui de PDB (Protein Data Bank, pour plus d'information consulter le site suivant http://www.pdb.bnl.gov) qui est le format le plus utilisé internationalement, voir la figure 1.8. L'algorithme est intégré à un autre algorithme déterminant les liens hydrogènes et la structure secondaire d'une macromolécule donnée, le tout

sera intégré au programme NANOTE [33] pour l'annotation de structure d'ADN-ARN. Dans ce travail [33], on présente alors le processus général d'une annotation automatique des structures tridimensionnelles de l'ADN et l'ARN. Tout d'abord, le programme obtient un fichier en format PDB décrivant la molécule, les atomes d'hydrogène seront alors ajoutés à ce fichier en se conformant aux exigences du format PDB et les bases azotées *modifiées* [1] seront remplacées par leurs bases azotées standards (déjà montrées dans la figure 1.2).

On fera en sorte que tous les paramètres exigés par NANOTE soient présents sinon un message d'avertissement sera affiché pour chaque élément manquant. Ainsi, on procédera à la détermination des appariements et empilements de bases (l'approche pour l'empilement de bases est présenté en détail dans le chapitre subséquent). Les structures en doubles hélices seront aussi identifiées selon un format PDB hybride. L'appariement et l'empilement sont montrés par deux nouvelles notations pour le format PDB. La figure 1.9 présente le genre de fichier de sortie montrant l'information complète ainsi obtenue. L'utilisateur pourra choisir selon ce qu'il veut modéliser entre la description des séquences de résidus ou les appariements ou l'empilement ou les doubles hélices.

Enfin cela servira comme outil important à l'ensemble du programme de modélisation des structures moléculaires MC-SYM [26]. MC-SYM est en soit un système utilisant la programmation symbolique et des techniques de satisfaction de contraintes. De cette manière, il est capable de générer, à partir d'une séquence polymérique, un ensemble de structures tridimensionnelles satisfaisant les contraintes chimiques tirées des expériences de laboratoire ou des hypothèses imposées par l'utilisateur.

[1] Bases azotées de types purines ou pyrimidines différentes des cinq bases azotées A, C, G, T, U et rencontrées occasionnellement dans les molécules d'ARN et d'ADN.

```
ATOM     12  C1'   A A   1      6.540   5.120  -1.419  1.00  0.00
ATOM     13  N9    A A   1      5.317   4.299  -1.193  1.00  0.00
ATOM     14  C8    A A   1      4.010   4.678  -1.299  1.00  0.00
ATOM     15  N7    A A   1      3.166   3.729  -1.036  1.00  0.00
ATOM     16  C5    A A   1      3.972   2.639  -0.733  1.00  0.00
ATOM     17  C6    A A   1      3.677   1.316  -0.366  1.00  0.00
ATOM     18  N6    A A   1      2.428   0.845  -0.236  1.00  0.00
ATOM     19  N1    A A   1      4.715   0.491  -0.136  1.00  0.00
ATOM     20  C2    A A   1      5.953   0.965  -0.267  1.00  0.00
ATOM     21  N3    A A   1      6.349   2.173  -0.602  1.00  0.00
ATOM     22  C4    A A   1      5.290   2.979  -0.826  1.00  0.00
ATOM     31  H8    A A   1      3.704   5.686  -1.579  1.00  0.00
ATOM     32 1H6    A A   1      2.273  -0.136   0.036  1.00  0.00
ATOM     33 2H6    A A   1      1.623   1.465  -0.408  1.00  0.00
ATOM     34  H2    A A   1      6.751   0.250  -0.068  1.00  0.00
ATOM     46  C1'   C A  10     -1.947   4.860  -1.350  1.00  0.00
ATOM     47  N1    C A  10     -2.324   3.471  -0.965  1.00  0.00
ATOM     48  C2    C A  10     -1.298   2.589  -0.720  1.00  0.00
ATOM     49  O2    C A  10     -0.123   2.900  -0.886  1.00  0.00
ATOM     50  N3    C A  10     -1.686   1.313  -0.366  1.00  0.00
ATOM     51  C4    C A  10     -2.981   0.855  -0.239  1.00  0.00
ATOM     52  N4    C A  10     -3.170  -0.408   0.112  1.00  0.00
ATOM     53  C5    C A  10     -3.988   1.853  -0.516  1.00  0.00
ATOM     54  C6    C A  10     -3.639   3.103  -0.863  1.00  0.00
ATOM     63 1H4    C A  10     -4.126  -0.776   0.214  1.00  0.00
ATOM     64 2H4    C A  10     -2.361  -1.022   0.283  1.00  0.00
ATOM     65  H5    C A  10     -5.043   1.589  -0.443  1.00  0.00
ATOM     66  H6    C A  10     -4.417   3.838  -1.067  1.00  0.00
TER
END
```

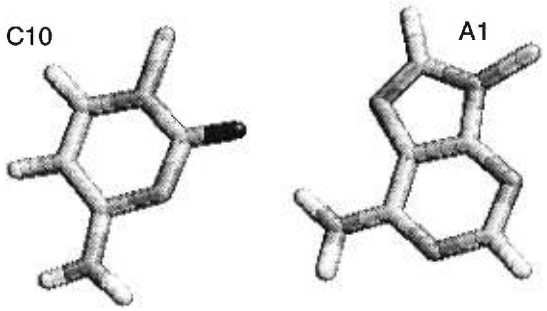


Figure 1.8: Exemple de fichier PDB, AC.pdb, d'une molécule formée par les deux bases A et C ainsi que leur représentation tridimensionnelle.

L'annotation automatique des structures tridimensionnelles servira dans le cas de MC-SYM à l'identification des sites importants des appariements et des empilements. Ces informations seront utilisées dans la modélisation pour prédire la structure tridimensionnelle d'une moléculaire donnée. NANOTE indiquera la voie à suivre dans la construction d'un certain modèle, en se basant sur des critères et propriétés qui aideront à une classification des empilements et appariements. Cette classification limitera le grand nombre de cas plausibles à explorer dans les arbres de recherche,

ce qui n'était pas faisable avant, même en présence de plusieurs contraintes. Chaque base azotée peut être paramétrisée par une matrice "*conformation*" définissant un repère local propre à la base où chaque atome de la base est déterminé par ses coordonnées tridimensionnelles dans ce dernier repère. Dans le cas d'empilement et d'appariement entre deux bases, on peut situer une base par rapport à l'autre par une matrice de rotation "*transformation*". Cette matrice est définie simplement par la matrice de passage entre les repères des bases et par la translation entre les origines des repères. Ainsi la classification se fait à l'aide d'une distance entre les matrices de conformations et de transformations.

Le programme NANOTE est un programme d'annotation automatique qui aidera à une modélisation mieux dirigée par une analyse structurale bien formalisée. Aussi cela permettra en découvrant des motifs structuraux d'établir des propriétés et des liens entre la structure d'une molécule et la fonction biologique qui la caractérise dans la nature.

```
01234567890123456789012345678901234567890123456789012345678901234567890
HYDBND N4  C     72   2H4 72   O6  G     1
HYDBND N6  A     5    1H6 5    O4  U     68

0  - 5   HYDBND
7  - 9   ATOM NAME (DONOR)
11 - 13  RES NAME (DONOR)
15 - 15  CHAIN ID (DONOR)
17 - 20  RES NUMBER (DONOR)
22 - 24  HYDRO NAME
26 - 29  RES NUMBER (HYDRO)
31 - 33  ATOM NAME (ACCEPTOR)
35 - 37  RES NAME  (ACCEPTOR)
39 - 39  CHAIN ID (ACCEPTOR)
41 - 44  RE NUMBER (ACCEPTOR)

01234567890123456789012345678901234567890123456789012345678901234567890
DBLHLX 1   H0  2  G     1    U     7    0
DBLHLX 2   H3  2  C     61   G     65   -1

0  - 5   DBLHLX
7  - 9   HELIX NUMBER
11 - 13  HELIX NAME
15 - 16  NUMBER OF HELIX
18 - 20  RES NAME (BEGINNING OF  HELIX)
22 - 22  CHAIN ID (RES BEGINNIG)
24 - 27  RES NUMBER
29 - 31  RES NAME  (END OF HELIX)
33 - 33  CHAIN ID (RES END)
35 - 38  RES NUMBER (RES END)
40 - 41  HELIX WAY (SENS)

01234567890123456789012345678901234567890123456789012345678901234567890
SEQRES  1    76     G   C   G   G   A   U   U   U   A   G   C   U   C
SEQRES  2    76     A   G   U   U   G   G   G   A   G   A   G   C   G

0  - 5   SEQRES
8  - 9   NUMBER OF LINES
11 - 11  CHAIN ID
13 - 16  NUMBER OF RESIDUS
20 - 22  RES NAME (FIRST OF THE LINE)
24 - 26  RES NAME
...      RES NAME
68 -70   RES NAME (LAST OF THE LINE)

01234567890123456789012345678901234567890123456789012345678901234567890
STACK G   C     71     72   ANTI ANTI
STACK C   A     72     73   ANTI SYN

0  - 4   STACK
6  - 8   RES NAME (FIRST RES)
10 - 12  RES NAME (SECOND RES)
14 - 14  CHAIN ID (FIRST RES)
16 - 29  RES NUMBER (FIRST RES)
31 - 31  CHAIN ID (SECOND RES)
33 - 36  RES NUMBER (SECOND RES)
38 - 41  (ANTI - SYN   FIRST RES)
43 - 46  (ANTY - SYN  SECOND RES)
```

Figure 1.9: Les nouveaux enregistrements PDB décrivant les appariements, les empilements, les doubles hélices et la séquence des résidus pour une molécule

Chapitre 2

Étude et analyse de l'empilement des bases azotées

2.1 Introduction

Trouver la structure tridimensionnelle d'une molécule est actuellement parmi les problèmes intrigants et fondamentaux en biochimie. Malgré la connaissance de structures finales moyennes et de la connectivité covalente pour les chaînes peptidiques et phosphodiestères, les détails du processus de repliement ne sont pas connus. Lorsque la structure de la double hélice a été proposée pour la première fois, il y avait une grande excitation, non seulement par le grand succès accompagnant la résolution du problème mais aussi pour la simplicité de la structure apparente. Une de ses principales particularités était d'être caractérisée par un nombre relativement petit de paramètres. Une autre grande excitation était attendue par la première solution de la protéine myoglobine. Cependant, ceci n'était pas accompagné par une joie concernant la simplicité de la structure mais par l'étonnement de la complexité de la structure. Dans les années suivantes, beaucoup de temps et d'efforts ont été dépensés pour essayer de comprendre cette complexité.

Dans cet esprit de vouloir donner des descriptions utiles, on tentera de parler de l'empilement entre deux bases dans les structures d'acide nucléique. Pour la première

fois, on analysera à l'aide de la géométrie Euclidienne l'empilement des bases deux à deux, ce qui aidera à la modélisation et à l'analyse de ces structures. Chaque empilement sera quantifié d'une manière efficace pour distinguer les différences entre tous les empilements qui se trouvent au sein d'une structure. Les données obtenues refléteront d'une manière paramétrique l'aspect de l'empilement qui a été jusqu'à maintenant analysé visuellement. Cette analyse visuelle ne permet pas de distinguer les différences entre les empilements présents dans une structure. La paramétrisation de l'empilement des bases sera définie en utilisant le principe de la surface moléculaire de van der Waals.

2.1.1 Surfaces moléculaires

Richards [39] a introduit le terme de surface moléculaire décrivant une enveloppe moléculaire accessible à une molécule du solvant. Cette surface peut être définie de plusieurs manières: surface de van der Waals, surface de Connolly et surface accessible au solvant (voir figure 2.1).

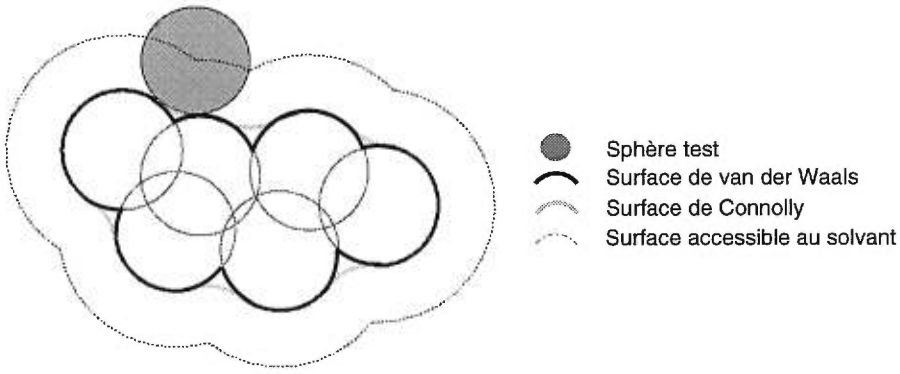


Figure 2.1: Les trois différentes définitions de surface moléculaire.

La surface de van der Waals

La surface d'une molécule n'a pas qu'une seule forme définie. Sa forme apparente dépend de l'approche technique utilisée pour la définir. Cela est dû à la distribution radiale diffuse de la densité d'électrons autour de chaque noyau atomique. Dans le cas des atomes liés chimiquement, cette distribution électronique et les autres propriétés de ces atomes isotropiques ne sont pas sphériquement symétriques. Chaque atome admettra alors une certaine sphère définie par la région de la densité électronique, le rayon de telles sphères est appelé *rayon de van der Waals.*

Les rayons de van der Waals sont utilisés pour le calcul de l'énergie potentielle d'une structure moléculaire. Pour cela, on fait varier plusieurs paramètres rotationnels dans des intervalles et on calcule les distances interatomiques r_{ij} entre les paires d'atomes i, j. Initialement, les calculs sont basés en supposant qu'il n'y a aucune attraction si $r_{ij} > r_0$ où r_0 est la somme des rayons de van der Waals des atomes i et j [41]. Le choix du rayon de van der Waals étant approximatif, chaque auteur possède sa liste favorite pour différents atomes (regarder par exemple dans [40, 8]). Les valeurs qu'on lui associe dans les prédictions structurales peuvent varier suivant la nature du problème. Les sphères associées à deux atomes liés d'une manière covalente se touchent; elles sont tronquées par un plan perpendiculaire au lien atomique et se trouvent dans la région d'intersection. La position exacte de ce plan est encore une fois approximative mais elle est normalement choisie pour couper le lien en deux segments proportionnels au rayon covalent de chaque atome impliqué. La surface complexe résultant de l'ensemble des sphères liées de cette façon est appelée *surface de van der Waals.* Cette surface a une aire et un volume définis. De plus, la construction est facile à visualiser et logiquement cohérente; aucune procédure chimique jusqu'à maintenant n'a mesuré directement cette aire ni ce volume.

Dans plusieurs approximations spécifiques, la considération de la plupart des atomes d'hydrogène est exclue. Par exemple dans les acides nucléiques, les lourds atomes C, N et O sont élargis en une série de groupes avec zéro, un, deux ou trois atomes d'hydrogène attachés de façon convenable. Chacun de ces groupes chimiques est considéré sphériquement symétrique. Le rayon associé à chaque groupe, ayant comme centre le plus lourd atome, est déterminé en tenant compte de la contribution des atomes d'hydrogène présents. Le groupe est identifié au carbone ou à l'azote sans considérer le nombre d'hydrogènes attachés. La figure 2.2 montre la surface de van der Waals d'une molécule hypothétique.

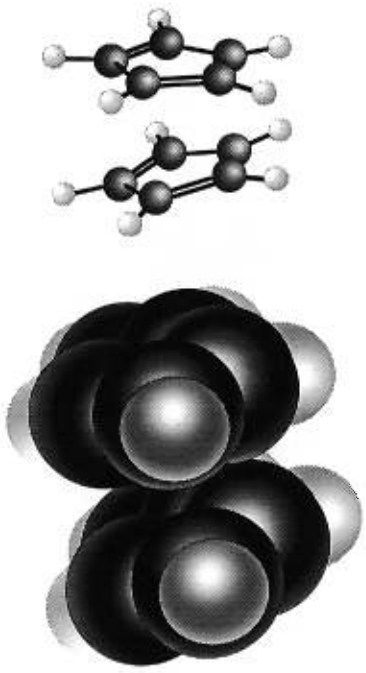


Figure 2.2: Surface de van der Waals d'une molécule hypothétique. La trace de l'enveloppe de van der Waals des atomes de la structure est montrée.

Surface de Connolly

La surface de Connolly est composée de deux sortes d'éléments de surface: premièrement, la surface de *contact* correspondant à la surface de van der Waals, lorsque la sphère test (de 1.4Å généralement, simulant une molécule d'eau) est en contact avec un seul atome; deuxièmement, la surface *réentrante* correspondant à la surface intérieure de la sphère test lorsqu'elle est en contact avec plus d'un atome.

Surface accessible au solvant

La surface accessible au solvant correspond à l'enveloppe moléculaire englobant le volume décrit par l'espace formé jusqu'au milieu de la sphère test lorsqu'on la roule sur les atomes représentés par des sphères de rayon de van der Waals.

2.1.2 La chimie de l'empilement des bases

L'association horizontale base-base par les liens hydrogènes est observée dans les solvants non-aqueux et dans les états gazeux et cristallins. En plus, à l'état solide les bases sont souvent empilées exclusivement de sorte que le plan d'une base est à une distance de van der Waals, ~3.4Å, et parallèle à celui de la base adjacente. Cet arrangement est dû aux interactions verticales plus qu'aux interactions horizontales. Dans les solutions aqueuses, cette forme d'empilement de bases se manifeste également [41]. Etant donné l'importance de la stabilisation des hélices dans une structure d'acide nucléique, les propriétés géométriques principales de l'empilement seront définies dans ce chapitre.

2.2 Méthode d'empilement des bases

L'empilement des bases est un phénomène dominant dans les structures cristallines. Il est à noter que les types d'empilement de bases trouvés et décrits par les moyens chimiques sont plutôt spécifiques. Les groupes polaires (ou halogènes) **-NH_2**, **=N-**, **=O** d'une base est superposé sur le composé aromatique (composé cyclique) de la base adjacente [41], voir la figure 2.3 . Ce type d'empilement inclut aussi les effets des liens hydrogènes.

Pour définir l'empilement entre deux bases, nous avons considéré, indépendamment du composé aromatique ou du groupe polaire, toute la surface moléculaire de van der Waals de ces bases. Pour établir s'il y a empilement, on déterminera la surface de van der Waals de deux bases et on cherchera à voir s'il y a une intersection entre ces surfaces en projetant l'une par rapport à l'autre et inversement.

Considérons alors deux bases A et B dans une certaine structure moléculaire d'un acide nucléique. La mesure de l'empilement entre A et B se fait en cinq étapes qu'on détaillera ci-dessous. Dans l'esprit de cette nouvelle approche, nous avons essayé de faire un programme basé sur les mathématiques pour analyser les structures d'acide nucléique et pour contrôler tout ce qui concerne l'empilement des bases. Le programme offre à l'utilisateur beaucoup de flexibilité. Une compréhension de la signification physique de notre méthode aidera l'utilisateur à prendre avantage des paramètres variables.

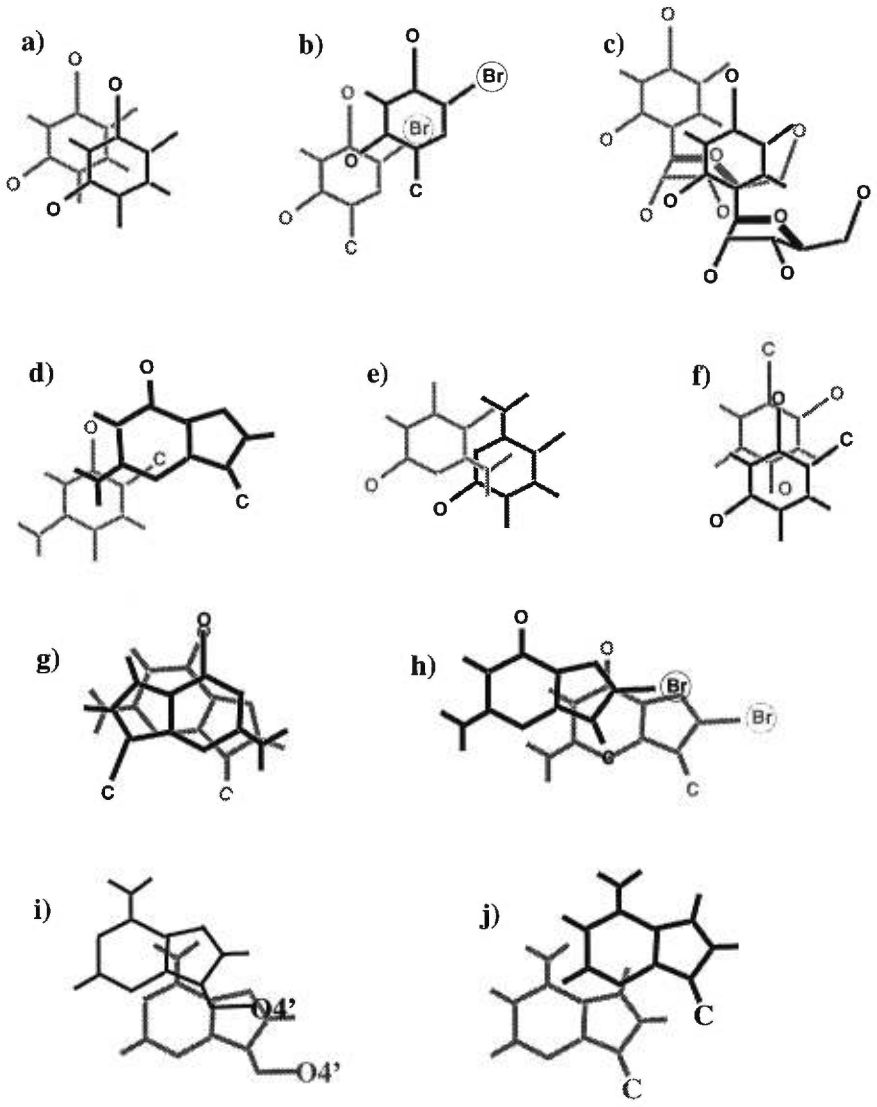


Figure 2.3: Exemples d'empilements de bases observés dans différentes structures cristallines. **a)** Empilement de deux uraciles impliquant l'azote du cycle; **b)** empilement de deux pyrimidines 5-bromouridines; **c)** interaction du ribose avec le composé aromatique, entre deux cytidines; **d)** interaction entre le 9-éthylguanine et le 1-méthyle-5-fluorocytosine; **e)** empilement entre deux cytosines monohydratées; **f)** empilement entre le 9-ethyladénine et le 1-méthyluracile; **g)** empilement entre une guanosine et une inosine; **h)** interaction entre deux 8-bromoguanosines; **i)** interaction entre deux adénosines; **j)** empilement entre deux purines protonées [41].

2.2.1 Paramétrisation des coordonnées des atomes d'une base

La figure 2.4 illustre comment les coordonnées d'état sont déterminées pour chaque base d'une structure expérimentale suivant les axes X, Y et Z. Ces trois axes définissent

un repère appelé *repère absolu* et les coordonnées d'un atome dans ce repère sont appelées *coordonnées absolues.*

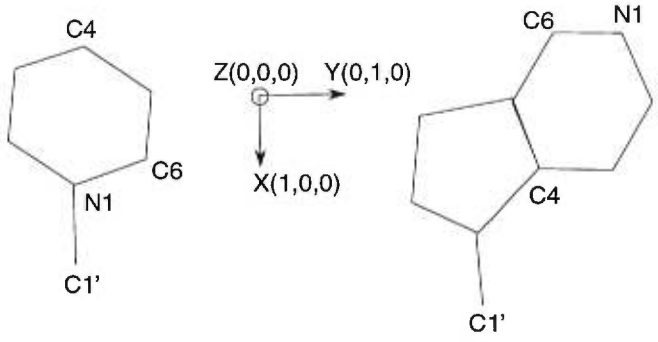


Figure 2.4: Les coordonnées des atomes de chaque base d'une structure expérimentale sont données suivant les axes X, Y et Z.

Pour dériver toutes les équations nécessaires, soit le vecteur colonne:

$$A_i = \begin{pmatrix} A_{ix} \\ A_{iy} \\ A_{iz} \end{pmatrix}$$

qui représente les valeurs des coordonnées de l'atome A dans une base i. En plus, soit:

$$O_i = \begin{pmatrix} O_{ix} \\ O_{iy} \\ O_{iz} \end{pmatrix}$$

le point origine des coordonnées de la base i et soient:

$$x_i = \begin{pmatrix} x_{ix} \\ x_{iy} \\ x_{iz} \end{pmatrix}, \quad y_i = \begin{pmatrix} y_{ix} \\ y_{iy} \\ y_{iz} \end{pmatrix} \quad \text{et} \quad z_i = \begin{pmatrix} z_{ix} \\ z_{iy} \\ z_{iz} \end{pmatrix}$$

les coordonnées absolues des vecteurs unitaires définissant le repère local de la base i. L'origine O_i pour chaque base i coïncide avec son atome N1. Les coordonnées

locales seront déterminées par rapport à un nouveau repère centré en O_i. La relation existante entre les coordonnées locales et les coordonnées absolues est illustrée dans la figure 2.5 pour le cas des coordonnées X et Y. Une relation similaire peut être établie pour les coordonnées Z.

Les coordonnées originales ou absolues d'un atome A peuvent être déterminées à partir des coordonnées A_i mesurées dans la base i en commençant par l'origine O_i et se déplaçant avec la valeur A_{ix} suivant l'axe des X engendré par le vecteur x_i, la valeur A_{iy} suivant l'axe des Y engendré par le vecteur y_i et A_{iy} suivant l'axe des Z engendré par le vecteur z_i. Ceci se manifeste dans les équations suivantes pour calculer A à partir de A_i:

$$\begin{aligned} A_x &= A_{ix}x_{ix} + A_{iy}y_{ix} + A_{iz}z_{ix} + O_{ix}, \\ A_y &= A_{ix}x_{iy} + A_{iy}y_{iy} + A_{iz}z_{iy} + O_{iy}, \\ A_z &= A_{ix}x_{ix} + A_{iy}y_{iz} + A_{iz}z_{iz} + O_{iz}. \end{aligned}$$

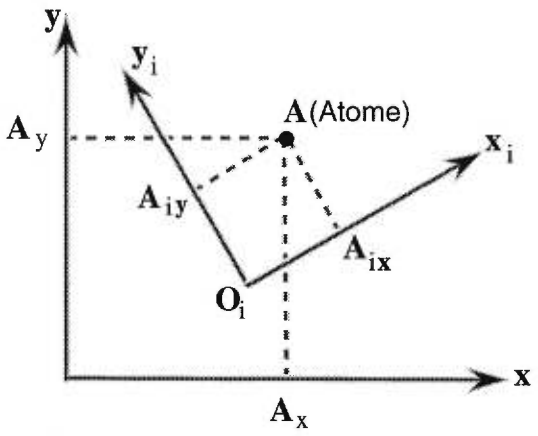


Figure 2.5: La relation entre les coordonnées locales et les coordonnées absolues est illustrée dans le cas des coordonnées X et Y.

Ces trois dernières équations peuvent être écrites vectoriellement comme suit:

$$A = A_{ix}x_i + A_{iy}y_i + A_{ix}z_i + O_i.$$

Aussi d'une manière plus concise, on peut les exprimer suivant la forme matricielle comme:

$$A = R_i A_i + O_i, \tag{2.1}$$

où R_i est la matrice dont les colonnes sont les coordonnées X, Y et Z des vecteurs unitaires du repère dans la base i:

$$R_i = \begin{pmatrix} x_{ix} & y_{ix} & z_{ix} \\ x_{iy} & y_{iy} & z_{iy} \\ x_{iz} & y_{iz} & z_{iz} \end{pmatrix}.$$

L'équation 2.1 nous permet de calculer les coordonnées absolues A d'un atome en fonction des coordonnées A_i associées au repère centré à l'origine O_i de la base i. La matrice R_i est notée pour nous comme une matrice de rotation parce qu'elle définit l'orientation angulaire des vecteurs de la base i. L'origine O_i sera appelée vecteur origine ou vecteur de translation, puisqu'il définit le déplacement de l'origine de la base i par rapport à l'origine des coordonnées absolues. Pour déterminer la position d'une base j par rapport à une base i et vice versa, l'équation 2.1 implique que:

$$R_i A_i + O_i = R_j A_j + O_j,$$

qui peut être écrite comme suit:

$$A_i = R_i^{-1} R_j A_j + R_i^{-1}(O_j - O_i). \tag{2.2}$$

où R_i^{-1} est la matrice inverse de R_i. L'équation 2.2 peut être simplifiée en définissant:

$$\begin{cases} R_{ij} &= R_i^{-1} R_j \\ O_{ij} &= R_i^{-1}(O_j - O_i). \end{cases} \tag{2.3}$$

D'où:

$$A_i = R_{ij} A_j + O_{ij}. \tag{2.4}$$

L'équation 2.4 permet de calculer les coordonnées d'un atome dans une base i à partir des coordonnées mesurées dans la base j. Ainsi, cette équation définit la position et l'orientation de la base j comme vue de la base i. En particulier, les colonnes de la matrice de rotation R_{ij} sont les vecteurs unitaires du repère de la base j représentés en fonction des vecteurs unitaires du repère de la base i. De la même manière, les entrées du vecteur O_{ij} sont les coordonnées de l'origine de la base j dans le repère de la base i.

Le calcul de la matrice inverse dans le système 2.3 se fait directement parce que la matrice de rotation possède la propriété suivante: son inverse est égale à sa transposée. C'est un cas particulier des matrices orthogonales. Ainsi, si R est une matrice de rotation alors

$$R^{-1} = R^T. \tag{2.5}$$

En utilisant l'équation 2.5, le système 2.3 se simplifie à:

$$\begin{cases} R_{ij} &= R_i^T R_j \\ O_{ij} &= R_j^T (O_j - O_i). \end{cases}$$

2.2.2 Les étapes de la méthode d'empilement

Etape 1

Les coordonnées expérimentales originales de la base seront transformées en les projetant sur son plan affine déterminé par les trois points des atomes N1(a_0, a_1, a_2), C4(b_0, b_1, b_2) et C6(c_0, c_1, c_2) (qui se trouvent dans chaque pyrimidine et purine). L'équation du plan affine est déterminée comme suit :

$$ax + by + cz + d = 0,$$

où

$$\begin{aligned} a &= (a_1 - b_1)(a_2 + b_2) + (b_1 - c_1)(b_2 + c_2) + (c_1 - a_1)(c_2 + a_2), \\ b &= (a_2 - b_2)(a_0 + b_0) + (b_2 - c_2)(b_0 + c_0) + (c_2 - a_2)(c_0 + a_0), \\ c &= (a_0 - b_0)(a_1 + b_1) + (b_0 - c_0)(b_1 + c_1) + (c_0 - a_0)(c_1 + a_1), \\ d &= -(aa_0 + bb_1 + cb_2). \end{aligned}$$

La projection d'un point quelconque U sur le plan en question défini par son vecteur normal $\vec{P}(a, b, c)$ passant par un point A est donc le point U_p appartenant au plan tel que:

$$\vec{AU_p} = \vec{AU} - < \vec{AU}, \vec{N} > \vec{N},$$

où $\vec{N} = \dfrac{\vec{P}}{||\vec{P}||}$ et $< \vec{AU}, \vec{N} >$ est le produit scalaire entre les deux vecteurs $\vec{AU}$ et $\vec{N}$. Dans notre calcul, on prend le point A comme le point représentant l'atome N1.

Une fois que les deux plans de chaque base respectivement sont déterminés ainsi que la projection de chaque atome d'une base sur le plan de cette dernière, on déterminera les nouveaux repères centrés au point représentant l'atome N1 de chacune des deux bases. Chaque repère sera engendré par une base orthonormée formée par trois vecteurs unitaires et orthogonaux deux-à-deux. Pour cela, on utilise trois atomes appartenant au même plan. Les trois atomes considérés sont N1, C4 et C6 pour chaque base. La figure 2.6 montre la position des bases et leurs plans respectifs.

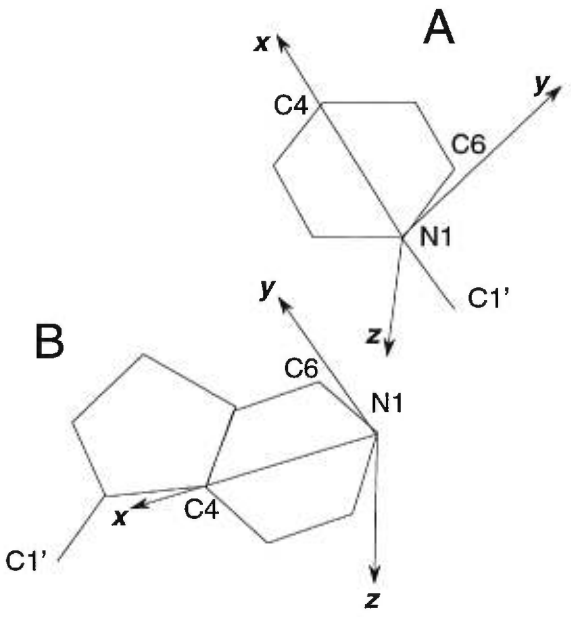


Figure 2.6: Bases hypothétiques appartenant aux plans engendrés par les atomes N1, C4 et C6 de chacune d'elles.

Etant donné que l'empilement chimique entre deux bases est détecté avec une distance ~3.4Å et que les plans respectifs de chacune des bases ne sont pas nécessairement parallèles, on va définir cette distance. Deux bases qui sont très éloignées ne peuvent pas être empilées puisqu'il y aura soit d'autres éléments chimiques entre les deux, soit que la projection d'une sur le plan de l'autre n'intersecte pas cette dernière. Le calcul de la distance en premier économisera beaucoup de calculs inutiles. Pour cela, on déterminera les surfaces de van der Waals pour chacune des bases. Ces surfaces vont être planaires, puisque tous les atomes d'une base se trouvent approximativement dans un même plan. Ainsi, chaque atome aura son propre rayon de van der Waals définissant un cercle. La surface planaire obtenue sera approximée par un polygone qui l'entoure. La figure 2.7 montre les surfaces de van der Waals en question avec les polygones qui les approximent.

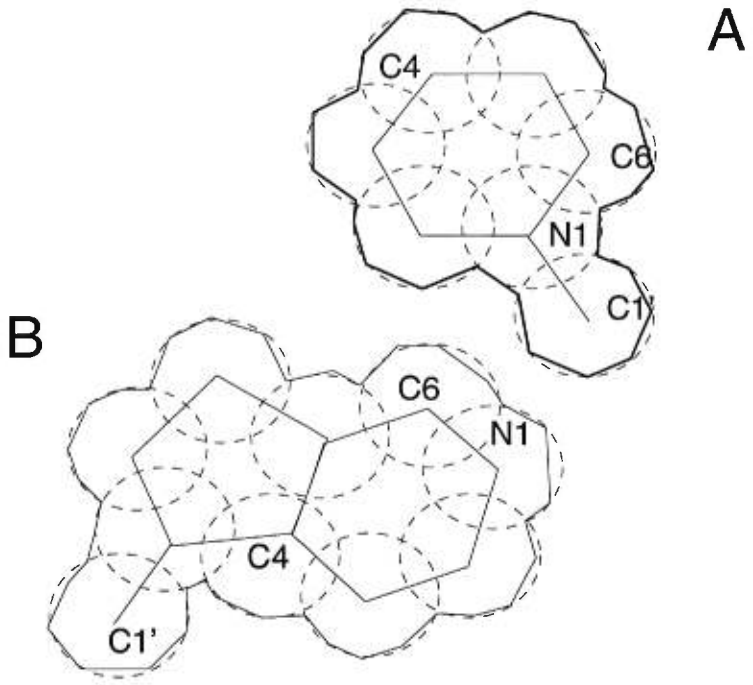


Figure 2.7: Surfaces de van der Waals suivant les plans des bases A et B avec les polygones qui les approximent.

Surfaces des bases utilisées

Le cas qui nous intéresse c'est l'analyse d'empilement des bases azotées A, C, G, T et U présentes dans une structure d'acide nucléique. Pour chaque base d'une structure donnée, on lui associera une surface de van der Waals planaire qui sera approximée par un polygone comme mentionné auparavant. On décrirera alors pour chaque base donnée les atomes considérés ainsi que le polygone approximant sa surface de van der Waals qui lui est associée. La table 2.1 et la figure 2.8 montrent les atomes qui sont pris pour chacune des bases et les surfaces de van der Waals générées par ces derniers.

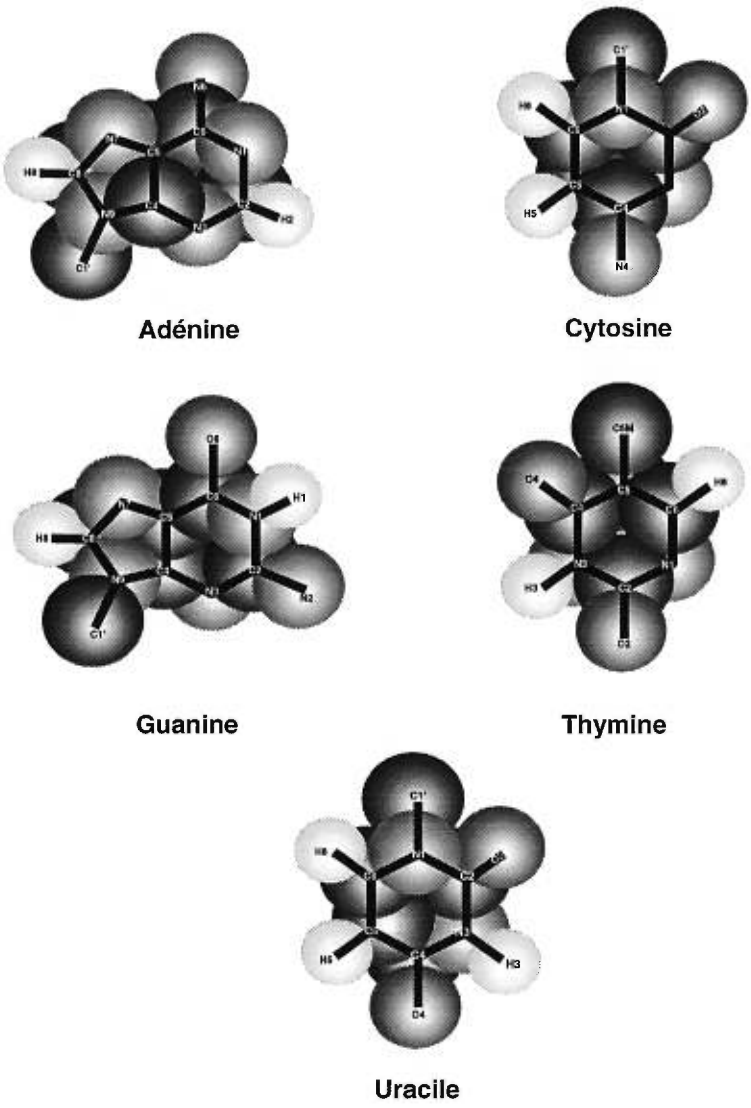


Figure 2.8: Les surfaces de van der Waals considérées des bases adénine, cytosine, guanine, thymine et uracile.

Adénine	{C2, C4, C5, C6, C8, C1', H2, H8, N1, N3, N6, N7, N9}
Cytosine	{C2, C4, C5, C6, C1', H5, H6, N1, N3, N4}
Guanine	{C2, C4, C5, C6, C8, C1', H1, H8, O6, N1, N2, N3, N7, N9}
Thymine	{C2, C4, C5, C6, C1', C5M, H3, H8, O2, O4, N1, N3}
Uracile	{C2, C4, C5, C6, C1', H3, H5, H6, O2, O4, N1, N3}

Table 2.1: Les atomes associés à chacune des bases A, C, G, T et U. L'adénine est considérée avec 13 atomes, la cytosine avec 11 atomes, la guanine avec 14 atomes, la thymine et l'uracile avec 12 atomes.

Etape 2

Après avoir déterminé les surfaces de van der Waals des deux bases dont on veut analyser l'empilement, l'étape suivante consiste à les approximer par des polygones qu'on définira pour chacune des bases. Pour cela, on trouve les points d'intersection des cercles qui forment l'enveloppe extérieure des surfaces de van der Waals. Par exemple, pour la base guanine, on considérera les cercles suivants:

$\mathcal{C}$(N2, R_{N}), $\mathcal{C}$(N3, R_{N}), $\mathcal{C}$(N7, R_{N}), $\mathcal{C}$(C2, R_{C}),

$\mathcal{C}$(C4, R_{C}), $\mathcal{C}$(C6, R_{C}), $\mathcal{C}$(C8, R_{C}), $\mathcal{C}$(C1', R_{C}),

$\mathcal{C}$(H1, R_{H}), $\mathcal{C}$(H8, R_{H}) et $\mathcal{C}$(O6, R_{O});

où $\mathcal{C}$(Xa, R_{X}) est le cercle centré en un certain atome Xa et admettant comme rayon de van der Waals R_{X} où X est la nature de l'atome Xa. Ainsi on détermine les intersections suivantes:

$\mathcal{C}$(N2, R_{N}) $\cap$ $\mathcal{C}$(C2, R_{C}), $\mathcal{C}$(C4, R_{C}) $\cap$ $\mathcal{C}$(C1', R_{C}),

$\mathcal{C}$(C8, R_{C}) $\cap$ $\mathcal{C}$(H8, R_{H}), $\mathcal{C}$(C6, R_{C}) $\cap$ $\mathcal{C}$(O6, R_{O}),

$\mathcal{C}$(C6, R_{C}) $\cap$ $\mathcal{C}$(H1, R_{H}) et $\mathcal{C}$(C2, R_{C}) $\cap$ $\mathcal{C}$(H1, R_{H}).

Les valeurs choisies pour les rayons R_{H}, R_{N}, R_{O} et R_{C} sont celles utilisées par Bondi [4]. La table 2.2 montre les rayons de van der Waals sélectionnés pour les atomes suivant les auteurs mentionnés.

De cette manière, on identifiera les arcs qui définissent la frontière de la surface de van der Waals. Chaque arc sera alors subdivisé uniformément et approximé par des segments. Ce qui définira un polygone pour chacune des bases. Dans notre méthode, l'adénine est donc approximée par un polygone ayant 32 points, la guanine 38 points, la cytosine 30 points, l'uracile 26 points et la thymine 44 points.

Type d'atomes	Rayon de van der Waals(Å)
H	1.20
O (simple ou double lien)	1.52 1.40*
C (aliphatique ou aromatique)	1.70
N (aliphatique ou aromatique)	1.55 1.50*
P	1.80 1.90*
=S	1.75 1.85*
-S-	1.80 1.85*
F	1.35*
Cl	1.80*
Br	1.95*
I	2.15*

Table 2.2: Sélection des rayons de van der Waals suivant Bondi [4] . Les rayons qui sont marqués par un $*$ sont de Weast [52].

Etape 3

Lorsqu'il y a présence d'empilement entre deux bases quelconques, la distance entre ces deux dernières est $\sim$3.4Å suivant les données chimiques. Chimiquement, lorsque cette distance est beaucoup plus grande que cette valeur, il y a deux possibilités : soit qu'il y a une autre molécule qui va se loger entre ces deux bases, soit que ces dernières ne peuvent pas présenter un empilement, c'est-à-dire que les projections successives d'une base sur le plan de l'autre donnent une intersection vide. De plus, le cas où la distance est beaucoup plus petite que $\sim$3.4Å ne présente pas une interaction chimique verticale susceptible de créer un empilement, les deux bases dans ce cas-là seront tout simplement décalées l'une de l'autre de sorte que les projections ne présentent pas d'intersection possible. Le fait que les deux plans des bases ne sont pas nécessairement parallèles nous a amené à définir une distance entre bases. Ainsi, on vérifiera si la distance est dans un intervalle donné; si oui, on procédera à l'étape suivante, sinon on pourra conclure tout simplement que les deux bases ne présentent pas d'empilement.

Ceci a pour but d'éviter de faire des calculs inutilement. L'intervalle ainsi pris est de 2.4 à 4.4Å.

Jusqu'à maintenant, on a réduit la notion de base chimique à une structure géométrique qui est un polygone appartenant à un certain plan donné. Ce polygone est défini par un certain nombre de points planaires. La définition de la distance entre les deux bases ainsi utilisée est:

$$\frac{1}{|A|+|B|}\left(\sum_{k\in A} d(k,P(k)) + \sum_{k\in B} d(k,P(k))\right),$$

où k est un point appartenant à une base X vue comme un ensemble de points formant un polygone (on note $k \in X$), $|X|$ représente le nombre de point dans la base X et $d(k,P(k))$ est la distance entre le point k et sa projection sur le plan de la base adjacente.

Etape 4

Chaque polygone associé à une base va être projeté sur le plan contenant la base adjacente. Nous montrerons dans la figure 2.7 en considérant deux bases A et B qu'on a approximé par des polygones, les deux projections orthogonales que l'on aura à faire. Chaque projection donnera lieu à la détermination d'une intersection entre deux polygones. L'intersection est alors un ou plusieurs polygones dont on tentera de calculer l'aire. En utilisant le théorème de Green, l'aire d'une région bidimensionnelle S est donnée par :

$$\text{Aire}(S) = \int_S dS = \int_{Frt(S)} \frac{1}{2}(p_x dy - p_y dx) = \int \frac{1}{2} p \wedge dp,$$

où $p(p_x, p_y)$ est un point sur la frontière $Frt(S)$ de la région S, dp est un vecteur infinitésimal le long de la frontière et $\wedge$ représente le produit vectoriel en deux di-

mensions. L'élément $\frac{1}{2}p \wedge dp$ est l'aire d'un triangle ayant comme base un morceau infinitésimal de la frontière et un sommet à l'origine. Pour les polygones, les frontières sont linéaires; l'intégration le long d'un côté d'un polygone donne l'aire d'un triangle ayant ce côté comme sa base et un sommet à l'origine. Ainsi, l'aire du polygone est la somme de toutes ces aires de triangles:

$$\text{Aire} = \sum_{1 \leq k \leq n} \frac{1}{2}(s_k \wedge s_{k+1}),$$

où les s_k sont les n sommets du polygone pris cycliquement, c'est-à-dire que $n+1 = 1$. Le nombre des multiplications nécessaires peut être réduit de moitié en réarrangeant les termes comme suit:

$$\text{Aire} = \sum_{1 \leq k \leq n} \frac{1}{2}(s_{k,x}s_{k+1,y} - s_{k,y}s_{k+1,x}) = \sum_{1 \leq k \leq n} \frac{1}{2}s_{k,x}(s_{k+1,y} - s_{k-1,y});$$

où $s_k(s_{k,x}, s_{k,y})$ est le $k^{ième}$ sommet du polygone S.

Le problème est réduit à la recherche de l'intersection entre deux polygones. Ce que nous voulons montrer dans cette section, ce sont les principes permettant de traiter graphiquement la procédure de l'opération intersection. Pour cela, on va considérer que les polygones constituent des surfaces pleines. Le résultat de l'intersection sera alors la figure délimitant la surface obtenue en appliquant l'opération de l'intersection. La figure 2.9 présente un exemple d'intersection entre deux polygones quelconques.

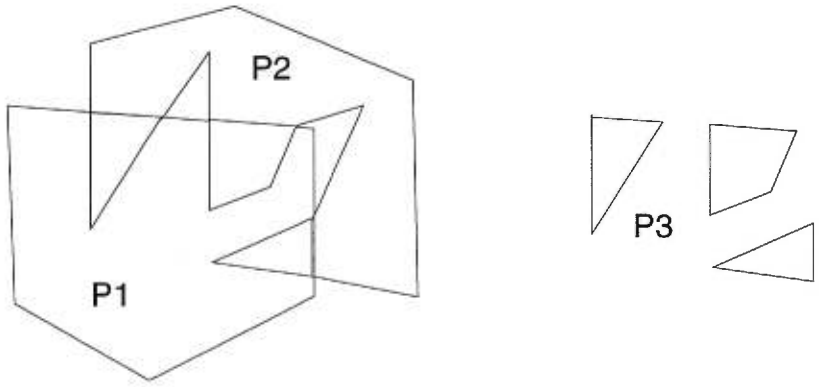


Figure 2.9: Exemple d'intersection entre deux polygones P1 et P2 donnant un graphe P3.

Dans cet exemple, on peut remarquer que l'intersection obtenue n'est pas nécessairement un polygone. L'intersection est alors un ensemble de polygones. Dans le cas où les deux polygones sont convexes, cette intersection est convexe. Pour réaliser l'opération intersection entre deux polygones P_1 et P_2, nous allons utiliser le même principe de la représentation par les diagrammes de Venn comme dans la figure 2.9. L'ensemble des deux polygones A et B sera représenté par un graphe où les noeuds correspondent aux sommets et les arcs aux arêtes. Chaque arc admet l'un des états suivants possibles:

1. arcP_1 : arc de A uniquement
2. arcP_2 : arc de B uniquement
3. InterP_1 : arc complètement à l'intérieur de A
4. InterP_2 : arc complètement à l'intérieur de B
5. InterP_1P_2 : arc complètement à l'intérieur de A et B, i.e. l'arc est redondant.

Au moment de la construction du graphe, si des arêtes se coupent, de nouveaux

sommets et de nouvelles arêtes sont créés. Nous montrons dans la figure 2.10 un exemple de la procédure.

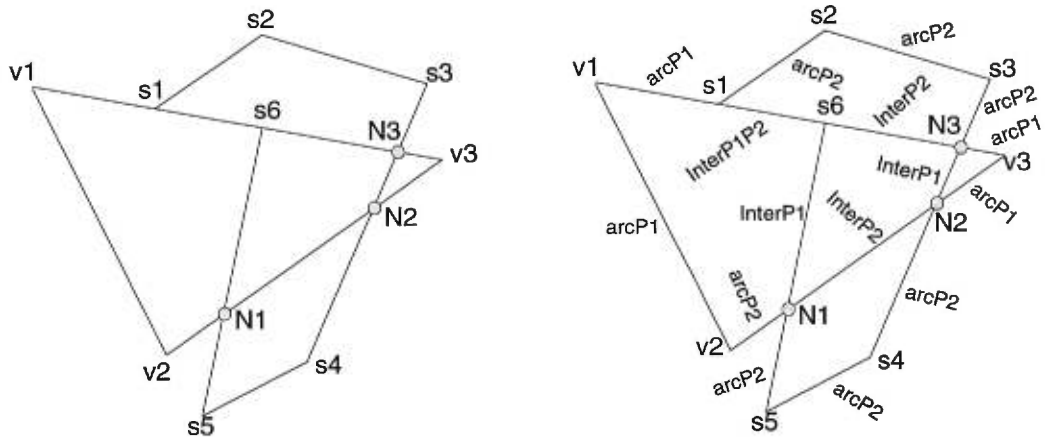


Figure 2.10: Les graphes associés aux deux polygones P_1 et P_2. P_1 est défini par les sommets s1, s2, ..., s6 et P_2 est défini par les sommets v1, v2, v3. Les sommets N1, N2 et N3 sont les points d'intersection entre P_1 et P_2. L'état de chaque arc du graphe est noté dans le graphe à droite.

Les points N_k sont les sommets ajoutés. La figure qui résulte de l'opération intersection est formée d'arêtes correspondant à des arcs ayant les états suivants:

InterA, InterB ou InterAB

Ainsi de l'exemple 2.10, on retiendra les arcs vérifiant les états ci-dessus de la figure résultante. Les parties obtenues du graphe sont représentées alors dans la figure 2.11.

L'algorithme de l'opération intersection est alors très simple et peut se résumer comme ainsi:

1. Construire le graphe commun de P_1 et P_2.

2. Pour chaque arc du graphe, si son état est:

InterP_1, InterP_2 ou InterP_1P_2,

on inclut l'arête correspondante dans la figure finale.

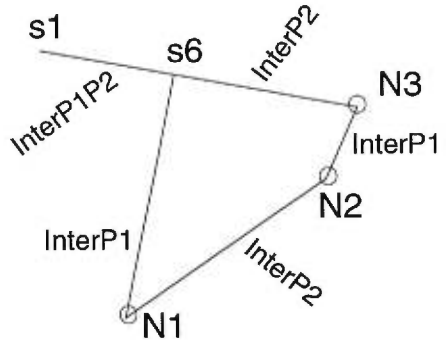


Figure 2.11: Graphe généré de l'opération d'intersection traitée dans la figure 2.10.

Sur cette base, nous avons utilisé la structure de données de la figure 2.12 pour représenter notre polygone.

```
class     Sommet
{
     Sommet();
     Vecteur   : V;  /* Coordonnées du sommet */
     Sommet    : *precedent, *suivant;
}

class     Arc
{
     Arc();
     Sommet  :     *sommet_precedent, *sommet_suivant;
               /* Pointeurs vers les sommets à l'extrémité de l'arc */
     Arc     :     *arc_precedent, *arc_suivant;
               /* Pointeurs vers les arcs associés aux mêmes sommets */
     char    :     *etat;
               /* "etat" appartient à l'ensemble {arcP1, arcP2, InterP1, InterP2, InterP1P2};
}

class     Polygone
{
     Polygone();
     Arc     :     *arcs;
     Sommet  :     *soms;
}
```

Figure 2.12: Structure de données employées pour un sommet, un arc et un polygone.

La fonction *Intersection* (voir figure 2.13) calcule l'intersection entre deux polygones $P1$ et $P2$. Elle appelle la fonction *Structure* qui construit le graphe. Les sommets et les arcs sont comptés et mémorisés dans les listes S et A des sommets et des arcs respectivement. OKvérifie que l'opération s'est bien passée.

```
Polygone Intersection(Polygone *P1, Polygone: *P3)
{
    Polygone :    *P3;
    Sommet   :    *S;
    Arc      :    *A;
    int      :    OK=0;
    Structure(&P1, &P2, &S, &A, OK);
    /* fonction qui construit le graphe en associant à chaque arc un état */
    if (OK)
    {
        Arc : *courant = A;
        while (!courant)
        {
            if ( Verifie(A->etat) )
              /* Verifier si A-etat est dans {InterP1, InterP2, InterP1P2} */
            {
                Ajoute(&P3, &A);
                  /* Insérer l'arc A dans le polygone P3 */
                courant = courant->arc_suivant;
                  /* Avancer à l'arc suivant dans la liste */
            }
        }
    }
    return(P3);
}
```

Figure 2.13: La fonction construisant l'intersection entre deux polygones $P1$ et $P2$.

Etape 5

La détermination de l'empilement entre les deux bases quelconques A et B dans une structure moléculaire repose sur le fait que l'intersection entre les polygones

représentatifs des bases soit non vide. La figure 2.14a nous montre l'intersection entre le polygone B et le polygone A projeté sur le plan contenant le polygone B. Contrairement à la figure 2.14b qui nous montre l'intersection du polygone A avec la projection du polygone B sur le plan contenant le polygone A.

Pour avoir une idée de l'importance de cet empilement, on a défini un paramètre qui mesurera le taux de chaque empilement présent entre deux bases. Ainsi, le taux d'empilement s'exprime comme suit:

$$\tau = \frac{1}{2}\left(\frac{\text{Aire(AB)}}{\text{Aire(A)}} + \frac{\text{Aire(BA)}}{\text{Aire(B)}}\right).$$

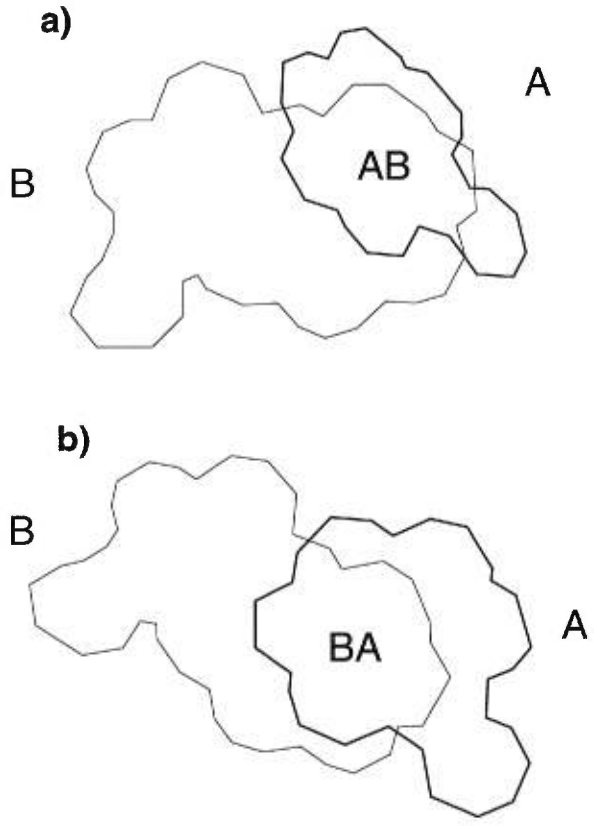


Figure 2.14: a) Intersection entre le polygone B et la projection de A sur le plan contenant B. b) Intersection entre le polygone A et la projection de B sur le plan contenant A.

2.3 Analyse des empilements dans la structure d'ARNt^{Phe} et du hammerhead

2.3.1 Les empilements dans la structure d'ARNt^{Phe}

Beaucoup d'études structurales concernant l'ARNt ont connu des changements révolutionnaires durant les dernières années. Parmi les informations biochimiques accumulées concernant cette molécule en plus de vingt ans avant 1973, aucune information valable ne permettait de déterminer sa structure tridimensionnelle. Cependant, au début de l'année 1973, l'ARNt^{Phe} de la levure a été cristallisée [20]. Les coordonnées atomiques ont été connues à 2.5Å de résolution, à l'aide de l'analyse de diffraction aux rayons X faite sur deux formes différentes de la même molécule [24, 37]. La connaissance de la structure tridimensionnelle de la molécule d'ARNt a été très utile pour les recherches qui permettent de poser des questions précises sur ses fonctions biologiques et chimiques. Le but de cette section est de décrire en détail les empilements de bases azotées dans la structure tridimensionnelle de l'ARNt.

En plus des quatre bases canoniques, l'ARNt possède des modifications qui sont désignées par les symboles : m^2G_{10}, m^5C_{40}, m^7G_{46} et m^1A_{58} (par exemple, m^2G_{10} indique un groupe méthyle à la position du carbone 2 du résidu guanine 10); $m_2^2G_{26}$ indique deux groupes méthyles au niveau de l'azote 2 de la guanine 26; la méthylation du 2'OH du ribose est indiquée par un "m" après le symbole comme $C_{32}m$, $G_{34}m$ et $U_{50}m$.

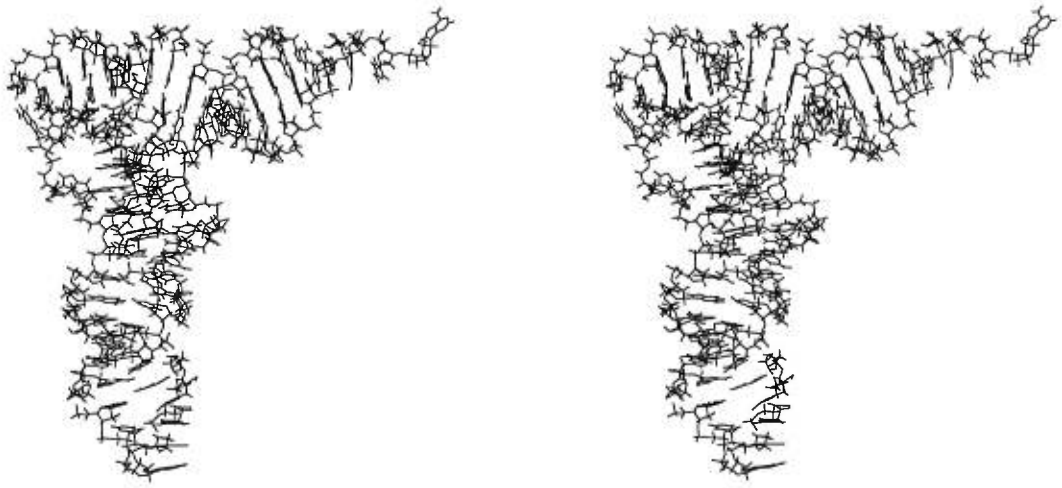


Figure 2.15: Vue stéréoscopique de l'ARNt^{Phe}. Ceci peut être mieux visualisé par des lunettes stéréoscopiques qui fusionneront les deux images. Cependant, les deux images peuvent être fusionnées sans l'utilisation de ces lunettes, simplement en relaxant les muscles oculaires jusqu'à ce que l'image tridimensionnelle apparaisse.

A. Rich et U. L. RajBhandary [38] ont déjà classé tous les empilements de bases de la molécule d'ARNt^{Phe}, observés par visualisation. Les empilements visualisés par A. Rich et U. L. RajBhandary, se résument dans la figure 2.16.

Nous allons utiliser l'ARNt et comparer nos résultats avec ceux de A. Rich et U. L. RajBhandary. Nos résultats sont montrés dans les tables 2.3 et 2.4.

Les interactions verticales entre les bases mentionnées dans [38] ont toutes été détectées à partir de $\tau > 20\%$. La seule base qui n'est pas impliquée dans de telles interactions verticales est U_{47} (i.e. le taux d'empilement de la base U_{47} et une autre base quelconque est $\tau = 0$), qui sort à l'extérieur du reste de la molécule. Une autre base, la guanine G_{20}, présente un taux d'empilement négligeable ($\tau = 0.284\%$) avec l'adénine A_{21}, coïncidant avec la description de [38].

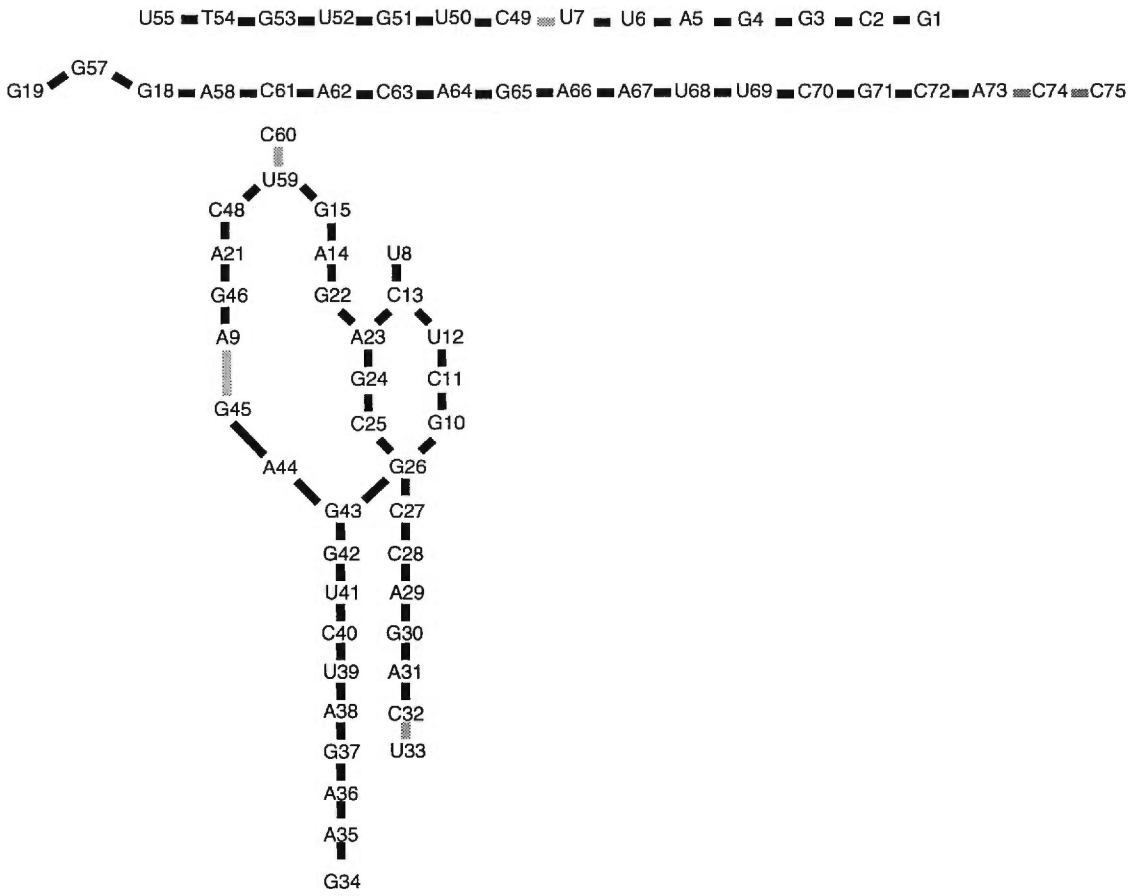


Figure 2.16: Diagramme montrant l'interaction hydrophobique entre les différents nucléotides dans l'ARNt^{Phe} de la levure. L'empilement entre bases est représenté par le trait noir. L'empilement partiel est montré par le trait gris.

résidus	angle (degrés)	distance (Å)	$d(C1'_1,C1'_2)$ (Å)	τ_{12} (%)	τ_{21} (%)	τ (%)
G_{51} / A_{62}	13.309	3.027	13.722	0.000	0.020	0.010
G_1 / G_{71}	10.967	3.467	13.884	0.000	0.022	0.011
G_{53} / A_{58}	9.116	3.175	11.464	0.284	0.000	0.142
G_{26} / G_{43}	47.197	3.715	14.162	0.000	0.403	0.201
A_5 / A_{67}	19.926	3.777	13.715	0.257	0.182	0.220
G_{20} / A_{21}	89.127	4.409	7.290	0.000	0.567	0.284
C_{11} / C_{25}	18.218	2.846	8.316	0.000	1.153	0.576
U_6 / U_{68}	16.987	2.331	7.502	1.891	0.000	0.946
G_{26} / G_{45}	32.046	3.100	11.089	2.122	0.041	1.081
G_{26} / A_{44}	37.093	2.031	12.750	0.168	3.179	1.673
A_9 / G_{10}	33.508	4.413	7.282	0.000	3.368	1.684
U_{52} / C_{63}	9.096	3.429	8.303	1.765	1.914	1.840
C_2 / C_{72}	6.126	3.333	7.406	3.488	0.292	1.890
U_{54} / C_{61}	8.518	3.457	7.392	0.649	3.927	2.288
U_{33} / A_{35}	19.105	2.003	6.901	4.784	1.227	3.006
A_{58} / C_{60}	56.673	2.635	6.531	7.977	0.000	3.989
U_{55} / A_{58}	20.137	3.781	8.180	4.579	4.243	4.411
U_{55} / G_{57}	22.468	2.925	6.527	8.782	1.961	5.372
U_8 / G_{22}	19.513	3.067	9.272	10.213	0.645	5.429
C_{13} / A_{14}	9.351	3.627	6.027	3.470	7.866	5.668
G_{15} / A_{21}	3.551	3.037	10.290	7.786	6.678	7.232
A_9 / G_{22}	10.675	3.484	9.723	16.465	11.660	14.063
A_{31} / C_{40}	19.849	2.987	8.968	6.359	24.335	15.347
A_{21} / G_{22}	6.501	3.215	5.876	13.556	17.247	15.401
C_{13} / A_{23}	9.516	2.933	8.463	25.571	12.011	18.791
A_5 / U_{69}	10.818	3.103	8.105	19.529	18.058	18.793
U_8 / G_{15}	16.324	3.690	9.129	12.442	25.888	19.165

Table 2.3: Résultats de tous les empilements présents dans l'ARNt^{Phe} ayant un taux d'empilement $\tau < 20\%$. On y exhibe les paramètres angle et distance entre deux bases données, distance entre les atomes C1' faisant partie des sucres associés à chacune des bases et les paramètres τ_{12}, τ_{21} et τ qui représentent successivement le taux d'empilement de la première base sur la deuxième, le taux d'empilement de la deuxième base sur la première et le taux d'empilement total entre les deux bases; qui est tout simplement la moyenne de τ_{12} et τ_{21}.

résidus	angle (degrés)	distance (Å)	$d(C1'_A,C1'_B)$ (Å)	τ_{12} (%)	τ_{21} (%)	τ (%)
U_{12} / G_{24}	14.392	2.708	8.724	25.257	17.353	21.305
C_{48} / U_{59}	13.993	3.305	4.600	21.706	22.331	22.019
U_{8} / C_{13}	6.520	3.287	9.305	24.160	21.712	22.936
C_{49} / A_{66}	12.621	2.708	8.031	32.882	16.517	24.699
U_{7} / A_{67}	7.671	3.351	8.348	29.959	21.232	25.596
C_{70} / G_{71}	12.297	3.032	5.465	27.252	23.971	25.611
C_{61} / A_{62}	9.329	3.342	5.687	26.256	25.765	26.010
C_{49} / U_{50}	22.925	3.555	5.979	13.107	39.676	26.392
U_{50} / G_{51}	12.411	2.886	5.559	34.896	23.366	29.131
U_{52} / G_{53}	9.804	3.187	5.300	32.528	27.579	30.053
G_{4} / C_{70}	7.180	3.156	8.219	30.114	31.298	30.706
C_{28} / G_{43}	11.926	3.351	8.274	37.143	25.338	31.241
G_{10} / G_{26}	5.331	3.492	8.054	33.361	33.546	33.453
U_{69} / C_{70}	8.058	3.459	5.412	34.249	36.304	35.277
C_{25} / G_{26}	20.762	3.792	5.279	43.725	26.977	35.351
G_{22} / A_{23}	4.733	3.298	6.323	37.976	33.278	35.627
U_{41} / G_{42}	4.254	3.273	5.279	38.705	32.938	35.821
A_{14} / G_{22}	3.868	3.301	7.946	33.314	38.675	35.994
G_{15} / U_{59}	6.256	3.810	9.204	35.794	37.086	36.440
A_{36} / G_{37}	27.903	3.406	6.127	37.147	36.198	36.673
C_{28} / A_{29}	17.652	3.330	5.128	36.558	37.624	37.091
C_{63} / A_{64}	20.744	3.462	5.589	39.790	35.320	37.555
C_{40} / U_{41}	16.145	3.849	5.699	28.858	47.946	38.402
G_{30} / U_{41}	9.478	3.157	7.622	36.922	41.565	39.244
C_{2} / G_{3}	13.744	3.632	5.398	43.150	37.139	40.144
C_{72} / A_{73}	10.716	3.350	5.292	47.669	32.817	40.243
G_{1} / A_{73}	11.079	3.399	7.363	46.643	34.027	40.335
U_{7} / C_{49}	12.321	3.798	5.725	49.638	34.999	42.319
A_{58} / C_{61}	7.796	3.553	8.286	39.984	45.581	42.782
A_{14} / G_{15}	7.706	3.506	5.063	43.364	42.719	43.042
C_{27} / C_{28}	25.790	3.963	6.056	37.124	51.995	44.559
U_{6} / U_{7}	28.095	3.780	5.754	32.781	57.711	45.246
U_{50} / G_{65}	11.715	3.750	7.599	43.597	47.368	45.482
U_{39} / C_{40}	12.876	3.575	5.541	43.254	52.775	48.014
C_{56} / G_{57}	9.901	3.577	5.175	52.793	43.596	48.194
A_{29} / G_{30}	4.975	3.636	5.682	48.133	48.900	48.516
C_{74} / C_{75}	30.132	3.716	5.317	26.384	71.341	48.863
A_{44} / G_{45}	13.827	3.724	5.461	52.331	46.180	49.255
G_{19} / G_{57}	22.082	3.651	6.647	38.076	60.711	49.393
A_{21} / C_{48}	4.732	3.372	8.502	42.967	55.936	49.452
C_{11} / U_{12}	17.047	3.557	5.612	55.732	44.047	49.890
A_{29} / G_{42}	5.101	3.341	8.408	50.076	49.711	49.894

résidus	angle (degrés)	distance (Å)	$d(C1'_A, C1'_B)$ (Å)	τ_{12} (%)	τ_{21} (%)	τ (%)
A_9 / G_{46}	5.450	3.433	5.482	45.115	55.116	50.116
G_{51} / A_{64}	10.431	3.516	8.094	55.389	48.403	51.896
C_{32} / U_{33}	19.790	3.635	5.453	40.245	64.624	52.434
A_{64} / G_{65}	4.004	3.410	5.287	50.153	58.258	54.206
A_{21} / G_{46}	3.496	3.767	8.228	55.110	56.213	55.662
G_3 / G_4	15.721	3.378	5.208	54.357	60.134	57.246
A_{66} / A_{67}	10.469	3.602	5.552	54.305	61.625	57.965
U_{68} / U_{69}	15.066	3.588	5.345	54.015	63.476	58.746
U_{12} / C_{13}	11.725	3.920	5.112	64.723	53.227	58.975
G_{18} / G_{57}	6.409	3.498	7.748	58.093	62.263	60.178
A_9 / G_{45}	16.371	3.567	5.956	60.183	62.219	61.201
G_3 / G_{71}	6.493	3.340	8.268	62.466	60.355	61.411
G_{10} / C_{11}	5.017	3.586	5.971	51.357	72.182	61.770
G_{42} / G_{43}	6.968	3.446	5.254	64.790	59.106	61.948
G_{65} / A_{66}	4.482	3.303	5.332	64.296	60.000	62.148
A_{38} / U_{39}	9.783	3.636	5.599	57.309	67.105	62.207
A_{73} / C_{74}	19.126	3.766	5.912	47.100	77.629	62.365
G_{53} / A_{62}	3.698	3.523	7.525	67.883	58.164	63.023
G_{18} / A_{58}	4.438	3.414	6.936	67.165	60.125	63.645
A_{23} / G_{24}	1.540	3.513	5.237	61.370	66.589	63.979
G_{30} / A_{31}	12.250	3.671	5.265	67.950	60.139	64.044
A_{62} / C_{63}	7.970	3.628	5.521	54.971	74.271	64.621
G_{71} / C_{72}	7.097	3.658	5.513	61.931	68.339	65.135
G_{24} / C_{25}	23.853	3.639	5.469	65.418	69.003	67.210
A_{31} / C_{32}	16.170	3.795	5.114	62.789	71.779	67.284
A_{67} / U_{68}	5.339	3.493	5.465	57.136	78.446	67.791
G_{34} / A_{35}	8.573	3.470	5.449	71.236	65.788	68.512
G_1 / C_2	1.822	3.464	5.694	61.836	76.059	68.947
G_{43} / A_{44}	15.093	3.625	5.135	70.615	67.336	68.975
G_4 / A_5	9.987	3.797	5.370	71.777	66.902	69.339
G_{53} / U_{54}	3.857	3.449	6.061	65.407	76.806	71.107
G_{37} / A_{38}	6.104	3.475	5.131	73.342	70.461	71.901
U_{54} / U_{55}	15.562	3.730	5.177	80.315	64.478	72.396
U_{59} / C_{60}	21.232	3.825	5.662	77.969	69.961	73.965
G_{51} / U_{52}	2.349	3.571	5.248	69.757	84.029	76.893
A_5 / U_6	6.077	3.645	5.472	66.601	87.643	77.122
G_{26} / C_{27}	11.819	3.687	5.530	69.332	89.511	79.422
A_{35} / A_{36}	8.070	3.627	4.678	80.755	79.103	79.929

Table 2.4: Tableau contenant tous les empilements mentionnés par les auteurs dans [38] ainsi que par notre programme.

Pour les régions où on trouve le cas suivant: soient quatre bases B1, B2, B3 et B4 telles que les deux premières et les deux dernières forment une paire de bases (les deux bases ont des liens hydrogènes), B1 est adjacente à B3 et B2 est adjacente à B4 (Voir la figure 2.17); les auteurs ont mentionné seulement l'empilement présent entre les bases B1 et B3 ou les bases B2 et B4. Il ne figure pas dans leur description l'empilement entre les bases B1 et B4 ou les bases B2 et B3. On appellera ce genre d'empilement *croisé*.

Notre approche a montré qu'il existe des empilements croisés assez signifiant atteignant dans le cas des bases G_{53} et G_{62} un taux $\tau = 63.023\%$ avec une distance de 3.523.

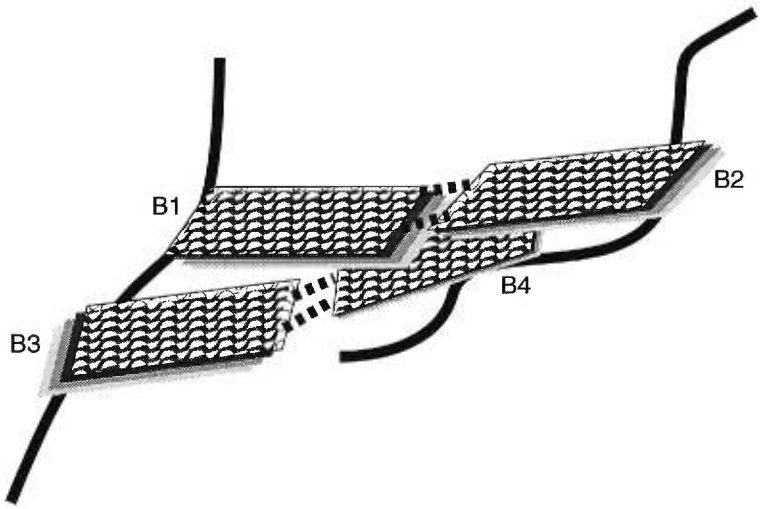


Figure 2.17: Schéma d'un exemple d'empilement croisé entre les bases B1 et B4. Les bases B2 et B3 sont empilées aussi. Les ponts hydrogènes entre les bases sont montrés par des lignes pointillées et l'adjacence des bases est indiquée par une ligne continue.

La table 2.5 contient les empilements croisés non mentionnés par A. Rich et U. L. RajBhandary. Un exemple est montré dans la figure 2.18 où les bases G_3 et G_{71} (G_{71} présente un lien hydrogène avec C_2) sont empilées d'une manière croisée.

résidus	angle (degrés)	distance (Å)	$d(C1'_1,C1'_2)$ (Å)	τ_{12} (%)	τ_{21} (%)	τ (%)
U_{12} / G_{24}	14.392	2.708	8.724	25.257	17.353	21.305
C_{49} / A_{66}	12.621	2.708	8.031	32.882	16.517	24.699
U_7 / A_{67}	7.671	3.351	8.348	29.959	21.232	25.596
G_4 / C_{70}	7.180	3.156	8.219	30.114	31.298	30.706
C_{28} / G_{43}	11.926	3.351	8.274	37.143	25.338	31.241
G_{30} / U_{41}	9.478	3.157	7.622	36.922	41.565	39.244
G_1 / A_{73}	11.079	3.399	7.363	46.643	34.027	40.335
U_{50} / G_{65}	11.715	3.750	7.599	43.597	47.368	45.482
A_{29} / G_{42}	5.101	3.341	8.408	50.076	49.711	49.894
G_{51} / A_{64}	10.431	3.516	8.094	55.389	48.403	51.896
G_3 / G_{71}	6.493	3.340	8.268	62.466	60.355	61.411
G_{53} / A_{62}	3.698	3.523	7.525	67.883	58.164	63.023

Table 2.5: Les empilements croisés ayant une évaluation $\tau > 20\%$.

On a pu détecter un empilement entre les bases C_{56} et G_{57} avec un taux $\tau_{C_{56}/G_{57}} = 48.194\%$ qui n'a pas été mentionné dans l'analyse de la structure tridimensionnelle de l'ARNt^{Phe} (voir figure 2.16).

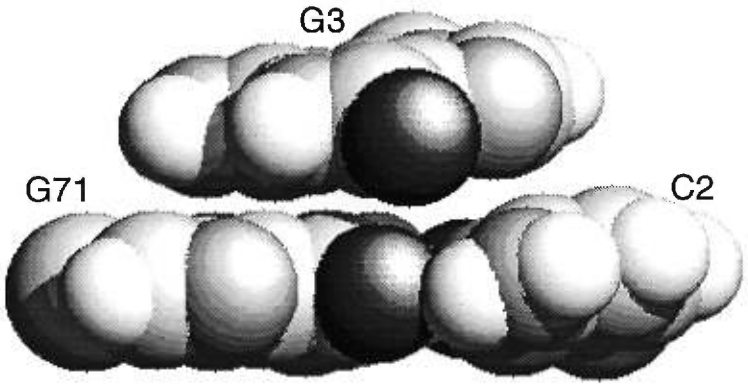


Figure 2.18: Visualisation de l'empilement de type croisé entre les deux bases G_3 et G_{71} dans l'ARNt^{Phe}. La base G_{71} a des liens hydrogènes avec C_2. On remarque visuellement que cet empilement ($\tau_{G_3/G_{71}} = 61.411\%$) est plus important que celui présent entre C_2 et G_3 ($\tau_{G_2/G_3} = 40.144\%$).

Rich et RajBhandary ont aussi affirmé qu'il existe un empilement croisé entre G_{26} et G_{43}. Cependant, notre méthode n'a pas identifié ce dernier; le taux calculé est

$\tau_{G_{26}/G_{43}} = 0.201\%$. On a aussi procédé à une vérification visuelle pour montrer que les deux bases mentionnées n'étaient pas empilées (voir figure 2.19).

En ce qui concerne les cas limites ou empilements partiels notés par les auteurs ($G_7/G_{49}, G_9/G_{45}, G_{32}/G_{33}, G_{73}/G_{74}$ et G_{74}/G_{75}), nos résultats révèlent qu'ils n'ont aucune caractéristique commune. Les paramètres calculés pour ces cas démontrent qu'il y a d'autres empilements qui leurs sont semblables, cependant ils ne sont pas considérés comme étant partiels par Rich et RajBhandary.

La figure 2.20 nous montre la distribution des empilements par rapport à la distance qui sépare les bases prises deux à deux. Sur ce diagramme, on constate qu'on peut diviser les points ayant un taux d'empilement plus grand que 20% en deux classes. Une première classe ayant des points avec $20\% \leq \tau \leq 36\%$, dans ce cas la distance entre les bases deux à deux varie entre 2.7 et 3.6Å. La deuxième classe est plus importante que la première, elle est représentée par les points qui sont définis avec $\tau > 36\%$, dans ce cas la distance varie de 3.2 à 3.9Å.

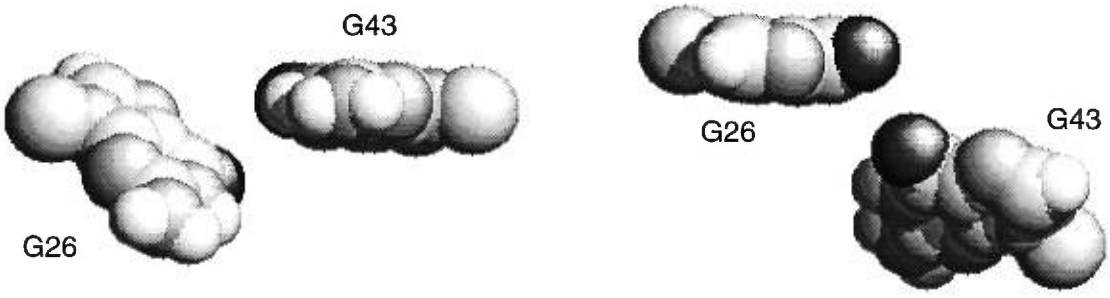


Figure 2.19: Visualisation des deux bases G_{26} et G_{43}. Cette figure montre que la projection orthogonale de G_{26} sur le plan support de G_{43} génère une intersection vide avec la surface de G_{43}. Par contre, G_{43} projetée orthogonalement sur le plan contenant G_{26}, intersecte cette dernière. Le taux d'empilement calculé dans ce cas est $\tau_{G_{26}/G_{43}} = 0.201\%$ et l'angle entre les deux bases est 47.197 degrés.

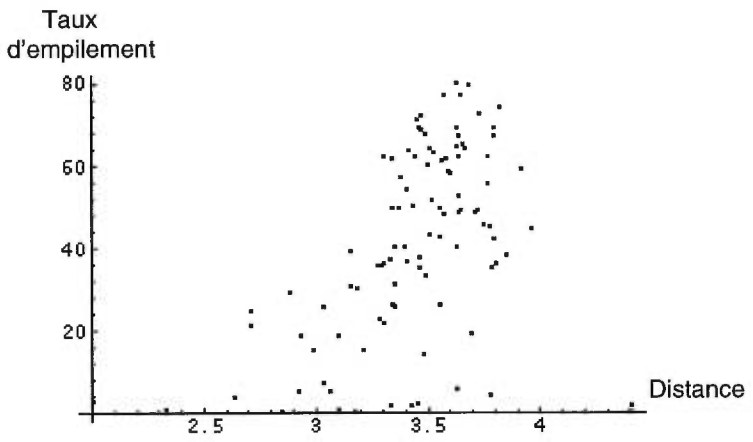


Figure 2.20: Distance entre les bases par rapport au taux d'empilement.

L'angle entre les deux bases empilées nous donne une bonne idée sur leur position. Pour les empilements retenus avec $\tau > 0\%$, on remarque que cet angle varie entre 1.54 (angle entre G_{23} et G_{24} avec $\tau = 63.979\%$) et 89.127 degrés (angle entre G_{20} et G_{21} avec $\tau = 0.284\%$). Dans la figure 2.21, on distingue nettement une classe de point avec un angle variant dans un intervalle $[2, 4.5]$. Par contre, l'empilement est significatif (dans notre cas pour $\tau > 20\%$) pour tous les points ayant un empilement avec un angle ≤ 30.132 degrés (angle entre C_{74} et C_{75} avec $\tau = 48.863\%$).

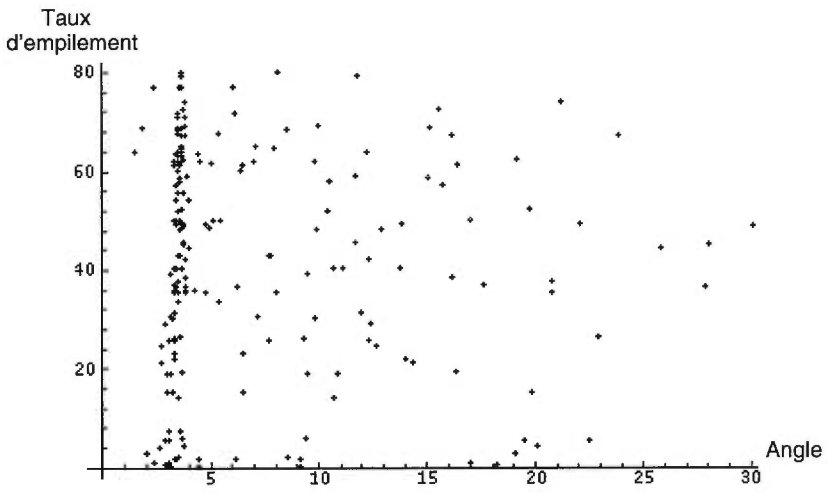


Figure 2.21: Angle par rapport au taux d'empilement.

Un autre paramètre $d(\text{C1'}_1,\text{C1'}_2)$, distance entre les deux atomes C1' des sucres associés à chacune des deux bases données 1 et 2, a été généré. On distingue deux classes de points dans la figure 2.22. Une première classe contenant les points avec 4.600Å $\leq d(\text{C1'}_1,\text{C1'}_2) \leq$ 6.323Å (où $d(\text{C1'}_{\text{C}_{48}},\text{C1'}_{\text{U}_{59}}) = 4.600$Å et $d(\text{C1'}_{\text{G}_{22}},\text{C1'}_{\text{A}_{23}}) =$ 6.323Å), est caractérisée par tous les empilements présents entre deux bases adjacentes (bases i et $i+1$ dans la séquence) dans la séquence d'ARNt^{Phe}, avec $\tau > 20\%$. Dans ce cas, la torsion au niveau du brin n'est pas importante, et les deux atomes C1' associés à chacune des bases sont plus ou moins superposés. Les seules bases appartenant à cette classe ne vérifiant pas cette propriété sont:

$$\text{C}_{48}/\text{U}_{59},\ \text{U}_7/\text{C}_{49},\ \text{A}_9/\text{G}_{46} \text{ et } \text{A}_9/\text{G}_{45}.$$

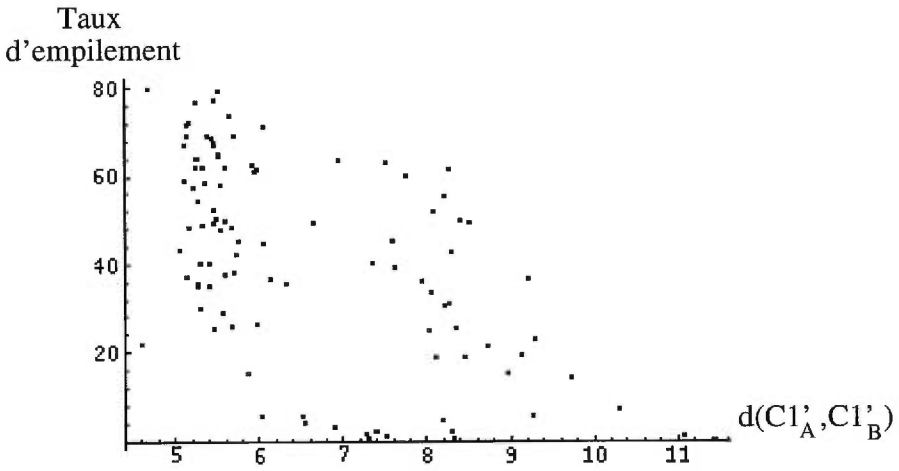


Figure 2.22: Distance entre les atomes C1' des sucres de chacune des deux bases considérées par rapport au taux d'empilement.

Cette région contient des courbures importantes causant ainsi ce rapprochement des atomes C1' des paires de bases mentionnées. La deuxième classe se distingue par les points avec 6.527Å $\leq d(\text{C1'}_1,\text{C1'}_2) \leq$ 11.464Å. Les seuls empilements entre bases adjacentes qui font partie de cette classe sont:

A_9/G_{10} et G_{20}/A_{21}.

Cependant, ces derniers ne sont pas retenus à cause de leurs taux négligeables, 0.284 et 1.684% respectivement.

On peut remarquer aussi trois sites intéressants où l'empilement se fait entre une base et deux autres bases adjacentes dans la séquence. Ainsi, A_9 est empilée entre G_{45} et G_{46}, G_{18} est empilée entre A_{58} et G_{57} et finalement G_{57} est empilée entre G_{18} et G_{19}. Aussi, les résultats nous montrent que le triplet de bases adjacentes $\{G_{34}, A_{35}, A_{36}\}$ est fortement empilé en comparaison à tout autre triplet de ce genre, avec une évaluation moyenne de $\frac{1}{2}(\tau_{A_{34}/A_{35}}+\tau_{A_{35}/A_{36}}) = 74.22\%$. Deux autres triplets sont aussi importants à mentionner: $\{G_4, A_5, A_6\}$ et $\{G_{53}, U_{54}, U_{55}\}$ avec des pourcentages moyens 73.23 et 71.75% respectivement. La figure 2.23 montre le comportement des trois paramètres (angle, distance et τ) d'empilements de bases dans la structure d'ARNt^{Phe}.

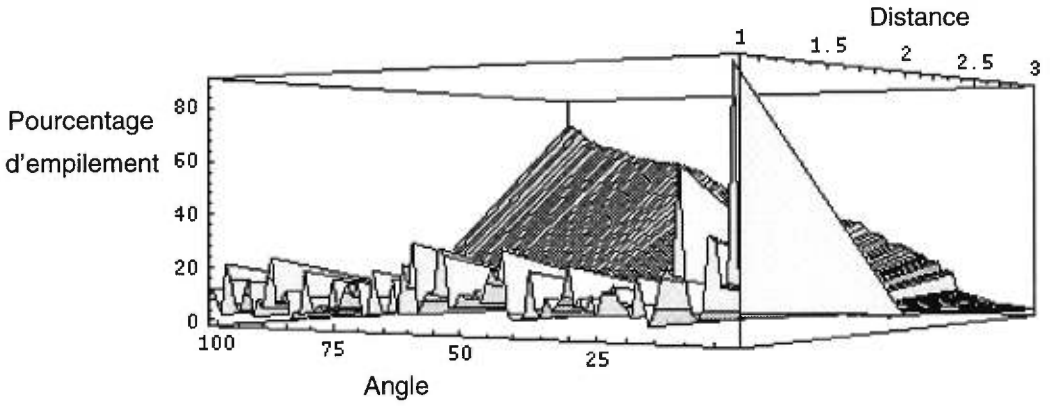


Figure 2.23: Représentation tridimensionnelle de paramètres d'empilements de bases présents dans la structure d'ARNt^{Phe}. Les paramètres sont l'angle, la distance et le taux de l'empilement entre deux bases suivant les axes X, Y et Z respectivement.

La figure 2.24 décrit tous les empilements présents dans la structure tridimensionnelle de la molécule d'ARNt^{Phe}. Ainsi, cette description complète et corrige celle de

Rich et RajBhandary [38]. Tous les empilements significatifs ont un $\tau \geq 20\%$, les cas ayant un τ inférieur mais proche de 20 sont aussi intéressants à analyser et à étudier comme cas limite. Ils peuvent aussi bien servir dans la modélisation locale d'une structure moléculaire, ce qui peut enrichir et diversifier les solutions générées par une telle modélisation.

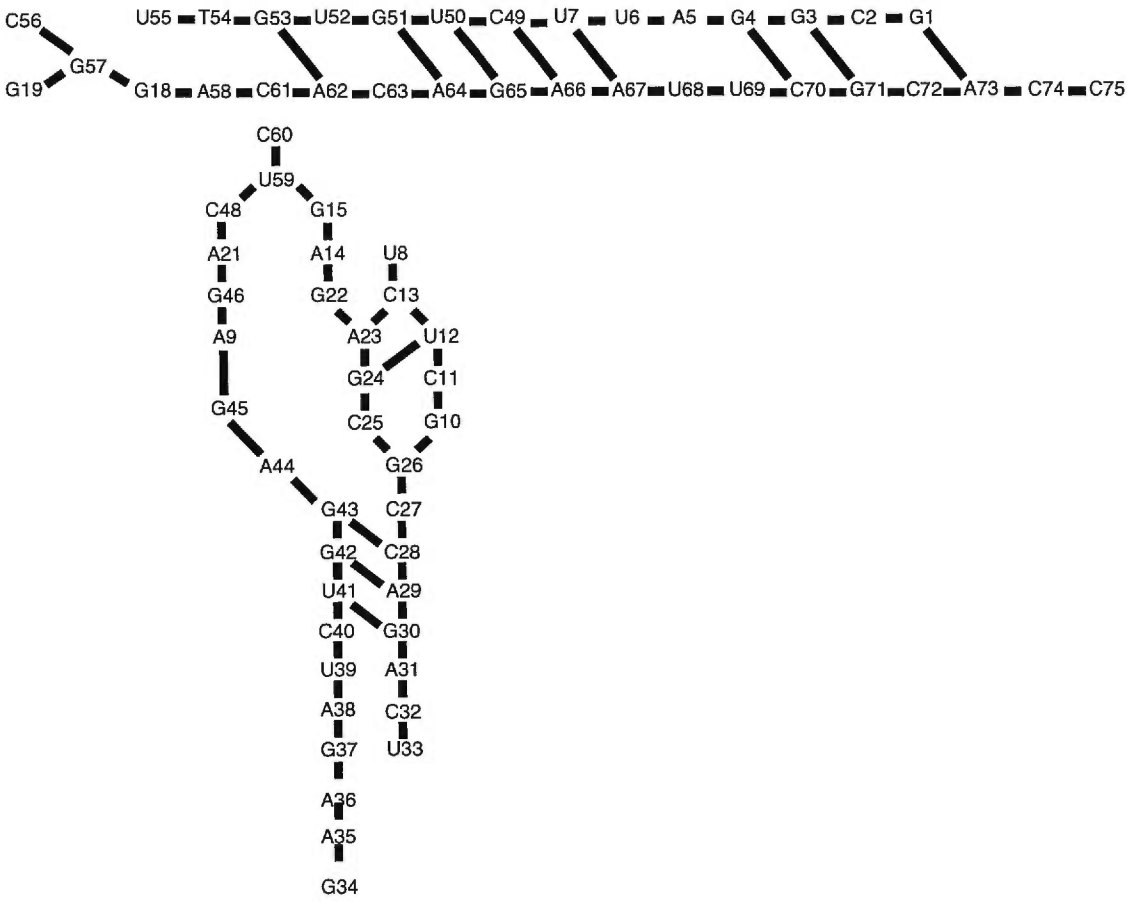


Figure 2.24: Diagramme montrant tous les empilements identifiés par notre méthode entre les différents nucléotides dans l'ARNt^{Phe} de la levure. L'empilement entre deux bases est représenté par un trait foncé. On remarque que cette structure est en majorité décrite à l'aide de l'empilement. Cette description se distingue de celle de Rich et RajBhandary [38] par la détection des empilements croisés en plus du cas de l'empilement des bases adjacentes C_{56}/G_{57}, par l'élimination de l'empilement entre les deux bases G_{26} et G_{43} et par l'absence du concept d'empilement partiel.

2.3.2 Les empilements dans la structure du hammerhead

Le mot "*hammerhead*" en anglais veut dire tête de marteau. Ce nom a été donné à un petit ribozyme catalytique ARN, formé par trois tiges de paires de bases et par un corps constitué de 15 nucléotides ayant la propriété d'être grandement conservé et ne présentant aucune complémentarité de séquences (voir la figure 2.25).

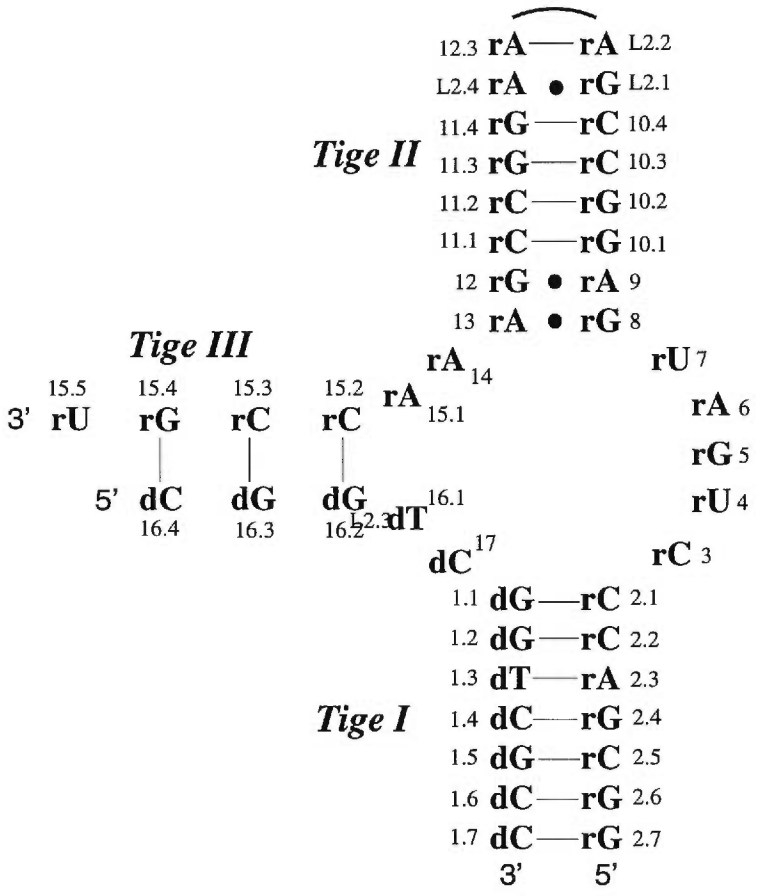


Figure 2.25: Structure secondaire du ribozyme ARN-ADN hammerhead. Les nucléotides de type ARN et ADN sont précédés respectivement par les lettres "r" et "d". Les traits désignent les appariements de bases de type Watson-Crick et les boules noires montrent les appariements non Watson-Crick.

La structure cristallographique par rayons X de l'inhibiteur ribozyme ARN-ADN hammerhead donnée avec 2.6Å de résolution par H. Pley et al. [36] est montrée dans

la figure 2.26. Le hammerhead a la capacité de se lier à une autre molélcule en formant une double hélice à l'aide de deux régions complémentaires et de couper cette autre molécule en un site précis entre les deux régions complémentaires. La spécificité de reconnaissance d'une autre molécule par le ribozyme peut être modifiée, en gardant l'activité catalytique du corps et en changeant un certain nombre de nucléotides impliqués dans les régions complémentaires. On peut cibler ainsi différentes molécules qui sont d'intérêt à subir la catalyse.

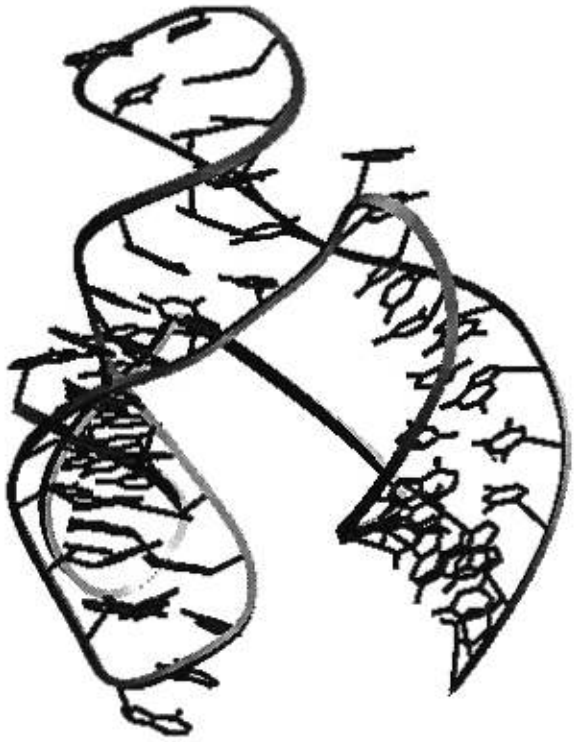


Figure 2.26: Structure tridimensionnelle du ribozyme ARN-ADN hammerhead [36].

On étudiera la structure du hammerhead de H. Pley publiée dans [36]. Il faut mentionner qu'il existe une autre structure par analyses cristallographiques du hammerhead donnée par W. Scott dans [42]. Cette dernière est presque identique à la première. Un empilement est signicatif si son taux d'empilement est supérieur 20% ($\tau > 20\%$). La table 2.7 montre les empilements avec un tel taux. L'analyse des

empilements présents entre les bases dans le hammerhead montre qu'il y a une seule base, U_7, qui n'est pas impliquée dans ce genre d'interaction verticale d'une manière significative. Elle présente une interaction de 2.927% et 17.995% respectivement avec $A_{15.1}$ et G_8, l'angle entre les deux bases U_7 et G_8 est 43.561 degrés; il est le plus grand angle enregistré entre les bases impliquées deux-à-deux dans un empilement. On présente dans la table 2.6 tous les empilements considérés comme non-significatifs, le taux d'empilement est alors inférieur à 20%. On note aussi seulement cinq empilements croisés qui sont tous présents au sein de la tige II (voir la structure secondaire montrée dans la figure 2.25): G_8 et G_{12} avec $\tau = 59.545\%$, A_9 et A_{13} avec $\tau = 73.05\%$ où les deux bases sont presque parallèles présentant un angle de 2.188 degrés, $G_{10.2}$ et $C_{11.1}$ avec $\tau = 29.907\%$, $C_{10.4}$ et $G_{11.3}$ avec $\tau = 37.752\%$ et finalement $G_{L2.1}$ et $G_{11.4}$ avec $\tau = 40.623\%$.

résidus	angle	distance	d(C1'$_1$,C1'$_2$)	τ_{12}	τ_{21}	τ
G_{12} / A_{13}	7.238	3.490	6.525	0.593	1.653	1.123
$C_{10.3}$ / $C_{11.2}$	11.258	2.481	7.499	0.909	1.483	1.196
$G_{L2.1}$ / $A_{L2.3}$	24.172	2.992	7.286	3.519	0.840	2.180
U_7 / $A_{15.1}$	19.175	2.892	7.263	4.312	1.542	2.927
G_8 / A_{14}	25.957	3.721	8.132	0.181	21.264	10.722
$A_{L2.4}$ / $G_{11.4}$	14.502	2.958	4.660	12.402	12.046	12.224
$G_{10.1}$ / G_{12}	2.486	3.102	7.847	17.280	17.264	17.272
U_7 / G_8	43.561	2.462	5.998	20.065	15.924	17.995

Table 2.6: Résultat de tous les empilements non significatifs présents dans le hammerhead, avec un taux d'empilement $\tau < 20$.

Tous les empilements considérés dans la table 2.7 ont une distance entre 2.882 (distance entre $G_{10.2}$ et $C_{11.1}$) et 3.751Å (distance entre les bases $C_{1.6}$ et $C_{1.7}$). L'angle varie graduellement entre 1.998 (angle entre les bases $G_{10.1}$ et $G_{10.2}$) et 19.879 degrés (angle entre les bases $C_{1.4}$ et $G_{1.5}$) pour grimper directement à la valeur 43.561 degrés qui est calculée pour les bases U_7 et G_8.

On remarque aussi que le long de la séquence des nucléotides de type ARN, il y a six endroits où des bases adjacentes dans la séquence ne présentent pas d'empilement, dont trois A_6/U_7, G_8/A_9 et $G_{L2.1}/A_{L2.2}$ ont un taux d'empilement nul et U_7/G_8, $A_{L2.4}/G_{11.4}$ et G_{12}/A_{13} qui possèdent respectivement 17.995, 12.224 et 1.123% comme taux d'empilement non-significatif. Cependant, pour la séquence de nucléotides de type ADN, il y a uniquement $T_{16.1}/C_{17}$ qui ne présentent pas d'empilement, le taux de ce dernier étant nul.

Les bases impliquées dans la formation des différentes hélices sont empilées les unes sur les autres. Certaines bases du corps central le sont aussi. La base U_7 est la seule base qui n'est pas impliquée dans aucun empilement significatif. Les trois bases $A_{L2.2}$, $A_{L2.3}$ et $A_{L2.4}$ sont fortement empilées en comparaison avec d'autres triplets. Cinq empilements croisés ont été identifiés, tous dans la tige II. La figure 2.27 résume l'analyse complète des empilements de bases dans la structure du hammerhead, ainsi que les différents appariements. On remarque comme dans le cas de l'ARN^{Phe} que la structure du hammerhead est en majorité décrite par l'empilement.

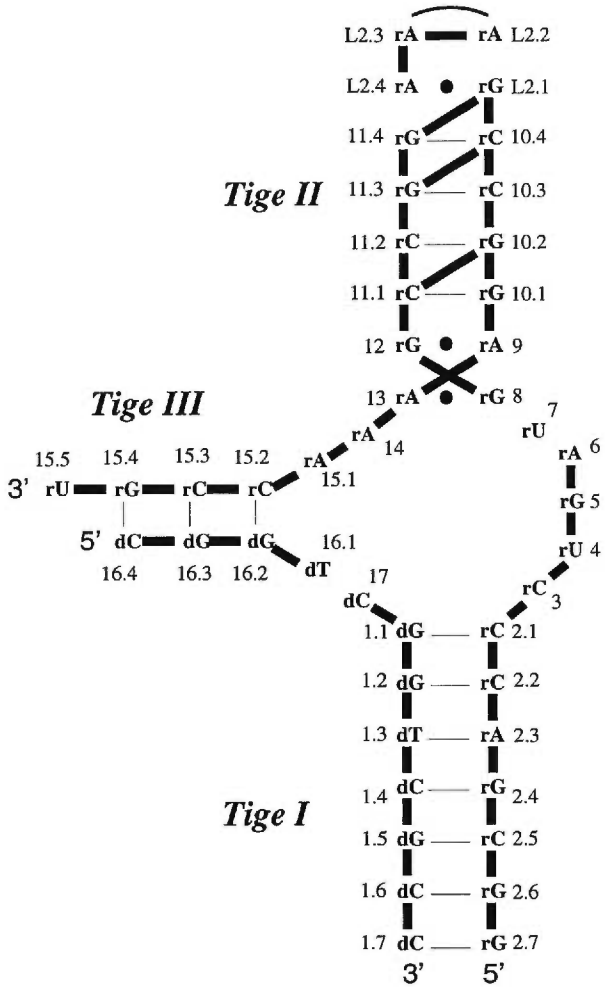


Figure 2.27: Structure secondaire du ribozyme ARN-ADN hammerhead. Les segments foncés désignent la présence d'empilement de bases. Les segments fins montrent les appariements de bases Watson-Crick, par contre les boules noires montrent les appariements non Watson-Crick. L'arc entre les bases $A_{L2.2}$ et $A_{L2.3}$ représente une boucle formée par douze nucléotides.

résidus	angle (degrés)	distance (Å)	d(C1'$_1$,C1'$_2$) (Å)	τ_{12} (%)	τ_{21} (%)	τ (%)
$C_{1.4}$ / $G_{1.5}$	19.879	3.306	5.650	25.452	27.008	26.230
$G_{10.2}$ / $C_{11.1}$	7.395	2.882	7.756	25.715	34.099	29.907
A_9 / $G_{10.1}$	6.438	3.406	5.008	32.960	28.507	30.734
$C_{16.4}$ / $G_{16.3}$	18.350	3.617	5.522	35.273	32.703	33.988
$C_{15.3}$ / $G_{15.4}$	12.851	3.020	5.219	42.419	26.816	34.617
$C_{2.5}$ / $G_{2.4}$	8.466	3.491	5.326	39.364	34.077	36.721
$C_{10.4}$ / $G_{11.3}$	3.476	3.286	7.649	43.273	32.230	37.752
$C_{11.2}$ / $C_{11.1}$	11.637	3.694	5.702	37.396	42.396	39.896
$G_{L2.1}$ / $G_{11.4}$	5.686	3.500	7.947	38.805	42.440	40.623
$C_{10.4}$ / $G_{L2.1}$	10.487	3.514	5.582	47.359	36.068	41.713
$C_{2.2}$ / $C_{2.1}$	8.207	3.537	5.380	41.555	45.810	43.683
C_3 / U_4	12.899	3.571	5.338	47.837	44.316	46.077
$C_{10.3}$ / $C_{10.4}$	12.375	3.707	5.393	44.975	52.454	48.714
C_{17} / $G_{1.1}$	14.700	3.599	5.589	55.826	42.589	49.207
$C_{15.2}$ / $C_{15.3}$	6.503	3.480	5.456	47.958	51.183	49.571
$C_{1.6}$ / $C_{1.7}$	13.413	3.751	5.629	49.043	51.498	50.271
$C_{11.1}$ / G_{12}	11.009	3.552	5.163	61.953	44.637	53.295
$G_{1.1}$ / $G_{1.2}$	6.630	3.328	4.948	50.367	57.594	53.981
$G_{11.4}$ / $G_{11.3}$	5.175	3.335	5.626	54.160	56.596	55.378
$G_{2.4}$ / $A_{2.3}$	9.161	3.439	5.530	57.139	54.530	55.834
A_{13} / A_{14}	2.805	3.418	5.007	58.321	54.946	56.633
$G_{2.7}$ / $G_{2.6}$	2.925	3.350	5.469	58.192	56.940	57.566
$G_{16.2}$ / $T_{16.1}$	11.295	3.502	5.076	48.553	69.892	59.223
G_8 / G_{12}	17.980	3.461	7.680	62.073	57.018	59.545
A_{14} / $A_{15.1}$	7.950	3.447	5.367	56.608	62.618	59.613
$G_{10.1}$ / $G_{10.2}$	1.998	3.461	4.973	61.108	59.518	60.313
$G_{10.2}$ / $C_{10.3}$	8.398	3.447	5.795	54.053	69.826	61.940
$G_{1.2}$ / $T_{1.3}$	7.804	3.467	5.432	58.592	70.015	64.303
$T_{1.3}$ / $C_{1.4}$	12.007	3.416	4.816	72.664	58.402	65.533
$G_{15.4}$ / $U_{15.5}$	14.864	3.550	5.371	69.584	63.080	66.332
$G_{16.3}$ / $G_{16.2}$	5.687	3.504	5.520	68.005	68.483	68.244
$C_{2.1}$ / C_3	6.099	3.598	5.086	72.995	64.195	68.595
$A_{15.1}$ / $C_{15.2}$	6.885	3.515	5.288	60.208	79.574	69.891
$G_{11.3}$ / $C_{11.2}$	9.782	3.398	5.231	66.562	74.341	70.451
$A_{L2.2}$ / $A_{L2.3}$	10.908	3.560	5.013	72.704	68.645	70.675
$A_{L2.3}$ / $A_{L2.4}$	5.013	3.576	4.995	73.257	71.315	72.286
A_9 / A_{13}	2.188	3.382	6.809	73.124	72.976	73.050
G_5 / A_6	14.332	3.529	5.422	75.271	72.545	73.908
$G_{1.5}$ / $C_{1.6}$	8.599	3.517	5.245	66.248	85.657	75.953
$A_{2.3}$ / $C_{2.2}$	9.934	3.656	5.170	66.168	86.071	76.119
$G_{2.6}$ / $C_{2.5}$	12.634	3.602	5.359	67.562	87.659	77.610

Table 2.7: Les empilements présents entre les bases deux-à-deux dans le hammerhead. On montre les différents paramètres, angle, distance, d(C1'$_1$, C1'$_2$), τ_{12}, τ_{21} et τ qui relient deux bases 1 et 2 représentant un empilement significatif (i.e. $\tau > 20$).

Chapitre 3

Modèle

3.1 Introduction

Dans ce chapitre, nous verrons la modélisation d'une molécule appelée *leadzyme* qui est une enzyme d'ARN (voir figure 3.1). Le but de la modélisation est de prédire la structure tridimensionnelle de cette molécule. Pour cela, nous nous sommes basé sur MC-SYM (modélisation faite dans [25]), pour gérer toutes les contraintes structurales disponibles. La complexité de cette modélisation est que les données struc-

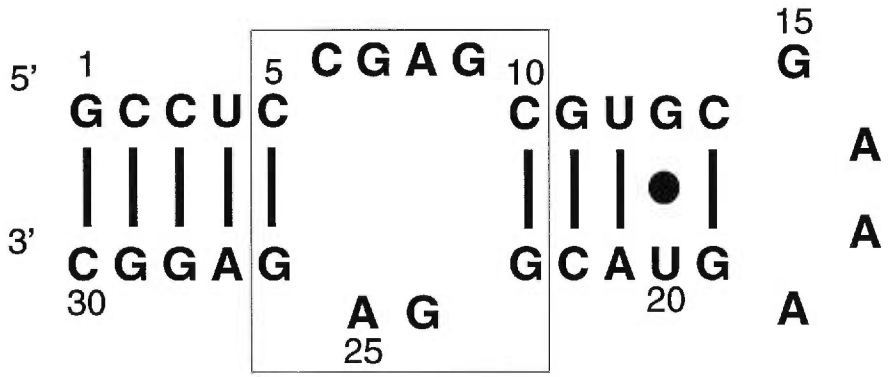


Figure 3.1: Structure primaire et secondaire du leadzyme. Les segments foncés désignent les appariements de bases de type Watson-Crick. La boule noire montre un appariement qui n'est pas de type Watson-Crick. La région encadrée indique le corps central comprenant le site de clivage entre C_6 et G_7.

turales permettent de générer plusieurs classes de conformations cohérentes; il faut donc déterminer la partie de l'espace conformationnel qui est susceptible de contenir la

structure désirée. Nous avons proposé un modèle mathématique basé sur la théorie de la logique floue [58] (suivant les informations que nous avons sur les différentes classes de l'espace conformationnel), dont on rappellera ultérieurement les notions de bases pertinentes. Par la suite, nous identifierons la partie conformationnelle recherchée à laquelle on associera un indice de préférence permettant de distinguer les différentes parties de l'espace conformationnel. La technique proposée est essentiellement l'intersection qu'on appellera *conformationnelle* entre toutes les conformations du leadzyme[1] et de ses huit séquences actives analogues[2], qui ont été isolées par la sélection *in vitro* [34]. Ceci a permis d'identifier les conformations communes pour toutes les structures générées par MC-SYM.

3.2 Modélisation du leadzyme à l'aide de MC-SYM

Aucune contrainte réelle n'est disponible pour la boucle intérieure composée de quatre nucléotides sur un coté et deux sur l'autre coté (voir figure 3.1). Par contre, moyennant les structures connues et l'application de notions thermodynamiques simples, on suggère que les empilements de bases et les appariements sont les caractéristiques dominantes dans de telles boucles. Ensuite, à cause de l'asymétrie de cette boucle, seulement deux appariements de bases (qui ne sont pas de type Watson-Crick) sont possibles. Ceci donne alors six hypothèses (voir figure 3.2):

(H1) $C_6 \bullet A_{25}$, $G_7 \bullet G_{24}$;

(H2) $C_6 \bullet A_{25}$, $A_8 \bullet G_{24}$;

(H3) $C_6 \bullet A_{25}$, $G_9 \bullet G_{24}$;

[1] La séquence initiale du leadzyme est appelée séquence *primaire* ou *de type sauvage.*

[2] Ce sont des séquences obtenues à partir de la séquence primaire en modifiant ses bases azotées et en s'assurant de maintenir partiellement ou totalement son activité (sa fonction biologique).

(H4) $G_7 \bullet A_{25}$, $A_8 \bullet G_{24}$;

(H5) $G_7 \bullet A_{25}$, $G_9 \bullet G_{24}$;

(H6) $A_8 \bullet A_{25}$, $G_9 \bullet G_{24}$.

Les liens hydrogènes impliqués dans chaque paire de bases vont être implémentés suivant la description de Saenger [41]

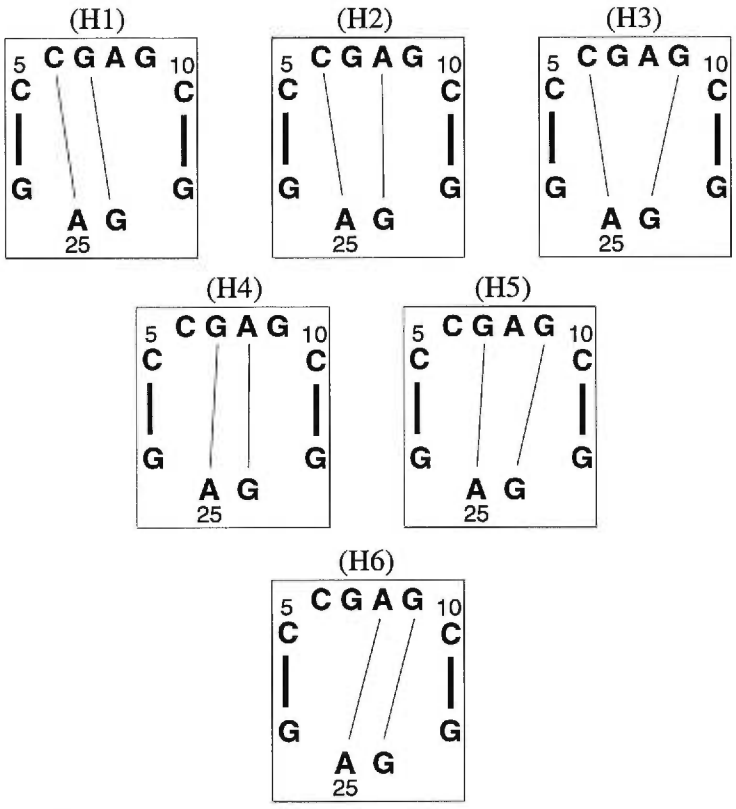


Figure 3.2: Les six hypothèses structurales décrivant l'appariement de bases dans la boucle intérieure du leadzyme. Les appariements ne sont pas de types Watson-Crick et ils sont indiqués par des segments étroits.

En premier lieu, à l'aide de MC-SYM, on procède à la génération de toutes les conformations de la séquence primitive. En un deuxième temps, on donne une évaluation à chacune des conformations suivant la capacité des séquence variantes d'être repliées dans cette conformation. Une valeur booléene est déterminée en changeant l'informa-

tion structurale décrite pour la séquence primitive dans MC-SYM pour qu'elle soit conforme à celle de la séquence variante donnée [25]. Par la suite, MC-SYM est appliqué de nouveau pour identifier les relations spatiales entre chacune des molécules. Les coordonnées atomiques de la plus part des nucléotides constituant les structures modélisées ont été générées à l'aide des relations spatiales existant entre les bases azotées des ARNs tirés des banques de données publiques. Ces relations spatiales sont l'empilement et l'adjacence entre les bases azotées. Elles sont représentées par des matrices de transformations homogènes [35]. Un repère local d'un nucléotide, A, est représenté par une matrice, R_A, déterminée par les coordonnées de trois atomes de A. Ainsi, la relation spatiale entre deux nucléotides, A et B, est une matrice homogène, $T_{AB} = R_A^{-1} R_B$, qui peut être tirée d'une structure tridimensionnelle connue et utilisée pour construire la structure à modéliser.

Par suite, un fichier a été formé et soumis comme entrée au programme MC-SYM et cela pour chaque hypothèse structurale (i.e. pour chacune des séquences variantes). Dans ce travail, il y avait plus de 14000 structures générées qui respectent les hypothèses H2, H3, H4, et H5. Cependant, aucune structure n'a été trouvée vérifiant les hypothèses H1 et H6. Dans une modélisation de ce genre, pour analyser les structures générées, on utilise généralement une analyse basée sur les méthodes interactives (qu'offrent certains programmes graphiques, par exemple InsightII [19]), ou sur la déviation **rmsd**[3] (*root-mean-square deviation*) entre les conformations de la même molécule prises deux-à-deux.

[3]La rmsd définit une mesure de ressemblance entre les conformations générées, c'est une distance sur l'espace des conformations, elle est donnée par la formule suivante:

$$rmsd(C_1, C_2) = \left[\left(\frac{2}{n(n-1)} \right)^2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(\left\| p_i^{C_1} - p_j^{C_1} \right\| - \left\| p_i^{C_2} - p_j^{C_2} \right\| \right)^2 \right]^{1/2} ;$$

où C_1 et C_2 sont deux conformations ayant n atomes dont les coordonnées tridimensionnelles sont respectivement p_i^{C1} et p_i^{C2}, $1 \leq i \leq n$.

Etant donné le grand nombre des structures, l'utilisation des approches mentionnées pour une telle analyse est complexe. Nous citerons dans ce qui suit les grandes étapes de la modélisation suivie dans [25]. Dans ce travail, les auteurs ont procédé moyennant un nouveau modèle de classification pour chaque hypothèse. Pour cela, MC-SYM a été alors modifié pour qu'à la sortie il donne, en utilisant les standards de Saenger [41], les informations structurales des paires de bases, la classification standard des liens angulaires présents dans les sucres des bases et l'orientation des bases par rapport au lien glucosile.

En se basant sur le principe de l'intersection conformationnelle pour identifier les conformations de la séquence primitive du leadzyme qui sont équivalentes avec celles des séquences variantes[4] actives, les fichiers d'entrée de MC-SYM associés à la séquence primitive ont été modifiés pour décrire les séquences variantes actives (voir [25]). Le résultat obtenu était que H3 est le seul modèle compatible à toutes les séquences variantes. Ceci résulte avec environ 2268 modèles générés par MC-SYM qui sont susceptibles d'engendrer les conformations actives sous l'hypothèse H3. Les modèles ainsi déterminés ont été classés en deux classes. Chaque classe a été subdivisée en deux sous-classes ayant la propriété suivante: la rmsd entre deux modèles appartenant à une même sous-classe est inférieur à 2.0Å [25]. En bas de cette valeur, la variation entre les modèles est négligeable.

L'étape suivante consistait à raffiner les structures tridimensionnelles générées par MC-SYM, le programme Discover, version 2.95, simulant la mécanique moléculaire, a été utilisé pour sélectionner les champs de forces [53, 54]. Pour la visualisation interactive des structures tridimensionnelles, on a utilisé le programme graphique moléculaire InsightII, version 2.3.

[4] Analogues (voir la note en bas de la page 65).

Pour la boucle **GAAA** (la boucle tétranucléotidique) du leadzyme, les auteurs de [25] ont basé leur modélisation sur les caractéristiques structurales de la structure solution associée à une classe de boucles tétranucléotidiques, **GNRA**, étudiées par Heus et Paradis dans [18]. Pour modéliser la boucle GAAA à l'aide de MC-SYM, les données structurales concernant l'empilement et les motifs d'appariement de bases A•G, tous les deux responsables de la stabilisation de la boucle intérieure, ont été tirées des informations des thermodynamiques d'ARN [5, 47, 49, 56]. L'orientation des bases autour du lien glucosile et les modes de conformation du sucre ont été prises des informations du spectre de résonance magnétique nucléaire **RMN** (méthode qui détermine les diverses propriétés spectrales d'une molécule; les spectres obtenus, correctement interprétés, donnent une foule de renseignements d'ordre structural) [22]. L'arbre de recherche, associé aux contraintes structurales de la boucle GAAA, que MC-SYM a exploré est de 28946 sommets. Seulement huit modèles ont été trouvés avec un maximum de la rmsd autour de 0.9Å, sans la considération des atomes d'hydrogène (pour plus de détails nous inviterons le lecteur à consulter [7, 25]).

La séquence primaire (ou de type sauvage) avec les huit autres séquences mutantes actives sont toutes susceptibles d'engendrer la structure tridimensionnelle cherchée. On associe alors une probabilité uniforme de $\frac{1}{9}$ à chacune des séquences. Les séquences seront nommées comme suit:

$$WT, A, B, C, D, E, F, G, H;$$

où WT est la séquence de type sauvage et les autres sont les séquences mutantes actives. La question qui se pose est comment donner une certaine mesure pour quantifier chacune des hypothèses H2, H3, H4 et H5 pour savoir laquelle des quatre hypothèses est la plus susceptible de contenir la structure tridimensionnelle désirée.

Cette mesure doit être basée sur un formalisme des résultats empiriques obtenus de la modélisation du leadzyme. Elle sera en quelque sorte un guide pour la prise de décision dans le domaine des conformations. Le fait de choisir une hypothèse au lieu d'une autre et de dire qu'elle est la plus susceptible de contenir la structure tridimensionnelle est à la fois imprécis et incertain dû au fait que les contraintes sont approximatives et imparfaites. La logique floue s'avère pertinente pour donner une mesure de préférence à chacune des hypothèses en plus de la modélisation imprécise [27]. Dans la section suivante, on rappellera les notions importantes de la logique floue qui vont s'appliquer à notre domaine d'hypothèse.

3.3 Mesure de la préférence des hypothèses à l'aide de la logique floue

3.3.1 Terminologie et notation pour le principe de l'incertitude

Les deux types d'imperfections, l'incertain et l'imprécis, dans les connaissances n'ont pas eu la même importance dans les préoccupations des scientifiques. Pour ce qui a rapport à l'incertain, il a été abordé par la notion de probabilité au XVII^{e} siècle par Pascal et Fermat. Cependant, celle-ci ne permet pas de traiter des croyances subjectives comme on a longtemps pensé qu'elle pouvait le faire, ni de résoudre le problème posé par les connaissances imprécises ou vagues. Ces dernières n'ont été prises en considération qu'à partir de 1965, lorsque L. A. Zadeh, connu internationalement pour ses travaux sur la théorie des systèmes, a introduit la notion de *ensemble flou* (en anglais: *fuzzy set*), à partir de l'appartenance partielle à une classe admettant des situations intermédiaires entre le tout et le rien. La *théorie des possibilités* qui a

été introduite par L. A. Zadeh en 1978 [58], constitue un cadre permettant de traiter des concepts d'incertitude de nature non probabiliste. On va la considérer à partir de la notion de l'ensemble flou ce qui va nous permettre d'exploiter, dans un même formalisme, imprécision et incertitude sur la préférence d'une hypothèse par rapport à une autre.

Mesures de possibilité et fonctions de croyance

Etant donné un ensemble de référence fini X, on attribue à chaque événement défini sur X, c'est-à-dire à tout sous-ensemble de X, un coefficient compris entre 0 et 1 évaluant à quel point cet événement est possible. Pour définir ce coefficient, on introduit une mesure de possibilité Π, qui est une fonction définie sur l'ensemble $\mathcal{P}(X)$ des parties de X, prenant ses valeurs dans $[0,1]$, telle que:

1. $\Pi(\emptyset) = 0, \quad \Pi(X) = 1,$

2. $\forall A_1 \in \mathcal{P}(X), \quad A_2 \in \mathcal{P}(X), \ldots$

$$\Pi(\bigcup_{i=1,2,\ldots} A_i) = \sup_{i=1,2,\ldots} \Pi(A_i),$$

où sup indique la plus grande valeur des valeurs concernées dans le cas fini.

Les fonctions de *croyance* concernent la modélisation et la quantification de la crédibilité attribuée à des faits. La *théorie de l'évidence* de Shafer [43] (voir aussi [21]) considère un univers de référence fini X sur lequel sont déterminés des coefficients de croyance, obtenus en répartissant une masse globale de croyance égale à 1 entre tous les événements possibles. Une masse m peut être définie comme suit

$$m : \mathcal{P}(X) \longrightarrow [0,1]$$

telle que

$$m(\emptyset) = 0 \quad \text{et} \quad \sum_{A \in \mathcal{P}(X)} m(A) = 1.$$

Pour chaque ensemble $A \in \mathcal{P}(X)$, la valeur $m(A)$ représente le degré avec lequel un groupe d'observateurs croit en la réalisation d'un événement par les éléments de A. Cette valeur, $m(A)$, concerne uniquement un seul ensemble, l'ensemble A; elle n'implique aucune information additionnelle pour les sous-ensembles de A. S'il y a une évidence additionnelle confirmant la réalisation du même événement dans un sous-ensemble de A, disons $B \subset A$, il faut l'exprimer par une autre valeur $m(B)$. Toute partie non-vide B de X telle que $m(B) \neq 0$ est appelée *élément focal.* Elle correspond à un événement auquel les observateurs croient, ne serait-ce qu'un peu. On définit alors le *degré* de croyance $Bel(A)$ (en anglais: *belief measure*) en une partie quelconque A de X en prenant en considération tous les éléments focaux qui entraînent A et cela s'exprime par la formule:

$$Bel(A) = \sum_{B \subseteq A} m(B),$$

et le *degré de plausibilité* de A (en anglais: *plausibility measure*) en prenant en compte tous les éléments focaux qui ont quelque chose à voir avec A:

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B).$$

Les deux mesures vérifient les relation suivantes:

$$Pl(A) = 1 - Bel(\bar{A}) \quad \text{et} \quad Bel(A) \leq Pl(A).$$

L'intervalle $[Bel(A), Pl(A)]$ encadre la probabilité mal connue $P(A)$ (probabilité imprécise [50, 15, 16]), quelque soit la partie A de X. Un cas particulier d'attribution de

la masse m est remarquable: supposons que les éléments focaux soient des singletons de X, les avis émis ne concernent que des événements élémentaires. Alors, toute partie A de X est telle que $Bel(A) = Pl(A)$ et cette valeur commune est également celle de la probabilité $P(A)$.

3.3.2 Les mesures de croyance des hypothèses du leadzyme

Dans le cas du leadzyme, considérons comme univers de référence l'ensemble fini X contenant toutes les conformations du leadzyme. Suite à la relation qui réside entre un ensemble des conformations tridmensionnelles et les hypothèses qu'elles vérifient (voir la figure 3.3), les éléments focaux sur lesquels on peut se baser sont:

$$WT, ADE, BF, CH \text{ et } G,$$

où $ADE = A \cap D \cap E$, $BF = B \cap F$, $CH = C \cap H$ et WT, A, B, C, D, E, F, G sont des sous-ensembles de X contenant successivement les conformations tridimensionnelles associées à la séquence du leadzyme de type sauvage WT et aux séquences mutantes actives A, B, C, D, E, F, et G. D'après les résultats obtenus par la modélisation du leadzyme, on a:

$$\begin{aligned} WT &= E_{\mathrm{H2}} \cup E_{\mathrm{H3}} \cup E_{\mathrm{H4}} \cup E_{\mathrm{H5}} \\ ADE &= E_{\mathrm{H3}} \\ BF &= E_{\mathrm{H2}} \cup E_{H3} \cup E_{H4} \\ CH &= E_{\mathrm{H2}} \cup E_{\mathrm{H3}} \\ G &= E_{\mathrm{H2}} \cup E_{\mathrm{H3}} \cup E_{\mathrm{H5}}, \end{aligned}$$

où $E_{\mathrm{H}i}$, $i = 2, 3, 4, 5$, représente l'ensemble des conformations tridimensionnelles satisfaisant l'hypothèse Hi. Remarquons que $X = WT$, puisque pour chaque hypothèse Hi, il existe des conformations dans X vérifiant cette dernière.

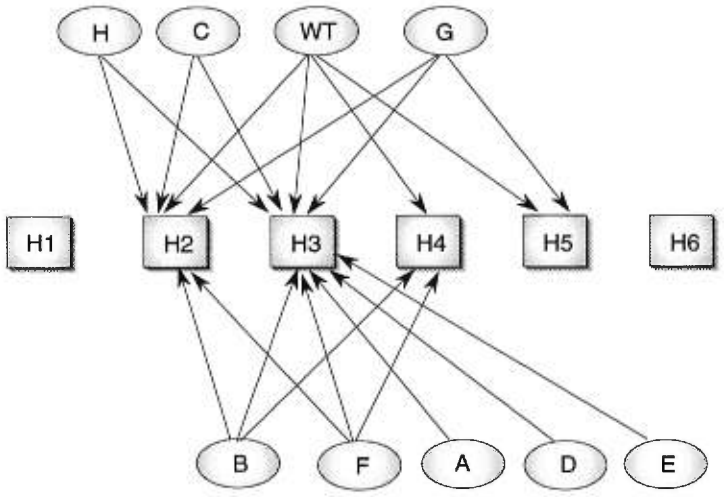


Figure 3.3: La relation définissant les éléments focaux WT, ADE, BF, CH et G.

Nous attribuons à chaque élément focal un coefficient de croyance en sa possibilité de contenir la structure tridimensionnelle que le leadzyme adopte dans la nature par:

$$m(WT) = m(G) = \frac{1}{9}, \quad m(BF) = m(CH) = \frac{2}{9} \text{ et } \quad m(ADE) = \frac{1}{3}.$$

Les masses ont été attribuées en se basant sur le fait qu'au départ on a neuf séquences (la séquence de type sauvage et les huit séquences mutantes actives) pour exprimer l'ignorance sur le fait qu'on ne peut dire laquelle est susceptible d'être associée à la structure du leadzyme, la distribution des possibilités, π, est alors $\forall x \in X \;\; \pi(x) = 1$, ce qui est équivalent dans le cas de la distribution des probabilités, p, à $\forall x \in X \;\; p(x) = \frac{1}{|X|}$. Mais, étant donné la supposition biochimique qui donne plus de croyance à l'ensemble de conformations satisfaisant certaines hypothèses associées au plus de séquences en question, on a attribué alors, par exemple, à ADE un coefficient de croyance $m(ADE) = \frac{3}{9} = \frac{1}{3}$, en se basant sur la fait qu'à partir des trois séquences A, D, et E parmi l'ensemble des neuf séquences, on a généré des conformations vérifiant

ensemble l'hypothèse H3. La figure 3.4 nous montre les éléments focaux dans l'univers de référence $X = WT$. Ainsi, on est en mesure maintenant de définir nos intervalles encadrant la probabilité imprécise pour chaque partie de l'espace conformationnel définie par une hypothèse donnée.

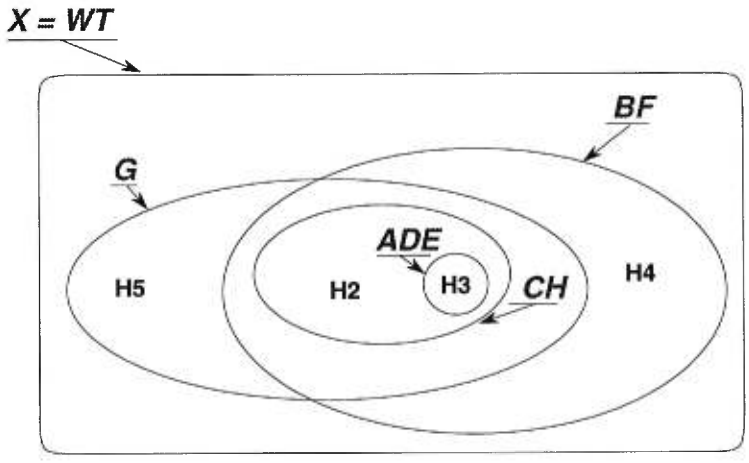


Figure 3.4: Les éléments focaux, WT, ADE, BF, CH et G, sont montrés avec les hypothèses que leurs conformations vérifient.

On se trouve alors dans la situation générale de définition des masses et on est en mesure de déduire l'intervalle de la probabilité mal connue sur chacun des ensembles de conformations associé à chacune des hypothèses à savoir lequel est susceptible de contenir la structure réelle du leadzyme. On obtient donc les degrés de croyance:

$$Bel(E_{\mathrm{H2}}) = Bel(E_{\mathrm{H4}}) = Bel(E_{\mathrm{H5}}) = 0,$$

$$Bel(E_{\mathrm{H3}}) = \sum_{B \subseteq E_{\mathrm{H3}}} m(B) = m(ADE) = \frac{1}{3};$$

et les degrés de plausibilité associés aux ensembles vérifiant chacune des hypothèses sont:

$$Pl(E_{\mathrm{H2}}) = \sum_{B \cap E_{\mathrm{H2}} \neq \emptyset} m(B) = m(WT) + m(BF) + m(CH) + m(G) = \frac{2}{3}$$

$$
\begin{aligned}
Pl(E_{\mathrm{H3}}) &= \sum_{B \cap E_{\mathrm{H3}} \neq \emptyset} m(B) \\
&= m(WT) + m(ADE) + m(BF) + m(CH) + m(G) = 1 \\
Pl(E_{\mathrm{H4}}) &= \sum_{B \cap E_{\mathrm{H4}} \neq \emptyset} m(B) = m(WT) + m(BF) = \frac{1}{3} \\
Pl(E_{\mathrm{H5}}) &= \sum_{B \cap E_{\mathrm{H5}} \neq \emptyset} m(B) = m(WT) + m(G) = \frac{2}{9}.
\end{aligned}
$$

Ceci se résume dans la table 3.1.

A	$Bel(A)$	$Pl(A)$
E_{H2}	0	2/3
E_{H3}	1/3	1
E_{H4}	0	1/3
E_{H5}	0	2/9

Table 3.1: Degrés de croyance et de plausibilité associés à chacun des ensembles $E_{\mathrm{H2}}, E_{\mathrm{H3}}, E_{\mathrm{H4}}$ et E_{H5} (vérifiant successivement les hypothèses H2, H3, H4 et H3 considérées pour la modélisation du leadzyme) mesurant l'intervalle de probabilité mal connue ou imprécise pour qu'une hypothèse soit susceptible d'être adoptée par la structure naturelle du leadzyme.

Les conformations sous l'hypothèse H3 ont été testées de nouveau expérimentalement au laboratoire, voir [7]. Il a été trouvé que la paire $G_9 \bullet G_{24}$ a été correctement prédite; par contre la paire $C_{26} \bullet A_{25}$ ne présentait aucune importance particulière pour l'activité catalytique du leadzyme étant donné que l'élimination des groupes chimiques du nucléotide A_{25} n'ont pas affecté le clivage [7]. Cependant, les groupes chimiques impliqués dans les interactions Watson-Crick du C_6 ont été trouvés importants pour l'activité [7]. Ces résultats ont suggéré ainsi de chercher un autre partenaire pour le C_6 pour former un appariement. On a alors trouvé un appariement de bases particulier $G_9 \bullet G_{24}$ qui permettait d'avoir des groupes chimiques du nucléotide G_{24} libres pour former un appariement avec le C_6. Ainsi, une nouvelle hypothèse H7 a été suggérée

formant ainsi un triplet comme le montre la figure 3.5.

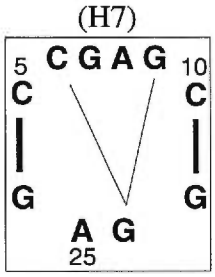


Figure 3.5: L'hypothèse structurale H7 suivant [7]. H7 a été obtenue à partir de H3 en éliminant l'appariement entre C_6 et A_{25} et en appariant C_6 avec G_{24}[7].

Cette hypothèse structurale a généré deux classes de modèles tridimensionnels, l'une des deux satisfait toutes les données structurales courantes valables (pour plus de détails voir [25]). Il faut noter que le leadzyme est un cas très simple pour l'application du modèle d'incertitude présenté. Le choix de l'hypothèse de H3 pouvait être déterminé directement de la figure 3.4. Par contre, la théorie sera utile dans d'autre cas plus complexe, où chaque étape de la modélisation nécessite un choix incertain suivant les contraintes approximatives imprécises disponibles.

Conclusion

Nous avons présenté une approche flexible pour trouver les empilements de bases dans les acides nucléiques. Notre algorithme n'utilise aucune hypothèse de préférence pour les positions relatives à une base, il ne se base pas non plus sur des fonctions d'énergie empiriques pour déterminer quelles bases sont empilées. Notre approche est strictement géométrique et elle est plus avantageuse étant donnée que les interactions électrostatiques et hydrophobiques responsables d'empilement sont jusqu'à maintenant mal définies.

Deux molécules classiques, l'ARNt et le hammerhead, ont été étudiées pour montrer la validité et la flexibilité de notre approche. Pour la première fois cette méthode nous a permis d'analyser des paramètres associées à l'empilement et d'identifier des empilements au sein d'une molécule qui ont été négligés par les méthodes de visualisation où celles se basant sur des fonctions d'énergie potentielles. Un nouveau paramètre, τ permet d'évaluer chacun des empilements présents dans une molécule d'acide nucléique, l'utilisateur pourra ainsi faire son choix sur la qualité des différents empilements où d'analyser les cas limites d'empilement plus facilement avec les différents paramètres disponibles. On a obtenu de bons résultats sur la distance d_{ij}, l'angle α_{ij} entre deux bases i et j empilées, et le paramètre τ à partir duquel on peut identifier un empilement significatif. A cet égard, dans un article à venir [12], nous allons analyser à l'aide de cet algorithme toute la banque de données des acides nucléiques, afin de définir d'une manière générale l'intervalle des valeurs des paramètres qui serviront comme propriétés importantes de l'empilement entre deux bases azotées. Cet algorithme de recherche d'empilement avec celui de l'identification des appariements forment le programme NANOTE, qui consistera à analyser les structures tridimen-

sionnelles des molécules d'acide nucléique. Le programme prend comme entrée un fichier de format PDB pour une certaine structure moléculaire et génère un nouveau fichier PDB décrivant la structure secondaire, les appariements et les empilements de la structure [33]. Plus d'une soixantaine de structures ont été analysées dont le $ARNt^{Phe}$ et le hammerhead. Les résultats seront accessibles, en versions française et anglaise, aux sites WEB suivants:

http://www-lbit.iro.umontreal.ca/LIST/list_fr.htm

http://www-lbit.iro.umontreal.ca/LIST/list_en.htm

L'utilisateur est invité à soumettre ses propres fichiers PDB à fin d'analyser les structures qu'ils contiennent et à utiliser un grand nombre de données mis à sa disposition pour différents besoins d'analyse et de modélisation. L'utilisateur peut aussi à sa guise varier les différents paramètres d'appariement et d'empilement. Le programme NANOTE est un outil important pour la modélisation. Grâce aux résultats de l'-analyse structurale obtenus de ce dernier, nous pouvons maintenant capable maintenant identifier automatiquement des cas d'empilements servant à la construction d'un modèle moléculaire qui n'étaient pas traité auparavant à cause de la complexité du calcul ou par le choix aléatoire des empilements ou appariements sans aucune classification possible. Ceci est une contribution à la compréhension du lien qui existe entre la structure tridimensionnelle d'une molécule donnée et la fonction biologique qui la caractérise.

Glossaire

A

Acide

acide carboxylique (-COOH) ou molécule comportant un groupe acide carboxylique.

Acide désoxyribonucléique

ADN.

Acide ribonucléique

ARN.

ADN

Acide désoxyribonucléique. Molécule formée de deux polymères de désoxyribonucléosides monophosphates . Ces deux chaînes sont associées et enroulées en double hélice. L'ADN contient les gènes constituant le génome et il se trouve dans les chromosomes et la chromatine. Le 2-désoxyribose est le glucide distinctif de l'ADN. L' adénine , la guanine, la thymine et la cytosine sont les principales bases azotées entrant dans la structure de l'ADN.

Acides nucléiques

ADN (acide désoxyribonucléique) et ARN(acide ribonucléique)

Adénine

Base azotée de la famille des purines, entrant, entre autres, dans la structure des acides nucléiques.

Adénosine

Nucléoside, de la famille des purines, constitué de l'adénine unie au ribose.

ARN

Acide ribonucléique. Polymère de ribonucléosides monophosphates. Le ribose est le glucide distinctif de l'ARN. L'adénine, la guanine, l'uracile et la cytosine sont les principales bases azotées entrant dans la structure de l'ARN. On distingue l'ARN messager, l'ARN de transfert et l'ARN ribosomique.

ARN de transfert

ARNt. ARN auquel un acide aminé est lié, ce qui permet son transfert sur les ribosomes, et qui contient une séquence spécifique de bases, l'anticodon, reconnaissant et s'associant à une séquence spécifique de bases, le codon, de l'ARN messager lors de la synthèse protéique.

ARN messager

ARNm. ARN qui contient une longue séquence de bases constituant le code servant à la synthèse d'une protéine.

ARNm ARN messager.

ARNt

ARN de transfert.

B

Base azotée

Molécule comportant des atomes d'azote (N) et ayant des propriétés basiques. Il y a deux familles de bases azotées : les purines et les pyrimidines.

C

Catalyseur

Molécule qui, en petite quantité, accélère la vitesse d'une réaction et qui revient à sa forme initiale à la fin de la réaction. Les enzymes sont des catalyseurs biologiques

Complexe

Association de plusieurs molécules dans des proportions définies.

Cytidine

Nucléoside, de la famille des pyrimidines, constitué de la cytosine unie au ribose.

Cytosine

Base azotée de la famille des pyrimidines, entrant, entre autres, dans la structure des acides nucléiques.

D

Désoxyribonucléotide

Nucléotide formé par l'union d'une purine ou d'une pyrimidine et du 2-désoxyribose (ce qui constitue un désoxyribonucléoside), auquel sont liés un, deux ou trois groupes phosphates.

E

Enzyme

Protéine catalysant une réaction biochimique. Il y a six classes d'enzymes : les oxydoréductases, les transférases, les hydrolases, les lyases, les isomérases et les ligases.

G

Gène

Segment d'ADN servant à la synthèse d'un ARN ribosomique, d'un ARN de transfert ou, le plus souvent, d'un ARN messager codant une ou plusieurs protéines.

Génome

Ensemble du matériel génétique d'une personne. Il est constitué des gènes se trouvant dans les 46 chromosomes, soit 22 paires d'autosomes et une paire de chromosomes sexuels. Un chromosome de chaque paire (23) provient de l'ovule maternel et l'autre chromosome de chaque paire provient du spermatozoïde paternel. L'ADN se trouvant dans les mitochondries porte également quelques gènes constituant le génome mitochondrial.

Groupe

Ensemble fonctionnel d'atomes liés les uns aux autres d'une façon définie, formant une partie d'une molécule.

Groupe phosphate

Groupe (-PO42-) lié à une molécule organique.

Guanine

Base azotée de la famille des purines, entrant, entre autres, dans la structure des acides nucléiques.

Guanosine

Nucléoside, de la famille des purines, constitué de la guanine unie au ribose.

I

Inhibition

Diminution de l'activité d'une enzyme ou de la vitesse d'une voie métabolique.

Inosine

Nucléoside, de la famille des purines.

L

Liaison glycosidique (-C-O-C-)

Liaison entre deux groupes hydroxyle (-OH) de deux monosaccharides, formée par élimination d'une molécule d'eau (H2O).

Liaison peptidique (-CONH-)

Liaison entre le groupe amine (-NH2) d'un acide aminé et le groupe acide carboxylique (-COOH) d'un autre acide aminé. Elle est formée par élimination d'une molécule d'eau (H2O).

M

Mitochondrie

Organite délimité par deux membranes. Les mitochondries sont localisées dans le cytoplasme des cellules et elles jouent un rôle important dans le métabolisme énergétique cellulaire.

Molécule

Ensemble d'atomes unis les uns aux autres par des liaisons chimiques.

Monomère

Molécule libérée lors de l'hydrolyse d'un polymère, tels le glucose et les acides aminés. Egalement, autre terme désignant une sous-unité.

Mutation

Modification transmissible, spontanée ou provoquée, d'un gène.

N

Nucléoside

Molécule formée par l'union d'une purine ou d'une pyrimidine et d'un glucide. Il en existe deux types selon le glucide : les ribonucléosides et les désoxyribonucléosides. Les principaux sont l'adénosine, l'inosine et la guanosine (nucléosides puriques), la cytidine, la thymidine et l'uridine (nucléosides pyrimidiques).

Nucléotide

Molécule formée par l'union d'une purine ou d'une pyrimidine et d'un glucide (ce qui constitue un nucléoside) à laquelle sont liés un, deux ou trois groupes phosphates. Il en existe deux types selon le glucide : les ribonucléotides et les désoxyribonucléotides. On distingue également les nucléosides monophosphates, les nucléosides diphosphates et les nucléosides triphosphates.

P

Protéine

Polymère constitué d'une ou plusieurs chaînes d'acides aminés unis par des liaisons peptidiques. Les enzymes, les récepteurs, les transporteurs et certaines hormones sont des protéines. Donc, elles sont des constituants structuraux et fonctionnels très importants dans l'organisme.

Purine

Base azotée libre ou faisant partie de la structure de nucléosides, de nucléotides et des acides nucléiques. Les principales bases puriques sont l'adénine et la guanine.

Pyrimidine

Base azotée libre ou faisant partie de la structure de nucléosides, de nucléotides et des acides nucléiques. Les principales bases pyrimidiques sont l'uracile, la cytosine, et la thymine.

R

Résidu d'un acide aminé

Portion distinctive d'un acide aminé. Lorsqu'un acide aminé fait partie de la structure d'une protéine, c'est la portion de l'acide aminé qui ne participe pas à la liaison peptidique et qui demeure libre.

Ribonucléoside

Nucléoside formé par l'union d'une purine ou d'une pyrimidine et du ribose.

Ribonucléoside monophosphate

Nucléoside monophosphate formé par l'union d'une purine ou d'une pyrimidine, du ribose et d'un groupe phosphate.

Ribonucléotide

Nucléotide formé par l'union d'une purine ou d'une pyrimidine et du ribose (ce qui constitue un ribonucléoside), auquel sont liés un, deux ou trois groupes phosphates.

Ribose

Monosaccharide (ose) de la famille des pentoses. C'est un glucide le plus simple non hydrolysable.

Ribosome

Particule constituée d'ARN et de protéines, servant à la synthèse des protéines.

S

Site actif

Centre actif, site catalytique. Site particulier dans une molécule d'une enzyme où se lient les substrats et où se produit la réaction enzymatique.

Site catalytique

Site actif.

Substrat

Molécule qui, après s'être liée au site actif d'une enzyme, est transformée en un ou plusieurs produits.

T

Thymidine

Nucléoside, de la famille des pyrimidines, constitué de la thymine unie au 2-désoxyribose.

Thymine

Base azotée de la famille des pyrimidines, entrant, entre autres, dans la structure des acides nucléiques.

U

Uridine

Nucléoside, de la famille des pyrimidines, constitué de l'uracile unie au ribose.

Bibliographie

[1] Aida M.

An ab initio molecular orbital study on the sequence-dependency of DNA conformation: an evaluation of intra- and inter-strand stacking interaction energy. *J. Theor. Biol.* 1988, **130**: 327-335.

[2] Anfinsen C. B.

Principles that govern the folding of protein chains. *Science* 1973, **181**: 223-230.

[3] Babcock M. S., Pednault E. P. D., Olson W. K.

Nucleic acid structure analysis. Mathematics for local Cartesian and helical structure parameters that are truly comparable between structures. *J. Bio. Mol.* 1994, **237**: 125-156.

[4] Bondi A.

Van der Waals volumes and radii. *J. Phys. Chem.* 1964, **68**: 441-451.

[5] Cai Z., Tinoco I. Jr.

Solution structure of loop a from the hairpin ribozyme from tobacco ring spot virus satellite. *Biochemistry* 1996, **35**: 6026-6036.

[6] Cieplak P., Kollman, P. A.

Calculation of the free energy of association of nucleic acid bases in vacuo and water solution. *J. Am. Chem. Soc.* 1988, **110**: 3734-3739.

[7] Chartrand P., Usman N., Cedergren R.

The effect of structural modifications on the activity of the leadzyme. *Biochemistry* 1997, **34**: 3145-3150.

[8] Chothia C.

Structural invariants in protein folding. *Nature* 1975, **254**: 304-308.

[9] Dickerson R. E. Bansal M., Calladine C. R., Diekmann S., Hunter W. N., Kennard O., Von Kitzing E., Lavery R., Nelson H. C. M. Olson W. K. et al.

Definitions and nomenclature of nucleic acid structure parameters. *J. Mol. Biol.* 1989, **208**: 787-791.

[10] El Hassan M. A., Calladine C. R.

The assessment of the geometry of dinucleotide steps in double-helical DNA: a new local calculation scheme with an appendix. *J. Mol. Biol.* 1995, **251**: 648-664.

[11] Frechet D., Ehrlich R., Remmy P., Gabarro-Arpa

Thermal perturbation differential spectra of ribonucleic acids. II. Nearest neighbour interactions. *Nucleic Acids Res.* 19179, **7**: 1981-2001

[12] Ftouhi A., Major F.

Nucleic acid base stacking. En préparation.

[13] Gabb H. A., Sanghani S. R., Robert C. H., Prévost C.

Finding and visualizing nicleic acid base stacking. *J. Mol. Graphics* 1996, **14**: 6-11.

[14] Gorin A. A., Zhurkin V. B., Olson W. K.

B-DNA twisting correlates with base pair morphology. *J. Mol. Biol.* 1995, **247**: 34-48.

[15] Guan J. W., Bell D. A.

Evidence Theory and Its Applications, Vol. 1. North-Holland, New York, 1991.

[16] Guan J. W., Bell D. A.

Evidence Theory and Its Applications, Vol. 2. North-Holland, New York, 1992.

[17] Hunter C. A.

Sequence-dependent DNA structure. The role of base stacking interactions. *J. Mol. Biol.* 1993, 230, 1025-1054.

[18] Heus H. A., Paradi A.

Structural features that give rise to the unusual stability of RNA hairpins containing GNRA loops. *Science* 1991, **253**: 191-194.

[19] InsightII User Guide, San Diego. Biosym/MSI, October 1995.

[20] Kim S. H., Quigley G. J., Suddath F. L., McPherson A., Sneden D., Kim J. J., Weinzierl J., Rich A.

Three-dimensional structure of yeast phenylalanine transfer RNA: folding of the polynucleotide chain. *Science* 1973, **179**: 285-288.

[21] Klir G. T., Foger T.

Fuzzy sets, Uncertainty, and Information. Prentice-Hall, Englewood Cliffs, NJ, 1988.

[22] Laing L. G., Hall K. B.

A model of the iron responsive element RNA hairpin loop structure determined from NMR and thermodynamic data. *Biochemistry* 1996, **35**: 13586-13596.

[23] Langlet J., Claverie P., Caron F., Boeuve J. C.

Interactions between nucleic acid bases in hydrogen bonded and stacked configurations: the role of the molecular charge distribution. *Int. J. Quant. Chem.* 1981, **20**: 299-338.

[24] Ladner J. E, Jack A., Robertus J. D., Brown R. S., Rhodes D., Clark B. F. C., Klug A.

Nucleic Acids Res. 1975, **2**: 1929-1937.

[25] Lemieux S., Chartrand P., Cedergren R., Major F.

Modeling active RNA structures using the intersection of conformational space: application to the lead-activated ribozyme (Soumis à RNA).

[26] Major F. et al.

The combination of symbolic and numerical computation of three-dimensional modeling of RNA. *Science*, 1991 **253**: 1255-1260.

[27] Major F., Lemieux S., Ftouhi A.

Computer RNA Three-Dimensional Modeling From Low Resolution Data and Multiple-Sequence Information. *American Chemical Society*, Symposium Series 1997. Sous presse.

[28] Mazur J., Jernigan R. L.

Comparison of rotation models for describing DNA conformations: application to static and polymorphic forms. *Biophys. J.* 1995, **68**: 1472-1489.

[29] Naher, S.

The LEDA Manuel Version 3.1, Max-Planck-Institut fur Informatik, Saarbrucken, Germany.

[30] Newcomb L. F., Gellman S. H.

Aromatic stacking interactions in aqueous solution: evidence that neither classical hydrophic effects nor dispersion forces are important. *J. Am. Chem. Soc.* 1994, **116**: 4993-4994.

[31] Norberg, J., Nilsson L.

Stacking-unstacking of the dinucleoside monophosphate guanylyl-3',5'-uridine studied with molecular dynamics. *Biophys. J.* 1994, **67**: 812-824.

[32] Norberg, J., Nilsson L.

Stacking free energy profiles for all 16 natural ribodinucleoside monophosphates in aqueous solution. *J. Am. Chem. Soc.* 1995, **117**: 10832-10840.

[33] Pageau M., Ftouhi A. Portillo J. R., Major F.

Structural Analysis of DNA/RNA Three-Dimensional. En Rédaction.

[34] Pan T., Uhlenbeck O. C.

In vitro selection of RNAs that undergo autolytic cleavage with Pb^{2+}. *Biochemistry* 1992, **31**: 3887-3895.

[35] Paul R. P.

Robot Manipulators: Mathematics, Programmings, and Control; MIT Press: Cambridge, MA, 1981.

[36] Pley H., Flaherty k. M., McKay D. B.

Three-dimentional structure of a hammerhead ribozyme. *Nature* 1994 Nov 3; **372** (6501): 68-74.

[37] Quigley G. J., Seeman N. C., Wang A. H., Suddath F. L., Rich A.

Yeast phenylalanine transfer RNA: atomic coordinates and torsion angles. *Nucleic Acids Res.* 1975, **2**: 2329-2341.

[38] Rich A., RajBhandary U. L.

Transfer RNA: molecular structure, sequence, and properties. *Ann. Rev. Biochem.* 1976, **45**: 805-860.

[39] RIchards F. M.

Areas, volumes, packing and protein structure. *Ann. Res. Biophys. Bioeng.* 1977, **6**: 151-176.

[40] Ramachandran G. N., Sasisekharan V.

Conformation of polypeptides and proteins. *Adv. Prot. Chem.* 1968, **23**: 283-437.

[41] Saenger W.

Principales of Nucleic Acid Structure. Springer-Verlag, New-York, NY, 1984.

[42] Scott, W., Finch J., Klug A.

The crystal structure of an all-RNA hammerhead ribozyme: a proposed mechanism for RNA catalytic cleavage. *Cell*, 1995, **81**, 991-1002

[43] Shafer G.

A Mathematical Theory Of Evidence. Princeton Univ. Press. Princeton, NJ, 1976.

[44] Sponer J., Kypr J. L.

Theoretical analysis of the base stacking in DNA: choice of the force field and a comparison with the oligonucleotide crystal structures. *J. Biomol. Struct. Dyn.* 1993,**11**: 277-292.

[45] Srinivasan A. R., Olson W. K.

J. Biomol. Struct. Dynam. 1987, **4**

[46] Sukhorukov B. I., Gukovsky I. Ya., Petrov A. I., Gukovskaya A. S., Mayevsky A. A., Gusenkova N. M.

Interactions and self-organization of nucleic bases, nucleotides, and polynucleotides into ordered structures. Effect of ionization and solvent salt composition. *Int. J. Quant. Chem.* 1980, **9**: 326-437.

[47] Tinoco I. Jr., Davis P., Hardin C. C., Puglisi J. D., Walker G.T., Wyatt J.

In Cold Spring Harbor Symposia on Quantitive Biology Cold Spring Harbor Laboratory, 1987 : 135-146.

[48] Topal M. D., Warshaw M. M.

Dinucleoside monophosphates. II. Nearest neighbor interactions. *Biopolymers* 1976, **15**: 1775-1793.

[49] Varani G., Wimberly B., Tinoco I. Jr.

Conformation and dynamics of an RNA internal loop. *Biochemistry* 1989, **28**: 7760-7772.

[50] Walley P.

Statistical Reasoning With Imprecise Probabilities. Chapman and Hall, London, 1991.

[51] Watson J. D., Crick, F. H. C.

A structure for Deoxyribose Nucleic Acid. *Nature (London)* 1953, **171**: 737-738.

[52] Weast R. C.

Handbook of Chemistry and Physics. CRC Press Cleveland, Ohio. p. D-178.

[53] Weiner S. J. et al.

A new forcefield for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* 1984, **106**: 765-784.

[54] Weiner S. J., Kollman P. A., Nguyen D. T., Case D. T.

An all atom force field for simulations of proteins and nucleic acids. *J. comp. Chem.* 1986, **7**: 230-252.

[55] Wells R. D., Harvey S. C.

Unusual DNA Structures. Springer-Verlag, New York, 1987.

[56] Wimberly B., Varani G., Tinoco I. Jr.

The conformation of loop E of eukaryotic ribosomal RNA. *Biochemistry*, 1993, **32**: 1078-1087.

[57] Wing R., Drew H., Takano T., Broka C., Takana S.

Crystal structure analysis of a complete turn of B-DNA. *Nature* 1980, **287**: 755-758.

[58] Zadeh L. A.

Fuzzy sets as a basis for a theory of possibility, *Fuzzy sets and Systems* 1978, **1**: 3-28.