THÈSE DE DOCTORAT DE L'UNIVERSITÉ
PIERRE ET MARIE CURIE

Spécialité :
BIOPHYSIQUE MOLÉCULAIRE

Présentée par
Guillaume SANTINI

pour obtenir le grade de Docteur de l'Université Paris VI

Sujet de la Thèse :

VERS LA PRÉDICTION DE LA STRUCTURE TRIDIMENSIONNELLE DES ÉPINGLES À CHEVEUX D'ADN ET D'ARN COMPORTANT UN APPARIEMENT DANS LA BOUCLE

À PARTIR DE LA THÉORIE DE L'ÉLASTICITÉ ET DE LA MÉCANIQUE MOLÉCULAIRE

Soutenance prévue le 29 Septembre 2005

Devant le Jury composé de :

Monsieur P. Auffinger	Rapporteur
Monsieur J. Cognet	Directeur de Thèse
Monsieur J. Delettré	Examinateur
Madame C. Etchebest	Examinateur
Monsieur J. L. Martiel	Examinateur
Monsieur O. Mauffret	Rapporteur

Á mon grand-père

Pour Barbara, Hugo et Axelle

Remerciements

Je tiens d'abord à remercier Jean Cognet pour m'avoir offert l'opportunité de travailler sur un sujet de recherche aussi riche et formateur. Je dois à la qualité de votre sujet et à votre soutien l'obtention d'un financement pour ce travail. Je vous remercie de ne pas avoir compté vos heures quand il fallait boucler un dossier, quand je m'arrachais les cheveux à lancer un calcul récalcitrant. C'est avec la patience et la rigueur des grands pédagogues, avec le goût de celui qui aime le savoir que vous avez su me conseiller et enrichir mes connaissances au delà du cadre du sujet de cette thèse, tant sur les notions théoriques que les points pratiques. La recherche n'est pas faite que de science, c'est aussi une aventure humaine. Je veux donc également vous remercier pour votre accueil, pour ces discussions enrichissantes sur tous les sujets, pour la chaleur dont vous savez nourrir les relations et pour la compréhension dont vous avez fait preuve lors des périodes agitées qui ont suivi les naissances de mes enfants.

Je veux ensuite remercier Christophe Pakleza. Dès le début, chose assez exceptionnelle, tu as mis à mon entière disposition, sans retenue, le résultat de plusieurs années de ton travail pour que je puisse construire le mien. Tu as pris le temps de me mettre le pied à l'étrillé alors que tu n'en avais pas. Tu as toujours été disponible à mes sollicitations. Tu as été mon gourou Linux. Tu m'as initié à Mathematica. //(d'ailleurs mon code en portera toujours la marque#)&. La rigueur et l'exigence dont tu fais preuve dans ton travail ont placé la barre très haut. Passer après toi fût donc une vraie chance. Je veux aussi te remercier aussi chaleureusement que possible pour ton accueil et ta gentillesse et ton aide. Grâce à toi j'ai tout de suite trouvé ma place dans un environnement dont je ne connaissais au début ni les us ni les coutumes. Je veux finalement te remercier pour toutes ces discussions empreintes de cet humour subtil et intelligent dont tu as le secret.

Jean, Christophe, grâce à vous deux cette aventure scientifique et humaine à été à la hauteur de mes attentes. Je vous en serais toujours reconnaissant.

Je veux ensuite remercier sans exception l'ensemble des membres du laboratoire. Merci pour cet accueil chaleureux dont vous avez fait preuve durant toutes ces années. Je remercie Messieurs Jacques Bolard et Mahmoud Ghomi, Directeurs du LPBC et du BioMoCeTi de m'avoir accueilli parmi leurs équipes, de la confiance qu'ils m'ont porté, du soutien et des conseils qu'ils m'ont accordé. Je veux également adresser des pensées chaleureuses à Belén Hernández pour sa gentillesse et son écoute et à Olivier Seksek pour nos discussions et pour l'echo toujours humoristique donné aux questions existentielles que soulève la paternité à l'orée du deuxième millénaire.

Je remercie P. Auffinger et O. Mauffret d'avoir accepté d'être les rapporteurs de ce travail. Je remercie par ailleurs Pascal Auffinger de l'aide qu'il nous a apporté pour engager des calculs de dynamique moléculaire sur plusieurs molécules et dont les études ne figurent pas dans ce manuscript. Je remercie également Olivier Mauffret et Serge Fermandjian de nous avoir communiqué certaines données facilitant la réalisation de cette étude. Je remercie Jean Louis Martiel, Catherine Etchebest et Jean Delettré d'avoir accordé de leur temps précieux pour participer au jugement de ce travail.

J'adresse ici une pensée particulière à Emmanuel Lemeulin, un ami fidéle qui a partagé avec moi les épreuves de l'accession au grade de Docteur, mais surtout de très nombreuses soirées parmi les plus festives de ma vie.

Je veux finir par remercier toute ma famille pour le soutien et l'attention dont ils ont fait preuve. Je pense spécialement à mon grand-père décédé avant la conclusion de ce travail et à la fierté qu'il y portait. Merci à mon père, à ma mère et à mon beau-père. Votre contribution à la réalisation de ce travail a été de premier ordre car c'est à vous que je dois ce goût pour la connaissance. C'est grâce à votre présence et votre amour durant ces longues années d'études que tout a été possible. Dans les succès comme lors des échecs je vous remercie pour la constance de votre soutien. Il a toujours été et sera encore longtemps un moteur fondamental et une source de dépassement.

Merci enfin à mes trois amours, Barbara ma femme, Hugo mon fils et Axelle ma fille. Durant ces cinq dernières années, vous avez été les comptables de mes interrogations, des mes hésitations, de mes fatigues et autres sautes d'humeur. Merci d'avoir tout supporté. Barbara, je ne te remercierais jamais assez de m'avoir soutenu dans cette étape de notre vie. Comme tout se mesure à l'aune de votre regard, il est naturel, à la fin, que tout cela vous soit dédié.

Table des matières

Table des figures

Liste des tableaux

Définitions, notations et abréviations

Nucléotide : Sous unité monomérique des chaînes d'acides nucléiques. Il est composé d'un sucre ribose ou désoxyribose, d'une base azotée (Adénine, Guanine, Cytosine, Thymine ou Uracyl), et d'un groupe phosphate permettant la formation de liaisons phosphodiesters entre les différents monomères.

Nucléoside : Nucléotide privé du groupement phosphate.

Base : Cycle azoté conjugué composant "l'alphabet" des polymères d'acides nucléiques. Par abus de langage, ce terme est parfois utilisé en lieu et place des termes "nucléotides" ou "nucléosides" d'une séquence d'une chaîne d'acides nucléiques.

Tige : partie en double hélice des structures en épingle à cheveux.

Boucle : Partie simple brin fermant les épingles à cheveux à une extrémité de la tige.

RMSd : Acronyme anglo-saxon pour "Root Mean Square deviation" ou "Root Mean Square displacement" (*cf.* ANNEXE : A.1).

$\mathbf{B}_i$, $\mathbf{N}_i$: i^{eme}base ou nucléotide d'une séquence d'acide nucléique.

Numérotation des Bases : La numérotation des bases (*i.e.* des nucléotides) suit la convention d'orientation de l'extrémité 5' à l'extrémité 3'. Dans la boucle, la numérotation débute avec la base qui suit le dernier plateau de paires de bases de l'hélice

Numérotation des brins des hélices : Dans les tiges des épingles à cheveux , le brin en 5' de la séquence est noté "brin I" et le brin en 3' est noté "brin II", pour suivre la convention de Cambridge [1].

Numérotation des plateaux de paires de bases : Dans la tige des épingles à cheveux, les plateaux de paires de bases sont numérotés comme les bases du brin I.

ADN : Acronyme pour "Acide DésoxyRibonucléique". C'est un polymère linéaire de désoxyribonucléotides.

ARN : Acronyme pour "Acide Ribonucléique". C'est un polymère linéaire de ribonucléotides.

RMSd : Acronyme pour "Root Mean Square Deviation" ou "Root Mean Square displacement" (*cf.* ANNEXE : A.1).

Chapitre I

Les structures en épingles à cheveux d'acides nucléiques

Les acides nucléiques sont les molécules support de l'information génétique chez les êtres vivants. Cette information séquentielle est codée par la chaîne d'un polymère linéaire. Avec la découverte du code génétique, la séquence est naturellement devenue le critère prépondérant de description de ces molécules.

La structure tridimensionnelle est un critère tout aussi fondamental pour différentes raisons. Elle participe au maintien d'une transmission stable et régulée de l'information génétique et assure la stabilité et la fonctionnalité de l'ensemble des molécules biologiques. L'étude de la structure et des interactions à l'échelle moléculaire, fait partie de la "modélisation moléculaire" au sens large. Ce champ d'étude en croissance constante est important pour comprendre les mécanismes biologiques associés aux molécules d'ADN et d'ARN.

Nos travaux portent sur le développement et l'utilisation d'une nouvelle approche de modélisation des acides nucléiques en épingles à cheveux. Nous définirons d'abord ces motifs particuliers de l'ADN et de l'ARN, puis nous aborderons le cadre général de la modélisation moléculaire des acides nucléiques. Nous résumerons les méthodes générales d'obtention et les problématiques associées au traitement de ces modèles moléculaires. Enfin, nous ferons un état des connaissances sur les structures tridimensionnelles des épingles à cheveux et sur les questions et difficultés qu'elles posent.

I.1 Présentation générale des structures en épingles à cheveux d'acides nucléiques

Dans cette partie nous présentons notre objet d'étude : les structures en épingles à cheveux. Après avoir défini la structure et l'intérêt biologique de ce motif, nous développerons brièvement l'évolution des connaissances de ces structures.

I.1.1 Définition de la structure primaire et secondaire des épingles à cheveux d'acides nucléiques

Les structures en épingles à cheveux sont des motifs fondamentaux et ubiquitaires des acides nucléiques. Constitué d'une seule et même chaîne moléculaire, ce motif ne peut se former qu'à partir de séquences globalement auto-complémentaires, comme le sont les séquences répétées-inversées ou quasi-palindromiques (*cf.* FIG. : I.1). La chaîne moléculaire repliée sur elle-même, se structure en une partie en double hélice - la tige - fermée à un bout par une partie simple-brin - la boucle. La partie en tige est constituée par l'appariement des portions de séquences complémentaires de la chaîne moléculaire, alors que la boucle est formée de quelques bases non appariées ou appariées de façon non canonique. Si l'ARN simple brin se prête, de par sa nature monomoléculaire, à la formation de telles structures, l'ADN double brin présente aussi de tels motifs.

De façon générale les polymères d'acides nucléiques tendent à former naturellement des structures de type hélice. Lorsqu'il n'y a pas de chaîne de séquence complémentaire avec laquelle s'apparier, une chaîne unique va donc tendre à se replier sur elle-même pour former le maximum de parties en double hélice. Ces dernières sont alternées avec des parties simple-brin en boucle qui lient les hélices entre-elles. Les structures en hélice ou en boucle sont donc les deux motifs structuraux fondamentaux des acides nucléiques.

Dans l'ARN, le mode de structuration tridimensionnelle d'une partie donnée de la molécule suit une alternative simple : soit elle s'associe à une autre partie qui lui est complémentaire et forme une double hélice, soit elle reste sous forme simple brin. Dans ce cas elle peut donner lieu à différents motifs tels qu'une boucle interne, une hernie (ou bulge en anglais) ou une épingle à cheveux (*cf.* FIG. : 18.1 dans [2]). L'ARN est dans une première description essentiellement architecturé autour de

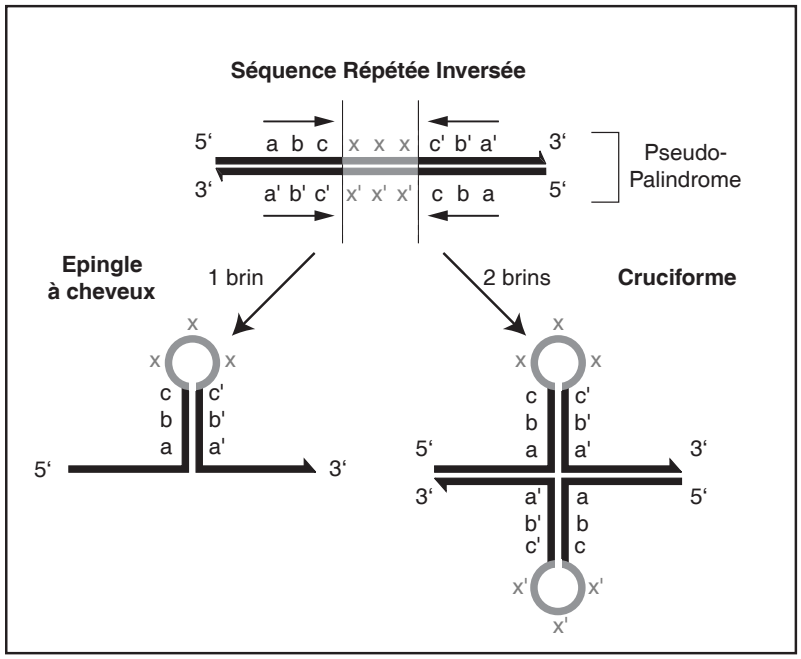


Fig. I.1 : ***Structure secondaire du motif en épingles à cheveux des Acides nucléiques :*** *Une séquence pseudo-palindromique peut former par réappariement local, sur un seul brin, une structure en épingle à cheveux, ou symétriquement sur les deux brins une structure de type cruciforme. La structure en épingle à cheveux est aussi appelée tige-boucle. La tige est la partie double brin en hélice formée par l'appariement des séquences répétées inversées complémentaires. La boucle est la partie simple brin constituée de nucléotides non appariés ou mésappariés. Elle est constituée par la séquence séparant les deux parties répétées-inversées formant l'hélice.*

parties en hélices reliées par des boucles en épingle à cheveux. (*cf.* les premières descriptions de structures secondaires de l'ARN 16S et 23S dans [3]). Les motifs stables en épingle sont donc aussi fondamentaux que les hélices pour l'architecture de ces polymères linéaires.

Dans l'ADN la double hélice peut former de façon transitoire et ponctuelle une épingle à cheveux sur un brin lors de réappariements locaux, ou de façon plus symétrique des structures dites cruciformes formées d'une épingle à cheveux sur chacun des deux brins [4] (*cf.* Fig. : I.1).

L'étude des structures en épingles à cheveux est aussi motivée par de nombreuses questions biologiques pour établir les liens entre la compréhension de processus biophysico-chimiques et celle des modes de structuration de ces molécules fondamentales à la vie.

I.1.2 Intérêt biologique

Dans l'ADN : L'importance biologique des structures d'ADN en boucle a longtemps été soupçonnée et commence à être clairement confirmée par une série de travaux récents [5, 6]. Les formes en épingles à cheveux apparaissent dans de nombreux aspects du métabolisme de l'ADN (réplication, réparation, et recombinaison). Elles sont très répandues dans les génomes des procaryotes et des eucaryotes, et sont des éléments des structures de contrôles : promoteurs, terminateurs. Chez les virus hôtes des procaryotes, eucaryotes et des cellules de mammifères, elles forment des origines de réplications du génome viral [7]. Elles sont fonctionnellement importantes pour l'initiation de la réplication de l'ADN dans les plasmides, les bactéries, les virus d'eucaryotes et les cellules de mammifères [8]. Elles peuvent former des structures cruciformes dans des conditions de superenroulement de l'ADN [6]. La formation des cruciformes *in vivo* a été démontrée chez les procaryotes et dans les cellules de mammifères [9, 10].

Dans l'ARN : L'importance biologique des boucles dans l'ARN de transfert, messager ou ribosomal n'est plus à démontrer et ne peut être détaillée ici. À titre d'exemple, notons l'importance de leur rôle structural pour faciliter un repliement et/ou une activité (ex : ribozyme ; interaction spécifique pour la réplication de virus HIV), ou pour reconnaître directement des ligands (ex. : ARNt). Ces structures en boucles sont parties prenantes d'un grand nombre de régulations, notamment lors de l'étape de l'épissage alternatif des régions non codantes des ARMm. Cette étape peut rendre compte de la variabilité génétique des isoformes protéiques qui pourraient expliquer les susceptibilités génétiques pour la maladie d'Alzheimer et la variabilité neuronale. Mentionnons aussi chez les virus des structures comme Hep C IRES et polio 5' UTR ou encore la boucle HIV1-TAR qui permet l'activation de la transcription virale.

I.1.3 Intérêt biologique et médical

Les boucles d'ADN et d'ARN sont des motifs structuraux très importants pour :

Comprendre les mécanismes moléculaires provoquant certaines maladies génétiques : Plusieurs maladies humaines génétiques : l'oedème angioneurotique héréditaire, la dystrophie musculaire de Duchenne, l'hypercholestérolémie familiale, ou l'une des douze maladies générées par l'instabilité de triplets répétés (la

dystrophie myotonique, le syndrome de l'X fragile, la chorée de Huntington, ...) ne sont pas dues à des protéines de réparation défectueuses ou absentes, mais à la formation de structure secondaire en épingle à cheveux qui permet à une boucle d'ADN d'échapper *in vivo* à la réparation [11].
En effet, des régions de l'ADN où la séquence est inversée-répétée peuvent adopter l'une ou l'autre des deux structures secondaires - en double hélice ou en structure cruciforme - et constituent ainsi une barrière à la fidélité de la réplication [5]. Ces structures cruciformes facilitent les mutations par décalage du cadre de lecture et peuvent provoquer des mutations par addition ou délétion donnant lieu à une transmission instable de matériel génétique.

Comprendre les mécanismes moléculaires afin de perfectionner des techniques de thérapies géniques : Les recherches pour utiliser des virus non pathogènes tels que l' "adenoassociated virus 2" (AAV2) comme vecteur en thérapie génique montre l'importance de certaines parties en épingles à cheveux du génome viral. En effet, la réplication du virus ainsi que sa capacité d'intégration spécifique en un site du chromosome humain 19 dépend directement de la présence de tige-boucles aux extrémités du virion. Le génome adopterait une jonction à trois branches ("three-way junction") avec deux bras portant respectivement une boucle T3 et une boucle A3 [12, 13] importante pour comprendre l'efficacité de AAV2 [14].

Renforcer la résistance des molécules anti-sens aux dégradations *in vivo* : Les simple-brins d'ADN anti-sens peuvent être protégés efficacement contre les dégradations des nucléases par l'ajout aux extrémités de la molécule de séquences se repliant en tige-boucles [15–17].

Comprendre les mécanismes moléculaires et développer de nouvelles molécules issues de la stratégie aptamère : Cette méthodologie permet d'obtenir très généralement des molécules d'acides nucléiques (ADN ou ARN) qui ont une très bonne affinité pour une cible donnée [18] et en particulier contre des ARN. À titre d'exemple, des aptamères ont été obtenus contre la partie en ARN de "HIV trans-activation-responsive" (TAR) qui fixe la protéine virale Tat [19]. Ils devraient théoriquement entrer en compétition avec cette protéine virale et empêcher la transcription du génome. Ces aptamères d'ADN ont une structure en tige-boucle qui s'ajuste à la structure en tige-boucle de TAR. L'ensemble donne lieu par le biais d'interaction boucle-boucle à des complexes dénommés "kissing complexes" entre deux épingles à cheveux [20].

Dans le cadre de la stratégie d'ARN interférence : Notons le succès

retentissant de l'utilisation de petits ARN en épingles à cheveux (shRNAs) sur des cellules de mammifères [21, 22] et ses applications potentielles dans la lutte contre le cancer et le SIDA. Un simple brin d'ARN exprimé par un plasmide transfecté, replié sur lui-même en épingle à cheveux, forme une portion d'ARN double brin (tige) capable d'induire une réaction d'inactivation spécifique d'un gène de séquence homologue par dégradation de son ARN messager [21].

En conclusion, l'ADN et l'ARN, qui sont les supports de l'information génétique, peuvent aussi, grâce à leurs propriétés physico-chimiques, structurales et dynamiques uniques, être considérés comme le support même de la régulation de l'expression des gènes. Dans ces processus de régulation de l'expression des gènes les boucles jouent un rôle essentiel.

I.1.4 Historique

I.1.4.1 Évolution des connaissances sur la structure des acides nucléiques :

Les premières résolutions structurales d'acides nucléiques proviennent du traitement de données de diffractions de fibres d'ADN [23–28]. Ces expériences donnent accès à la structure moyenne des paires de bases. Par exemple, pour un duplex poly(dG).poly(dC), seule la conformation moyenne d'un plateau de paire de bases (G$\cdots$C) est accessible, sans aucune autre information particulière sur les variations locales de la forme de la molécule. La structure en double hélice est obtenue par extension des données d'une paire de bases moyenne. Les paires de bases sont répétées et empilées autour d'un axe rectiligne dans une géométrie hélicoïdale parfaite. Les plateaux de paires de bases (A$\cdots$T) et (G$\cdots$C) sont isomorphes. Il est donc possible de construire des hélices régulières où chaque plateau est isomorphe à tout autre plateau du même type comme dans une hélice idéale [29–31].

Dans cette représentation, l'ADN apparaît comme une molécule très structurée selon un motif très régulier. Cette vision de la conformation de l'ADN implicitement indéformable renforçait l'idée d'une molécule très stable dont la principale fonction est de garantir la conservation de l'information génétique. Cette vision forte de la structure des acides nucléiques, profondément ancrée dans les esprits, a longtemps perduré jusqu'aux progrès de la biologie moléculaire, des techniques physiques et informatiques d'études structurales. Les développements parallèles des techniques

d'obtention d'oligonucléotides (purification, clonage, PCR, synthèse), les progrès dans l'acquisition de données structurales (radiocristallographie aux rayons X et RMN bidimensionnelle), ainsi que les progrès conjoints de l'informatique et des techniques de modélisation moléculaire, ont permis de renouveler complètement le domaine [32].

Aujourd'hui, la possibilité d'accéder à la structure complète et globale de molécules de tailles diverses (allant de quelques nucléotides à de très gros complexes protéines/acides nucléiques), la croissance exponentielle du nombre de structures résolues et la qualité croissante de celles-ci, ont profondément modifié notre vision de ces macromolécules très structurées, initialement perçues comme des "colloïdes" avant la fin des années 50. Leurs descriptions sont beaucoup plus complexes (courbure des hélices, triples et quadruples brins, épingles à cheveux, boucles internes, complexe ADN-ARN/protéines) qu'auparavant, et le spectre de leurs rôles biologiques s'enrichit de fonctions de régulation de l'expression génétique, de contrôle de la réplication et de la réparation ou encore d'activités catalytiques.

Ainsi, d'une conception figée de la structure de la double hélice support stable de l'information génétique, nous sommes passés aujourd'hui à une complexité structurale et une grande diversité fonctionnelle qui est à l'origine de l'essor exponentielle de l'étude de la structure des acides nucléiques depuis une dizaine d'années environ [33].

I.1.4.2 Les épingles à cheveux

Historiquement les parties simple-brins en boucle étaient considérées comme peu ou pas structurées relativement aux doubles-brins en hélices. Les premières études thermodynamiques portant sur les épingles à cheveux montraient que les boucles les plus stables dans l'ADN devaient comporter entre quatre et cinq nucléotides, et que dans l'ARN un maximum de stabilité était obtenu pour des boucles de six à sept nucléotides [34,35]. Cette vision a évolué avec la découverte, notamment dans l'ARN, de structures en épingles à cheveux hyperstables présentant des boucles à seulement quatre nucléotides (UUCG et GNRA) dont la température de demi-dénaturation était très élevée [36].

Leur résolution structurale a montré que ces boucles étaient en fait très ordonnées et très compactes. La grande stabilité de certains motifs s'expliquait par la formation

d'un appariement entre les bases extrémales de la boucle avec des interactions stabilisantes de type :

- liaisons hydrogène entre les bases appariées, et
- empilements et interactions hydrophobes entre le plateau apparié de la boucle et les plateaux de la tige.

Par la suite, de nombreuses structures en tige-boucles à trois ou quatre nucléotides contenant ou non des appariements dans la boucle ont été découvertes dans l'ADN et dans l'ARN. Des similarités de conformations et de séquences ont permis de classer les motifs en épingles à cheveux en familles structurales et d'établir les boucles à trois et quatre nucléotides comme des boucles de référence intervenant dans les processus biologiques [36, 37].

I.2 Méthodes d'obtention des conformations d'acides nucléiques

Les buts de la modélisation moléculaire sont multiples selon les domaines d'intérêt. Mentionnons pour la compréhension des comportements moléculaires, la localisation et la dynamique des ions, des molécules d'eau, de l'orientation des groupements chimiques qui participent à des liaisons ou des interactions (liaisons hydrogène, liaisons CH...O [38], interactions ions - π [39, 40], interactions hydrophobes [41], ...), la dynamique du squelette, les mouvements des chaînes latérales, ... Citons aussi pour la pharmacologie, la recherche de molécules ligands pour bloquer un site fonctionnel d'un enzyme ou d'une macromolécule biologique. Ces listes non exhaustives correspondent à des utilisations de structures établies de macromolécules biologiques ou des explorations autour de structures résolues.

Ces buts doivent être distingués des méthodes de calcul de modélisation qui visent à établir des structures moléculaires à partir de données dérivées de l'expérience. Dans ce dernier cas, on cherche au contraire à éliminer l'incertitude ou l'insuffisance des données expérimentales en rajoutant des informations à partir de bases de données sur les encombrements stériques des atomes, les longueurs et angles de liaisons, et essentiellement à partir d'arguments de mécanique et de dynamique moléculaire.

I.2.1 Diversité des données structurales dérivées de l'expérience

Les données structurales expérimentales qui servent à établir les structures des macromolécules proviennent principalement de la Résonance Magnétique Nucléaire (RMN) et de la radiocristallographie aux rayons X. Ces méthodes donnent accès à une vision détaillée de la molécule à l'échelle atomique de façon différente.

Les expériences de RMN bidimensionnelle, regroupées sous le terme générique de RMN, donnent accès à certaines distances interatomiques de la molécule en solution. La RMN du proton permet d'identifier les protons et d'évaluer les distances moyennes qui les séparent s'ils sont suffisamment proches. Les progrès de cette technique expérimentale (puissance des appareils, préparation des échantillons marqués, complexité des séquences impulsionnelles, traitement des signaux, ...) permettent aujourd'hui d'accéder à de plus en plus d'informations (marquage C^{13}ou N^{15}, angles de torsion portant sur des atomes autres que les atomes d'hydrogène, orientation absolue de groupes chimiques avec la technique des couplages résiduels dipolaires, ...). Les données structurales dérivées de la RMN conservent leur caractère relatif : la position d'un atome est évaluée en fonction des distances à d'autres atomes de la même molécule.

Avec ce type de données expérimentales le travail de modélisation consiste à élaborer un modèle moléculaire qui donne l'ensemble des positions atomiques avec un jeu partiel voire très incomplet de données de distances et d'angles interatomiques. Pour lever certaines incertitudes ou certaines incompatibilités dans les données structurales dérivées de l'expérience, le modélisateur fait appel aux principes de bases de la physicochimie des molécules (angles et longueurs de valence, potentiels d'interactions entre atomes, ...) avec les champs de force, et aux connaissances accumulées sur des structures proches (angles de torsions les plus fréquemment rencontrés, comportements global de la molécule) avec les bases de données de structures.

Avec la radiocristallographie aux rayons X, la nature des données structurales change considérablement. Cette technique permet *via* le traitement de cartes de densités électroniques d'accéder aux positions cartésiennes des atomes de la molécule. Le travail du modélisateur consiste à trouver une structure qui satisfait à la fois les positions cartésiennes données par l'expérience et les règles de géométrie des liaisons atomiques. Des étapes d'ajustement sont donc nécessaires car parfois plusieurs

arrangements différents sont possibles. Le modélisateur doit également extrapoler la position des atomes d'hydrogène qui n'apparaissent pas bien sur les cartes de densité électronique. Pour ce faire, il doit, là aussi, tenir compte des principes de géométrie des molécules et chercher les orientations optimales des atomes qui permettent la mise en place d'interactions stabilisantes.

Les données expérimentales sont en général insuffisantes et un traitement de modélisation complémentaire est nécessaire pour atteindre la résolution détaillée d'une structure moléculaire. Le modélisateur doit traiter des informations à l'échelle atomique pour élaborer une structure correcte tant au niveau local que global. Il doit également décrire cette structure complètement, *i.e.* proposer une position acceptable pour l'ensemble des atomes qui tient compte à la fois des principes de géométrie des liaisons atomiques et des données structurales fournies par l'expérience. Le grand nombre d'atomes à placer dans les macromolécules biologiques rend ce travail non-trivial.

I.2.2 Principes généraux de modélisation moléculaire

Un modèle moléculaire tridimensionnel est une représentation simplifiée de la structure d'une molécule. Il doit décrire complètement la topologie (*i.e.* les liaisons interatomiques) et la position spatiale de chaque atome de la molécule dans un système de coordonnées donné (coordonnées internes ou cartésiennes : *cf. infra*). Pour que le modèle ait un sens, il doit être énergétiquement stable mais également compatible avec les données structurales dérivées de l'expérience. Il doit respecter les règles de géométrie et d'énergie minimale à l'échelle locale des atomes (longueurs et angles de valence), et aussi présenter des interactions stabilisantes à l'échelle globale de la molécule pour être thermodynamiquement stable.

L'élaboration d'un modèle moléculaire n'est pas une chose aisée. Les molécules modélisées en biologie sont des structures complexes du fait de leur taille, du grand nombre d'interactions faibles et de leur mode de structuration optimisé par l'évolution. Pour les modéliser, c'est à dire pour les représenter de façon simple et quantitative, diverses approches de déformations et de description sont utilisées.

Pour résoudre ce problème non trivial, l'approche générale consiste à définir dans un premier temps une structure de départ dont la topologie correspond à celle de la molécule à modéliser : composition et structure stéréochimique correcte.

Cette structure de départ dont la conformation tridimensionnelle est incorrecte, est soumise à des déformations successives qui respectent la topologie de la molécule, pour la faire converger vers une structure finale qui va respecter les contraintes structurales dérivées de l'expérience.

Cependant, déformer intelligemment un objet aussi complexe n'est pas une chose aisée. Les déformations à l'échelle des atomes sont naturellement des événements complexes et concertés qui impliquent un grand nombre de paramètres. Les approches de modélisation généralement utilisées rencontrent les mêmes difficultés mais ne disposent pas nécessairement des méthodes et du temps pour explorer ces espaces de déformation. Le grand nombre d'atomes multiplie les modes de déformation possibles (grand nombre de degrés de liberté de déformation). Les espaces conformationnels ainsi définis sont complexes et les chemins de repliement-déformations possibles sont trop nombreux pour être explorés systématiquement. Il faut donc recourir à des astuces ou de nouvelles méthodes pour réaliser des déformations efficaces. Ceci est réalisé notamment en limitant le nombre de degrés de liberté de déformation des structures et en contraignant l'exploration des espaces conformationnels par l'introduction de contraintes dérivées de notre connaissance des mouvements atomiques et des données structurales dérivées de l'expérience.

Comme nous allons l'aborder, il existe principalement deux systèmes de coordonnées pour décrire les molécules : les coordonnées cartésiennes et les coordonnées internes en angles de torsion. Ces deux systèmes sont associés à différentes façons de "voir" les déformations des molécules et de modéliser les degrés de liberté naturels de déformation de celles-ci. Chacune donnent lieu à différentes approches de déformations et de modélisation.

I.2.3 Choix du système d'expression des coordonnées atomiques et degrés de liberté de déformation des molécules

Afin de décrire les positions des atomes et de définir un jeu de transformations, le modélisateur doit définir un système d'expression des coordonnées de la molécule. Le choix du système de coordonnées est souvent lié au mode de déformation (*i.e.* degrés de liberté) utilisé pour manipuler et déformer la molécule. Il existe deux principaux modes de déformation des molécules associés à deux systèmes de coordonnées différents.

I.2.3.1 Coordonnées cartésiennes et translation des atomes

Un premier mode de description des molécules en biologie consiste à donner l'ensemble des positions cartésiennes des atomes qui la constituent. Ce premier formalisme décrit la molécule comme un ensemble de particules (les atomes) liés par des liaisons et des interactions physiques ou quantiques (fonctions de potentiel), soumises à l'agitation thermique (translations, vibrations et rotations des atomes). Comme les atomes sont considérés individuellement, la déformation de la molécule passe nécessairement par une modification de la position de chaque atome. Le déplacement d'un atome est décomposé en une somme de trois translations le long des axes d'un repère cartésien (*cf.* FIG. : I.2). Les plateformes de modélisation moléculaire (AMBER [42], CHARMm [43, 44]) sont des exemples d'applications de ce formalisme. Dans ces approches, les champs de force décrivent les énergies à l'échelle des atomes. La direction et l'amplitude de la translation des coordonnées d'un atome à l'itération n dans un calcul de dynamique moléculaire sont données par la force égale au gradient de la somme des énergies d'interactions qui s'appliquent à l'atome à l'issue de l'itération n-1, avec :

$$\overrightarrow{F} = -\nabla\overrightarrow{E_{total}} \tag{I.2.3.1}$$

Les énergies d'interactions qui s'appliquent à un atome de la molécule sont diverses en nature et en amplitude (*cf.* EQ. : I.2.3.2 du champ de force AMBER [45, 46]).

$$\begin{aligned} E_{total} \quad = \quad & \sum_{liaisons} K_r(r - r_{eq})^2 + \sum_{angles} K_\theta(\theta - \theta_{eq})^2 + \\ & \sum_{torsion} \frac{V_n}{2}(1 + \cos(n\Phi - \gamma)) + \sum_{i<j} \left(\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\varepsilon R_{ij}} \right) \end{aligned} \tag{I.2.3.2}$$

Dans cette fonction de potentiel unique les termes qui prédominent sont ceux dont les constantes sont très élevées (entre ∼300 et ∼600 kcal $Å^{-2}mol^{-1}$ pour la constante de liaison, K_r, et entre ∼60 et ∼120 kcal $rad^{-2}mol^{-1}$ pour la constante d'angle de valence, K_θ) et les termes en $\frac{1}{R^n}$ pour R petit. Ces termes conditionnent, plus que toutes les autres contributions, les modes de déformation accessibles à la molécule. Le seul terme qui ne peut pas prendre de grandes valeurs est la contribution énergétique des angles de torsion (entre ∼0,1 à ∼6 kcal mol^{-1} pour $\frac{V_n}{2}$

). Au total il en résulte que les déformations sont plutôt restreintes à des rotations autour des angles de torsion qui préservent les longueurs et angles de valence des liaisons atomiques.

Les contraintes structurales dérivées des données expérimentales sont introduites comme autant de termes d'énergie supplémentaires. Ils introduisent des forces locales additionnelles sur les atomes concernés par les contraintes structurales dérivées de l'expérience. Ils influent sur les déformations de la molécule à chaque étape de la dynamique, en orientant les déplacements dans une direction favorable à la mise en place d'une structure globale compatible avec l'expérience.

Modéliser une molécule avec des énergies introduit donc une hiérarchie dans le protocole de modélisation. Les contraintes géométriques associées aux plus fortes énergies sont prioritaires pour le champ de force et sont donc rectifiées avant les termes de plus basses énergies. Travailler directement avec une fonction unique d'énergie potentielle atomique revient donc à prendre d'abord en compte la modélisation des liaisons atomiques et leur géométrie, avec pour conséquence principale de réduire les modes de déformation possible à des jeux de rotation autour des angles de torsion. Or les rotations autour des angles de torsion ne sont pas des modes de déformation faciles à manipuler pour élaborer une structure tridimensionnelle aussi complexe que celle des macromolécules biologiques, comme nous le verrons au paragraphe suivant.

L'avantage de la description cartésienne est de permettre des modifications locales qui préservent la structure globale de la molécule (*cf.* FIG. : I.2). Son inconvénient est qu'il est difficile d'imprimer des déformations globales efficaces car le nombre de degrés de liberté (3N-6 pour N atomes) et leur interdépendance rendent les calculs très complexes.

I.2.3.2 Coordonnées internes et rotation autour des liaisons atomiques

Un second mode de description des molécules en biologie consiste à exprimer les positions des atomes au moyen de coordonnées internes (angles de torsion). Ce formalisme décrit la molécule comme un ensemble d'atomes liés entre eux par des liaisons autour desquelles des rotations sont possibles. Ce degré de liberté de rotation autour des liaisons correspond à un degré de liberté naturel de déformation des molécules. Pour déformer celles-ci il suffit de modifier la valeur des angles de

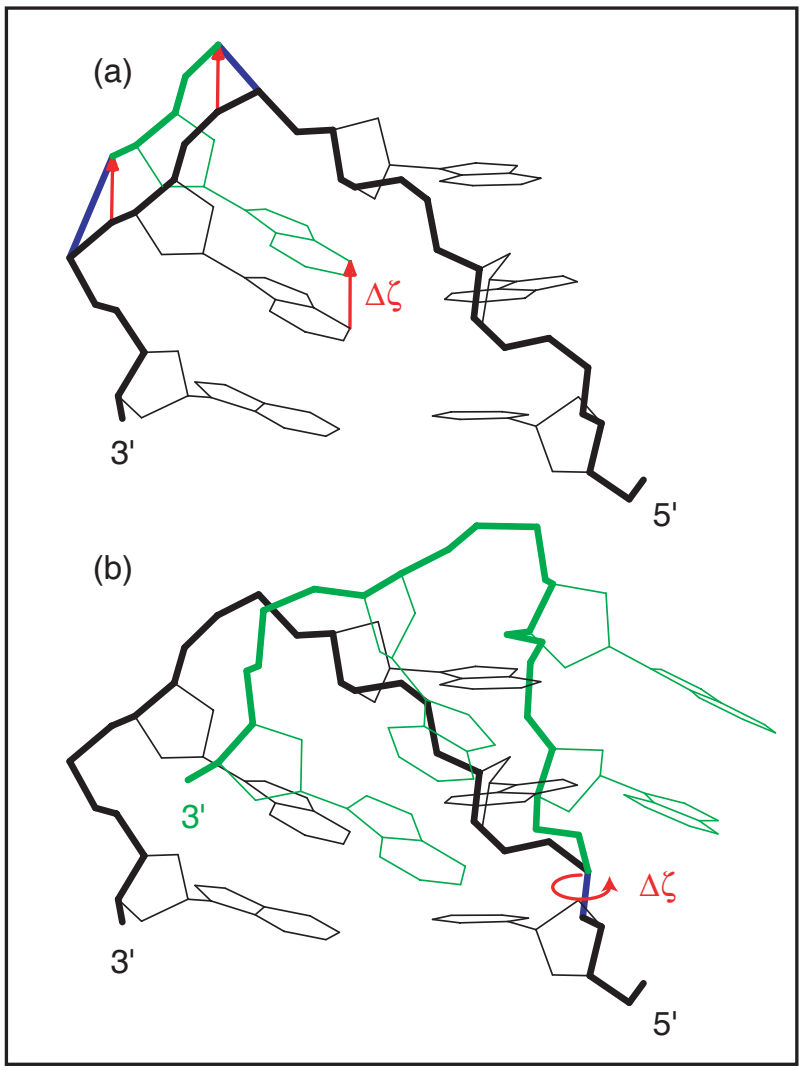


Fig. I.2 : *Exemples de déformation des macromolécules en coordonnées cartésiennes (a) et en angles de torsion (b) :* *L'épingle à cheveux 1BJH-AAA [47] est utilisée pour illustrer les modes de déformation en coordonnées cartésiennes et en coordonnées internes. La molécule originale est représentée en noir et la chaîne sucre-phosphate en trait noir épais. Les blocs déplacés sont en vert et les liaisons qui les relient au reste de la molécule sont en bleu. (a) déplacement par translation d'un vecteur rouge du quatrième nucléotide : ce déplacement déforme les liaisons flanquantes au bloc déplacé. La géométrie des liaisons n'est pas conservée. De faibles déplacements conservent la structure globale. (b) rotation de l'angle de torsion ζ du premier résidu autour de la liaison bleue. La géométrie des liaisons est conservée, mais la structure globale est fortement déformée à longue distance de la liaison de référence.*

torsion. Cette approche est utilisée dans des programmes comme DIANA [48], JUMNA [49] et DYANA [50]. Elle utilise comme précédemment des fonctions d'énergie potentielles.

Un avantage de cette approche est de réduire considérablement le nombre de degrés de liberté de déformation de la molécule. Il est réduit de 3N à N-3, où N est le nombre d'atomes (*cf.* Eq. : I.2.3.3), car les élongations des liaisons et les déformations des angles de valences ne sont plus prises en compte. Un autre avantage est de préserver, lors de chaque déformation, les longueurs et angles de liaison. L'inconvénient de cette approche est que de faibles variations locales de la valeur d'un angle de torsion peuvent avoir un impact très fort sur la conformation de la molécule à longue distance (*cf.* Fig. : I.2).

$$\begin{array}{rrl} & 3N & coordonnées\,atomiques \\ - & (N-1) & liaisons\,simples \\ - & (N-2) & angles\,de\,liaison \\ - & 3 & translations\,de\,l'ensenble\,de\,la\,molecule \\ - & 3 & rotations\ \ de\,l'ensenble\,de\,la\,molecule \\ \hline = & (N-3) & angles\,de\,torsion \end{array} \qquad \text{(I.2.3.3)}$$

avec,

N le nombre d'atomes de la molécule reliés par des liaisons covalentes.

Il faut donc (N-3) angles de torsion pour décrire la molécule. En pratique, les angles de torsion associés à la description des atomes dans les cycles des bases sont nuls ou de 180°. Si l'on suppose en outre que les bases sont parfaitement planes, ce nombre peut être encore abaissé.

I.2.3.3 Des descriptions atomiques

On remarque donc qu'à chaque formalisme de coordonnées moléculaires correspond une approche différente de déformation des molécules. Ces formalismes, dérivés d'une vision des molécules à l'échelle atomique rendent la description et la déformation des structures très complexes du fait du grand nombre de degrés de liberté. De telles descriptions des déformations moléculaires, peuvent sembler trop détaillées pour des molécules aussi grosses que les polymères biologiques. De plus, elles ne prennent pas en compte ni n'utilisent une caractéristique très importante des macromolécules biologiques : ce sont des polymères linéaires. Dans les chapitres qui vont suivre nous montrerons comment l'utilisation de ce caractère fondamental

simplifie les modes de description et de déformation des structures, et comment il est mis à contribution dans notre nouvelle approche de modélisation : BCE (Biopolymer Chain Elasticity).

I.3 Les outils de description et de comparaison des structures, complexité et échelle de travail

La description et la comparaison des structures font partie des objets de la modélisation moléculaire. Les approches de modélisation fondées sur les principes de mécanique moléculaire, proposent généralement plusieurs conformations qui satisfont les critères de moindre énergie et de respect des données structurales dérivées de l'expérience. Afin d'analyser ces résultats, il est nécessaire de comparer les différentes conformations entre elles. Pour que les critères de description et de comparaison soient pertinents, il faut qu'ils expriment de façon intelligible les différentes conformations des molécules. Plusieurs outils existent : les approches qualitatives ou quantitatives.

I.3.1 L'approche qualitative

L'approche qualitative consiste à définir des familles de structures sur la base des orientations générales qu'adoptent les bases de la boucle relativement à la double hélice de la tige. De nombreux travaux ont été réalisés sur ce sujet, et une synthèse claire a été proposée par C. Pakleza dans sa thèse [4]. Nous n'aborderons donc pas cet aspect si ce n'est pour en rappeler les conclusions. La position des bases de la boucle est le premier critère de caractérisation de ces structures. Elles se placent dans l'espace selon des règles que l'on commence à entrevoir qui sont diverses qualitativement et quantitativement selon la composition de la séquence. Les tentatives de classement se sont heurtées à une grande diversité de formes observées. Sans cesse remises à jour [4] les familles de structures sont devenues de plus en plus complexes pour tenir compte des exceptions observées au cours des nouvelles résolutions structurales.

I.3.2 Les approches quantitatives

Pour obtenir des informations quantitatives sur la conformation des molécules deux types de grandeurs sont généralement employées. La première consiste à exprimer ou comparer les valeurs des positions, des longueurs ou des angles associées aux différents degrés de liberté utilisés pour déformer et décrire la molécule. La deuxième approche consiste à définir et à calculer des paramètres locaux de description géométriques des conformations. Ces deux méthodes permettent d'accéder à des informations différentes, de nature locale ou globale selon les cas.

I.3.2.1 Les outils dérivés de la description en coordonnées internes

Une approche pour comparer les conformations est d'utiliser les angles de torsion des molécules. Très répandue, cette approche permet notamment de comparer les chaînes sucre-phosphates de deux molécules. L'intérêt de cette méthode est de ne pas imposer l'identité des séquences, mais seulement le même squelette sucre-phosphate. Mais la limite de cette approche est justement son caractère local et restreint au squelette de la molécule. Elle ne permet pas de comparer l'ensemble de la structure. Pour cela, il faut se tourner vers d'autres outils tels que les comparaisons en coordonnées cartésiennes ou les comparaisons d'orientation de bases.

I.3.2.2 Les outils dérivés de la description cartésienne

La comparaison des coordonnées atomiques de deux molécules peut être un autre outil de comparaison des conformations. Le modélisateur est alors confronté à plusieurs difficultés :

La première est le nombre extrêmement important de coordonnées atomiques. Dans une macromolécule, leur grand nombre rend la comparaison des coordonnées, une à une, inintelligible et donc difficilement exploitable. Ceci peut être contourné par le calcul d'une quantité moyenne globale tel qu'un RMSd (*cf.* ANNEXE : A.1) sur l'ensemble ou une partie des coordonnées atomiques. Cependant même ce calcul n'est pas évident, car sa pertinence est conditionnée par l'étape préalable et indispensable de superposition des structures. Le choix du référentiel commun de placement des structures soulève une autre difficulté. Il a en effet un impact direct sur la comparaison des coordonnées de deux systèmes. Si les deux molécules sont très

éloignées ou orientées de façon différente, le calcul d'une telle grandeur n'a aucun intérêt scientifique. Pour comparer deux structures, il faut donc les superposer. Or la superposition de deux structures déformées et déformables est d'abord une question de point de vue.

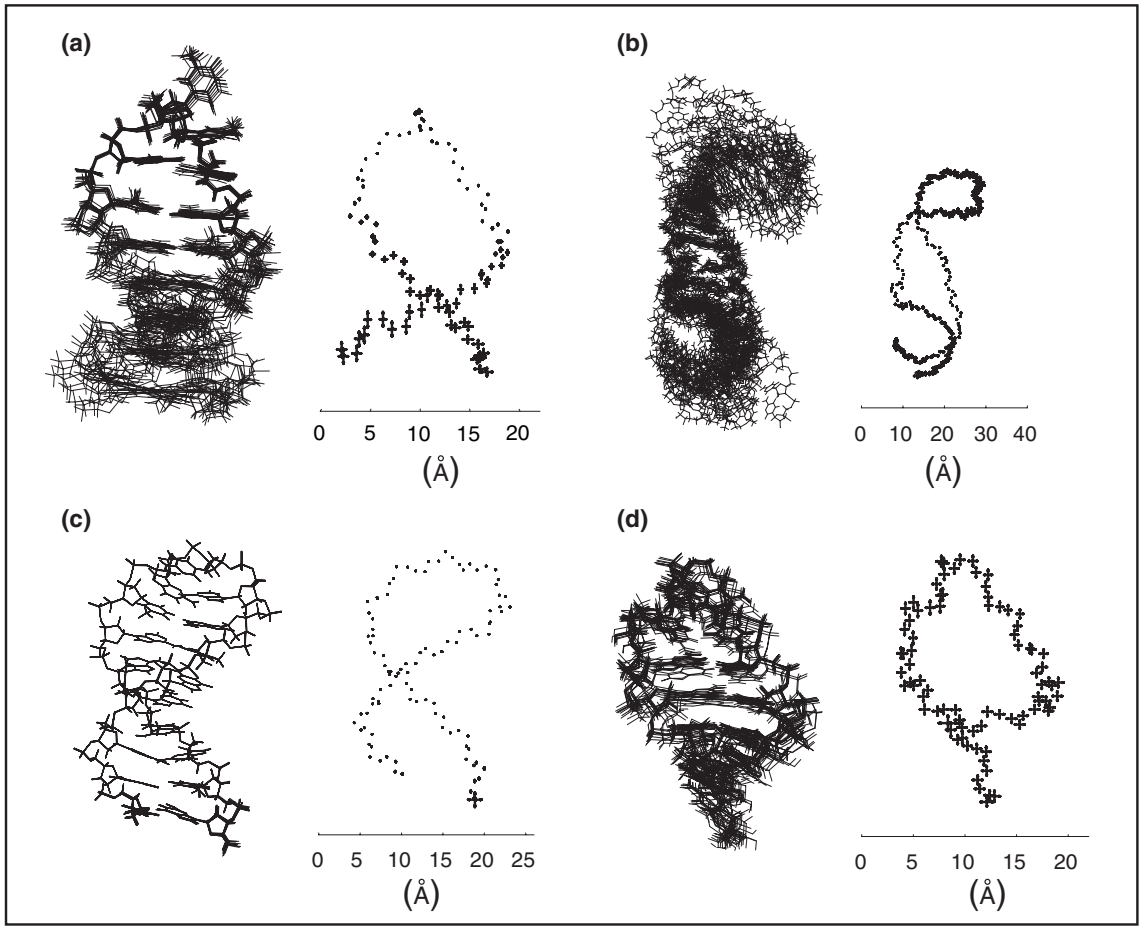


Fig. I.3 : ***Superposition des conformations données par les fichiers PDB :*** *Représentation des dix premières conformations (à gauche) et des déviations standards des positions des atomes principaux de la chaîne sucre-phosphate centrées sur leur position moyenne (à droite) calculées sur les n conformations du fichier PDB. (a) n=10, tétra-boucle d'ADN 1AC7-GTTA, (b) n=10, tétra-boucle d'ARN 1B36-UUCG, (c) n=10, tri-boucle d'ADN 1XUE-GCA, (d) n=19, tétra-boucle d'ARN 1C0O-UUCG. Les échelles des graphes sont différentes. Une échelle en Å est donnée à titre de comparaison.*

Comme le montre les superpositions des conformations contenues dans plusieurs fichiers PDB, les auteurs font des choix non neutres qui influent sur la vision que nous avons de la déformation des molécules. Dans la figure I.3, nous présentons 4 jeux de structures issues de quatre fichiers PDB. Dans chaque cas, les conformations sont affichées comme les auteurs les ont placées. Selon les cas, les structures peuvent sembler présenter une flexibilité localisée plutôt sur la tige (Fig. : I.3.a), plutôt dans

la boucle (FIG. : I.3.b), sur l'ensemble de la structure (FIG. : I.3.d). Dans certains autres cas aucune flexibilité n'est apparente (FIG. : I.3.c). Ces observations semblent à chaque fois pertinentes, pourtant elles sont le fruit d'artefacts (dans les cas a, b et d au moins). En effet, lorsque l'on optimise la superposition sur une partie d'une molécule, il est normal que les autres parties soient moins bien ajustées. Il est donc important d'avoir à l'esprit que la comparaison n'a de sens que sur les parties superposées. Ainsi, si l'on superpose les tiges, le seul RMSd ayant un sens porte sur les atomes de la tige. Réciproquement, si l'on veut calculer un RMSd sur la boucle, il est impératif de superposer préalablement les structures sur les atomes de la boucle et seulement ceux-là. Seuls les atomes utilisés pour superposer les structures doivent être utilisés pour calculer un RMSd. Il est juste ici de dire que l'on ne peut comparer deux parties d'un objet déformé et déformable que si, ces deux parties ont été préalablement superposées.

Ces contraintes proviennent en partie de ce que les déformations d'une structure par rapport à une autre peuvent être de nature locale ou globale. C'est-à-dire que la déformation peut être ponctuelle et avoir une incidence sur l'aspect global de la molécule, toutes les autres parties de la molécule étant identiques, ou réciproquement provenir de faibles modifications locales distribuées sur l'ensemble de la molécule. Pour identifier ces différentes composantes, il faut comparer les conformations à différentes échelles. C'est-à-dire qu'il faut étudier localement les différences ou les ressemblances avec des calculs locaux, et de même à l'échelle globale. Seule une approche multi-échelle peut rendre compte des différentes contributions à la déformation des structures avec une méthode utilisant des coordonnées cartésiennes.

I.3.2.3 Les outils dérivés de l'analyse des structures en hélices

Historiquement, l'analyse des structures d'acides nucléiques s'est d'abord focalisée sur les structures en double hélice. Pour caractériser les géométries des appariements et des empilements des plateaux de paires de bases dans les hélices, des paramètres descriptifs quantitatifs ont été définis (*cf.* ANNEXE : A.2). Ils correspondent au calcul de 3 paramètres de translations et 3 paramètres de rotations dans le repère cartésien de Cambridge (*cf.* PART. : II.2.2). Ils permettent de définir la position d'un plateau de paire de bases par rapport à un autre, ou de définir la position d'une base par rapport à une autre à l'intérieur d'un plateau de paire de bases appariées.

L'avantage de ces paramètres, à l'image des angles de torsion et à la différence des

coordonnées cartésiennes, est qu'il n'y a pas besoin de placer les molécules comparées dans un référentiel commun. Ces grandeurs sont internes à une molécule donnée et sont indépendantes du référentiel dans lequel elles sont calculées, à l'image des coordonnées internes. Elles ne sont pas, toutefois indépendantes de la méthode de calcul de ces paramètres [51]. Le corrolaire est que ces grandeurs ne permettent pas de comparer les conformations globales. En effet, bien que définies sur des objets plus gros (un nucléotide ou une paire de bases au lieu d'un angle impliquant quatre atomes), elles ne caractérisent que des positions relatives locales par rapport à l'entité précédente (base ou plateau) ou à l'axe de l'hélice local. Une autre limite importante pour notre étude est que ces grandeurs ont été définies pour analyser des doubles hélices dans lesquelles les appariements sont bien définis. Elles ne sont pas adaptées à l'étude des boucles où les appariements peuvent être non Watson-Crick voire inexistants. Il est donc nécessaire de définir de nouveaux outils pour caractériser quantitativement les conformations des parties en boucle.

I.3.3 Structures résolues et bases de données

Aujourd'hui, les bases de données de structures tridimensionnelles sont un outil indispensable dans l'analyse et la comparaison des structures. Elle s'accroissent de façon exponentielle comme le montrent les statistiques de la Protein Data Bank (PDB) [33], la banque générale de données de structures tridimensionnelles résolues (radiocristallographie aux rayons X et RMN) des molécules biologiques (*cf.* FIG. : I.4).

Dédiée à l'origine aux protéines, la PDB recueille également les structures d'acides nucléiques. Le 18 janvier 2005, sur un total de 29211 entrées, 26571 sont des structures de protéines et seulement 2627 concernent les acides nucléiques (1401 structures d'acides nucléiques seuls + 1226 complexes protéine-acide nucléique). Les structures d'acides nucléiques sont donc plus de 10 fois moins nombreuses que celles de protéines. Cette différence s'explique par un retard de vingt années environ entre l'étude des protéines et celle des acides nucléiques. C'est ce que l'on observe effectivement pour toute une série de dates repères comme le premier séquençage, la première synthèse automatique ou les premières déterminations de structures tridimensionnelles (*cf.* TAB. : I.1). La NDB (Nucleic Data Bank) est une banque de données de structures tridimensionnelles dédiée aux acides nucléiques et à leur description spécifique.

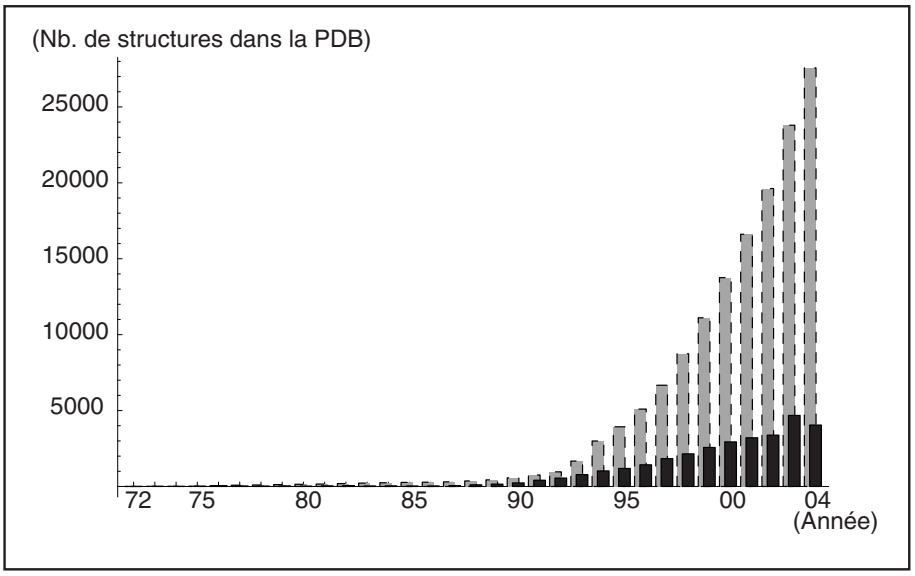


Fig. I.4 : ***Statistiques du nombre de structures accessibles dans la PDB :*** *en noir, le nombre de structures déposées chaque année depuis 1972 ; en gris, le total cumulé des structures disponibles chaque année jusqu'en 2004.*

Méthodes	**Protéines**	**ADN/ARN**
Premiers Séquençages	1953 Insuline (Sanger)	1975-77 (Maxam & Gilbert, Sanger)
Premières Synthèses Automatiques	1962 (Merrifield)	1981 Phosphoramidite
Premières Structures 3D	1959 Myoglobine	1976 ARNt (Klug, Kim & Rich) 1981 ADN-B (Dickerson) [52]
Début de la croissance exponentielle du nombre de structures déposées dans la PDB	1975 1972 : Création de la PDB	1995 1992 : Création de la NDB [53]

Tab. I.1 : ***Dates des grandes étapes de l'étude des protéines et des acides nucléiques.***

Longtemps cantonnée essentiellement aux variations de la double hélice d'ADN et aux ARNt, la croissance exponentielle du nombre de structures résolues d'acides nucléiques s'accompagne d'une diversification de leur taille, dont le spectre va désormais de petites structures (boucles, pseudo-noeuds, ribozymes, ...) à, récemment, de très grosses structures (le nucléosome et le ribosome). Aujourd'hui, même si le nombre de structures en épingles à cheveux est encore restreint, la PDB devient depuis 1995-2000 une source formidable de données pour étudier la variabilité

des conformations des acides nucléiques en fonction de la séquence.

Dans le cadre de notre étude nous avons utilisé de nombreuses structures issues de la PDB. Nous avons ainsi pu confronter nos résultats de modélisation aux structures obtenues par différents auteurs qui ont utilisé des approches différentes de la nôtre. Dans le chapitre suivant nous présentons ces structures dans l'ordre d'apparition chronologique de notre travail. Certaines de ces structures sont étudiées dans plusieurs chapitres de la thèse.

I.4 Les structures tridimensionnelles d'épingles à cheveux

Dans cette partie nous allons caractériser la structure tridimensionnelle des structures en épingle à cheveux d'acides nucléiques. Nous présentons dans un premier temps les structures utilisées au cours de notre étude, puis nous aborderons successivement les caractéristiques de la trajectoire de leur chaîne sucre-phosphate et la géométrie des appariements dans les tiges et les boucles.

I.4.1 Les structures étudiées

I.4.1.1 Première exploration générale : (Chapt. III)

Notre première étude vise à évaluer la capacité de notre approche BCE (*cf.* Part.: II)à calculer correctement la trajectoire de différentes boucles en épingle à cheveux. Nous avons restreint notre première étude aux structures comportant un appariement non Watson-Crick entre les bases extrémales de la boucle (nucléotides N_1-N_3 pour les tri-boucles et N_1-N_4 pour les tétra-boucles). Nous avons sélectionné huit molécules résolues par différents auteurs et disponibles dans la PDB (*cf.* Tab. : I.2). Nous avons comparé leurs structures à celles obtenues avec notre approche. Afin de prendre en compte différents facteurs importants pour la géométrie des boucles comme la nature de la chaîne (ADN ou ARN) et la longueur de la boucle, nous avons sélectionné quatre épingles à cheveux d'ADN dont la longueur de la boucle varie entre trois et quatre nucléotides et quatre épingles à cheveux d'ARN de séquence *c*-UUCG-*g*.

Code PDB	Auteurs & Année & Référence	Séquence de la molécule PDB	N1	N2	N3	N4
ADN tétra-boucle						
1AC7	van Dongen *et al.* (97) [54]	d(...ccta-GTTA-tagg...)	e/G	G	G/sol	e/p
ADN tri-boucle						
1BJH	Chou *et al.* (96) [47]	d(gtac-AAA-gtac)	e/G	G	e	-
1XUE	Zhu *et al.* (96) [55]	d(...gaat-GCA-atgg...)	e/G	G	e	-
1ZHU	Zhu *et al.* (95) [56]	d(caat-GCA-atg)	e/G	G	e	-
ARN tétra-boucle						
1AUD	Allain *et al.* (97) [57]	r(...gucc-UUCG-ggac...)	e/G	p/sol	G	e
1B36	Butcher *et al.* (99) [58]	r(...gcgc-UUCG-gcgc...)	e/G	p/sol	G	e
1C0O	Colmenarejo & Tinoco (99) [59]	r(...gguc-UUCG-gguc...)	e/G	p/sol	G	e
1HLX	Allain & Varani (95) [60]	r(...uaac-UUCG-guug...)	e/G	p/sol	G	e

Tab. I.2 : ***Structures sélectionnées de la PDB étudiées dans le chapitre III :*** *Code d'identification des structures PDB, auteurs et années de publication de la structure, séquence de la boucle et des 4 paires de bases de la tige; Orientation des bases de la boucle en référence à la double hélice de la tige. G : dans le Grand sillon; p : dans le petit sillon; e : empilées sur la partie centrale de la double hélice; sol : dans le solvant.*

Les trois molécules en tri-boucles d'ADN sélectionnées ont des séquences différentes mais elles appartiennent au même groupe structural car elles comportent un certain nombre de similarités (*cf.* Fig. : I.5). Elles possèdent toutes des appariements isomorphes Purine-Purine entre les bases N_1 et N_3 de la boucle (*cf.* Part. : I.4.3.4). Pour former l'appariement dans la boucle, les bases N_1 et N_3 s'empilent sur le dernier plateau de paire de bases de la tige. La base N_1 semble un peu déplacée dans le grand sillon (G), alors que la base N_3 semble empilée directement au dessus de la dernière base de l'hélice. La base centrale de la boucle N_2, s'empile sur le plateau formé par les bases N_1 et N_3, plutôt du côté du grand sillon sur la base, c'est-à-dire au dessus de la base N_1. Du point de vue fonctionnel, ces trois molécules présentent des intérêts biologiques importants et très différents.

Sous l'influence de contraintes de torsion sur la double hélice d'ADN [6] des structures cruciformes formées de 2 motifs en épingle à cheveux peuvent apparaître. La séquence 1BJH-AAA [47] est complémentaire à la séquence -TTT- qui forme un motif en épingle à cheveux. Ces deux motifs et plus particulièrement le motif TTT, sont importants pour l'intégration du génome de l'adénovirus AAV2 (Adénoassociated virus 2), et est étudié dans le cadre de la recherche de vecteurs non pathogènes en thérapie génique.

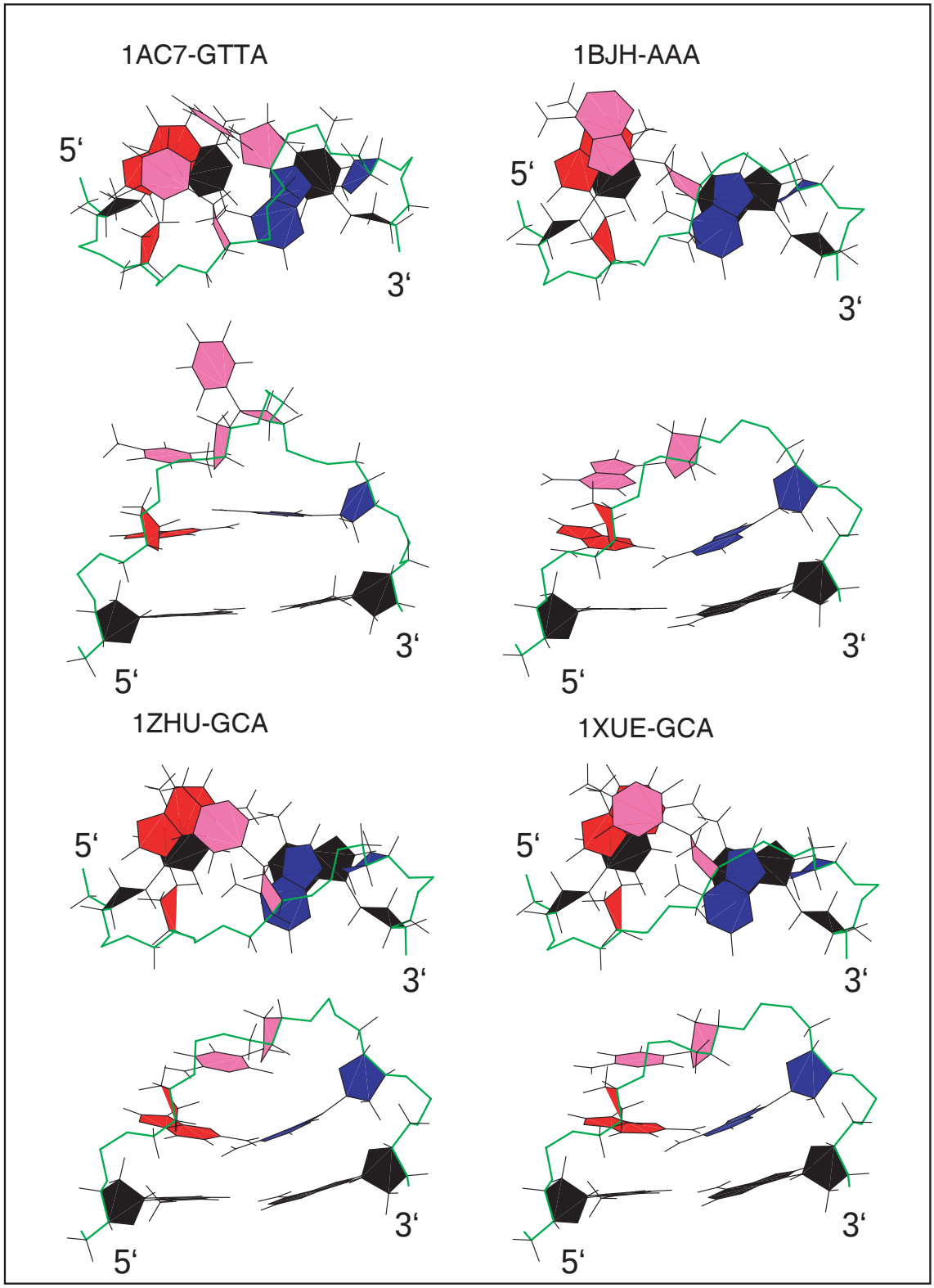


Fig. I.5 : *Représentation des structures PDB des épingles à cheveux d'ADN sélectionnées pour la première étude :* *Représentation de la première conformation des fichiers PDB des molécules 1AC7-GTTA [54], 1BJH-AAA [47], 1ZHU-GCA [56] et 1XUE-GCA [55]. En haut : le long de l'axe de la double hélice. En bas : vue depuis le petit sillon. Le dernier plateau de paire de bases de la tige est coloré en noir, la base en 5' de la boucle en rouge, la ou les bases centrales en violet et la base en 3' de la boucle en bleu. La chaîne sucre-phosphate est figurée en vert.*

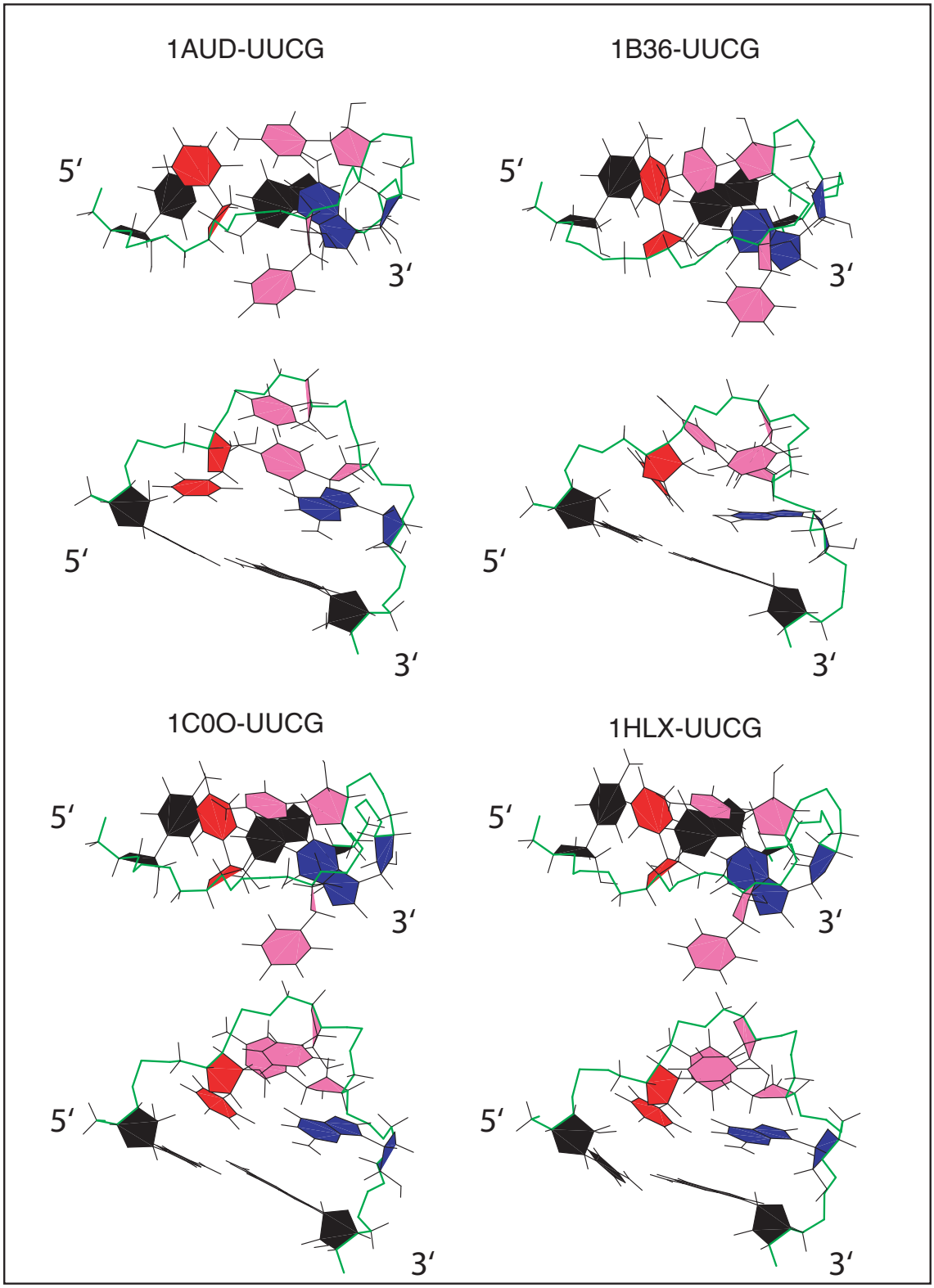


Fig. I.6 : ***Représentation des structures PDB des épingles à cheveux d'ARN sélectionnées pour la première étude :*** *Représentation de la première conformation des fichiers PDB des molécules 1AUD-UUCG [57], 1B36-UUCG [58], 1C0O-UUCG [59] et 1HLX-UUCG [60]. En haut : le long de l'axe de la double hélice. En bas : vue depuis le petit sillon. Le dernier plateau de paire de bases de la tige est coloré en noir, la base en 5' de la boucle en rouge, la ou les bases centrales en violet et la base en 3' de la boucle en bleu. La chaîne sucre-phosphate est figurée en vert.*

Les séquences -GCA- des molécules 1XUE-GCA [55] et 1ZHU-GCA [56] sont rencontrées dans les séquences répétées des centromères. Elles participent à l'apparition de pathologies génétiques liées à la répétition de triplets dans certains gènes qui codent alors pour des protéines défectueuses.

La structure de la molécule 1AC7-GTTA [54] présente des similarités avec les structures des tri-boucles d'ADN (*cf.* FIG. : I.5). L'appariement entre les bases extrémales de la boucle est le même avec la différence d'un très léger déplacement de la base N_4 dans le petit sillon. La base N_2 s'empile comme dans les tri-boucles sur le plateau formé des bases N_1 et N_4, plutôt au dessus de la base N_1. La base additionnelle N_3 est placée dans le solvant du côté du grand sillon. Du point de vue fonctionnel, ce type de séquence est rencontré dans les centromères et les télomères.

Les structures des tétra-boucles d'ARN 1AUD-UUCG [57], 1B36-UUCG [58], 1C0O-UUCG [59] et 1HLX-UUCG [60] sont toutes très proches (*cf.* FIG. : I.6). Les bases appariées U_1 et G_4 s'empilent sur le dernier plateau de paire de bases de la tige. Cet appariement présente un fort angle de buckle (*cf.* ANNEXE : A.2). La base du nucléotide C_3 s'empile sur la base du nucléotide U_1 du côté du grand sillon, et la base U_2 est tournée dans le solvant du côté du petit sillon. Du point de vue fonctionnel, ces structures sont intéressantes à divers points de vue.

L'épingle à cheveux de la molécule 1AUD-UUCG est présente sur un site d'inhibition de la polyadénylation reconnu par le domaine RNP (Ribo Nucléo Protéique)de la protéine humaine UA1 [61]. La structure 1B36-UUCG a été utilisée pour stabiliser la structure des deux sous-domaines essentiels pour l'ativité catalytique du ribozyme. De la même façon, l'épingle à cheveux de la molécule 1C0O-UUCG a été utilisée pour faciliter l'étude d'un ARN en stabilisant sa structure. Cet ARN est issu de l'hélice P5 du centre catalytique d'un intron du groupe I du ribozyme de Tetrahymena. Cette étude visait à étudier la fixation sur ce site d'un ligand métallique hexamine Cobalt III. La séquence de la molécule 1HLX-UUCG est la séquence de fermeture de l'hélice P1 des introns autoexcisables du groupe I.

I.4.1.2 Deuxième exploration : (CHAPT. IV)

Dans ce chapitre nous avons restreint notre étude aux tri-boucles d'ADN comportant un appariement dans la boucle. Nous avons complété la première exploration qui comportait 3 structures de ce type - 1BJH-AAA, 1XU-GCA et 1ZHU-GCA - avec

5 nouvelles structures plus récentes (*cf.* TAB. : I.3). Les structures 1JVE-GAA [62], 1PQT-GAA [63] et 1P0U-GAC [64] proviennent de la PDB. La structure ATC [65] nous a été communiquée par ses auteurs[1]. Elle est également modélisable par minimisation à partir des valeurs d'angles de torsion publiées [65, 66]. La dernière structure, AGC [67], a été reconstruite par minimisation sous contraintes d'angles de torsions d'après les données publiées.

Code PDB	**Auteurs & Année & Référence**	**Séquence de la molécule PDB**	**N1**	**N2**	**N3**	**App. N_1-N_3**
A···A						
1BJH	Chou *et al.* (96) [47]	d(gtac-AAA-gtac)	e/G	G	e	ANTI/ANTI
G···A						
1XUE	Zhu *et al.* (96) [55]	d(...gaat-GCA-atgg...)	e/G	G	e	ANTI/ANTI
1ZHU	Zhu *et al.* (95) [56]	d(caat-GCA-atg)	e/G	G	e	ANTI/ANTI
1PQT	Padrta *et al.* (02) [63]	d(gc-GAA-gc)	e/G	G	e	ANTI/ANTI
1JVE	Ulyanov *et al.* (02) [62]	d(...taac-GAA-gtta...)	e/G	G	e	ANTI/ANTI
G···C						
1P0U	Chin et al. (03) [64]	d(...gatc-GAC-gatg...)	e/G	G	e	ANTI/SYN
A···C						
ATC	Amir-Aslani *et al.* (96) [65]	d(caat-ATC-atg)	e/G	G	e	ANTI/ANTI
AGC	Chou *et al.* (99) [67]	d(...gtac-AGC-gtac...)	e/G	G	e	ANTI/ANTI

TAB. I.3 : ***Structures sélectionnées de la PDB étudiées dans le chapitre IV :*** *Code d'identification des structures PDB, auteurs et années de publication, séquence de la boucle et des 4 paires de bases de la tige, orientation des bases de la boucle en référence à la double hélice de la tige ; G : dans le Grand sillon; p : dans le petit sillon; e : empilées sur la partie centrale de la double hélice; sol : dans le solvant.*

Du point de vue structural, ces nouvelles structures en tri-boucles ressemblent fortement aux structures de la première exploration (*cf.* FIG. : I.7). Les bases de la boucle sont positionnées de façon similaire au-dessus de la double hélice de la tige comme le montre le tableau I.3. Une exception notable dans la conformation des bases est à noter pour l'épingle à cheveux 1P0U-GAC [64] (*cf.* FIG. : I.8) dont la dernière base de la boucle (*i.e.* base en 3' de la boucle) est en conformation SYN, ce qui correspond à une rotation de près de 180° de la base autour de la liaison glycosidique. Ce retournement de la base n'affecte pas la forme globale de la boucle. Pour toutes les autres, les géométries d'empilement sont similaires et les appariement se ressemblent à la séquence près.

[1]Nous remercions chaleureusement MM. S. FERMANDJIAN et O. MAUFFRET pour les fichiers de coordonnées de la structure -ATC-.

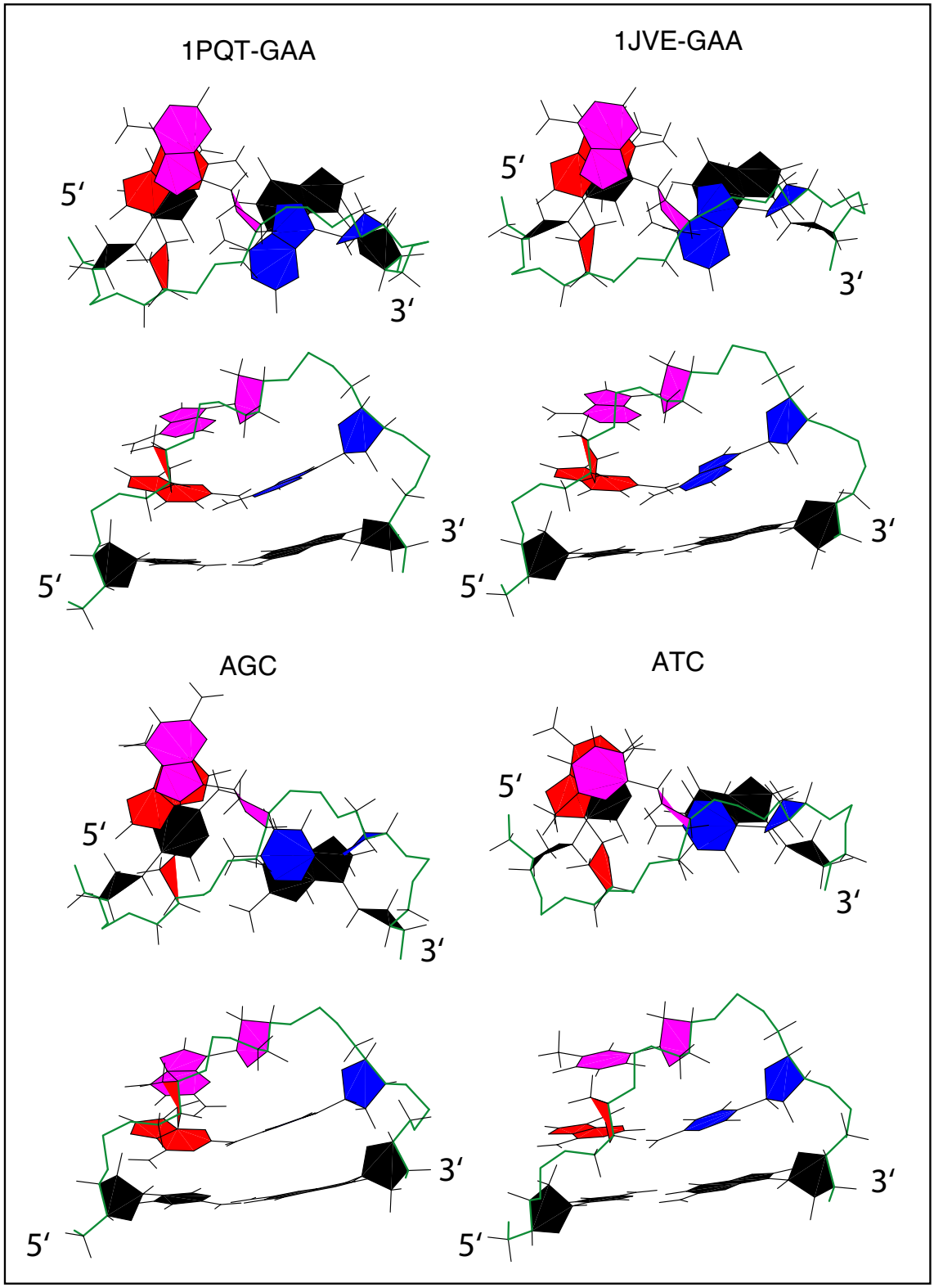


Fig. I.7 : *Représentation des structures PDB des tri-boucles d'ADN ajoutées dans la seconde étude (suite) :* *Représentation des premières structures des fichiers PDB des molécules 1PQT-GAA [63], 1JVE-GAA [62], AGC [67] et ATC [65]. En haut : le long de l'axe de la double hélice. En bas : vue depuis le petit sillon. Le dernier plateau de paire de bases de la tige est coloré en noir, la base en 5' de la boucle en rouge, la ou les bases centrales en violet et la base en 3' de la boucle en bleu. La chaîne sucre-phosphate est figurée en vert.*

Ces cinq nouvelles structures ont des fonctionnalités biologiques très différentes : L'épingle à cheveux 1PQT-GAA [63] donne la conformation d'une boucle extraordinairement stable (Tm de 76°C [17, 68]). Cette stabilité extrême a été mise à profit dans l'épingle à cheveux 1JVE-GAA [62] afin de faciliter l'étude par RMN d'une séquence riche en AT provenant des régions télomériques d'un virus à ADN double brin, le vaccinia virus. Ses propriétés de résistance aux nucléases, sa présence dans les origines de réplication du phage $\phi\chi174$ et du virus herpes simplex, dans la région promotrice d'un gène de réponse à la chaleur d'E. Coli, dans les gènes des ARN ribosomaux, son rôle majeur dans les répétitions des maladies à triplets répétés confèrent à cette séquence un grand intérêt biologique. La séquence de l'épingle à cheveux 1P0U-GAC provoque des décalages du cadre de lecture des gènes et peut conduire à l'apparition de maladies neurodégénératives graves comme le syndrome de l'X fragile, lorsque la séquence est répétée de nombreuses fois. Les épingles à cheveux AGC et ATC font partie de domaines structuraux invariants importants. La boucle AGC a été trouvée dans une enzyme d'ADN clivant l'ARN [67]. La séquence ATC est présente à un site de coupure de la topoisomérase II du thymus de veau [65,66].

En résumé, ces boucles peuvent être directement impliquées dans des processus biologiques ou bien utilisées pour leur stabilité afin de faciliter l'étude de séquences avoisinantes. L'étude de ces structures est donc fondamentale pour mieux comprendre et pour aborder les phénomènes de reconnaissance et leurs stabilités.

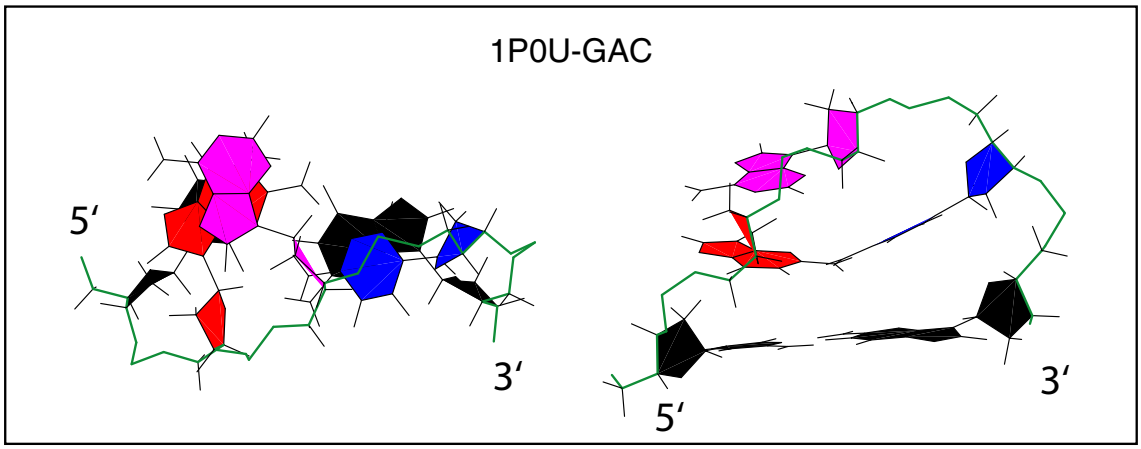


Fig. I.8 : ***Représentation des structures PDB des tri-boucles d'ADN ajoutées dans la seconde étude :*** *Représentation de la première structure du fichier PDB de la molécule 1P0U-GAC [64]. À gauche : le long de l'axe de la double hélice. À droite : vue depuis le petit sillon. Le dernier plateau de paire de bases de la tige est coloré en noir, la base en 5' de la boucle en rouge, la ou les bases centrales en violet et la base en 3' de la boucle en bleu. La chaîne sucre-phosphate est figurée en vert.*

I.4.2 Caractéristiques des trajectoires de la chaîne sucre-phosphate

Les tiges des épingles à cheveux ne se distinguent pas *a priori* des doubles hélices habituellement rencontrées et décrites dans la littérature. À ce titre elles peuvent potentiellement présenter toutes les variations conformationnelles rencontrées dans les doubles hélices. La description des déformations des hélices a fait l'objet de très nombreuses publications [1, 69, 70] et ne sera pas rappelée ici.

Dans les épingles à cheveux, l'observation des structures tridimensionnelles des boucles d'ADN ou d'ARN publiées met en évidence un certain nombre de caractéristiques communes (*cf.* FIG. : I.9) :

- la trajectoire de la chaîne sucre-phosphate est déformée continuement.
- les valeurs des angles de torsion dans la boucle sont proches de celles rencontrées dans l'hélice à quelques exceptions près (1, 2 ou 3 angles de torsion) au voisinage du resserrement de la boucle aussi appelé "sharp-turn".

Pour l'ADN, la trajectoire de la chaîne sucre-phosphate de la boucle passe au-dessus des plateaux de paire de bases de la tige. Vue le long de l'axe de la double hélice, c'est à dire dans le cercle perpendiculaire à l'axe, elle adopte une forme caractéristique dite en "S" ou en yin-yang. Pour l'ARN, la trajectoire de la chaîne sucre-phosphate longe le côté du cylindre associé à la double hélice. La trajectoire de la chaîne sucre-phosphate adopte une forme en "langue de chat" qui s'enroule sur le côté du cylindre. Dans tous les cas la structure générale de la boucle est très compacte et tend, à minimiser les interactions hydrophobes déstabilisantes.

Cette homogénéité des trajectoires des chaînes sucre-phosphates se retrouve dans les tri-boucles d'ADN étudiées, comme le montre la superposition (*cf.* FIG. : I.10) et le RMSd très faible de 0,63 Å calculé sur le squelette de la boucle et des deux derniers plateaux de paires de bases de la tige de huit des molécules sélectionnées.

Du fait de la symétrie pseudo-dyadique de la géométrie des doubles hélices, la trajectoire de la chaîne sucre-phosphate de la boucle doit, pour joindre les deux extrémités de la tige, effectuer un "aller-retour" dans l'espace. Ce repliement de la boucle se traduit par une une courbure brusque de la trajectoire appelé "sharp-turn"

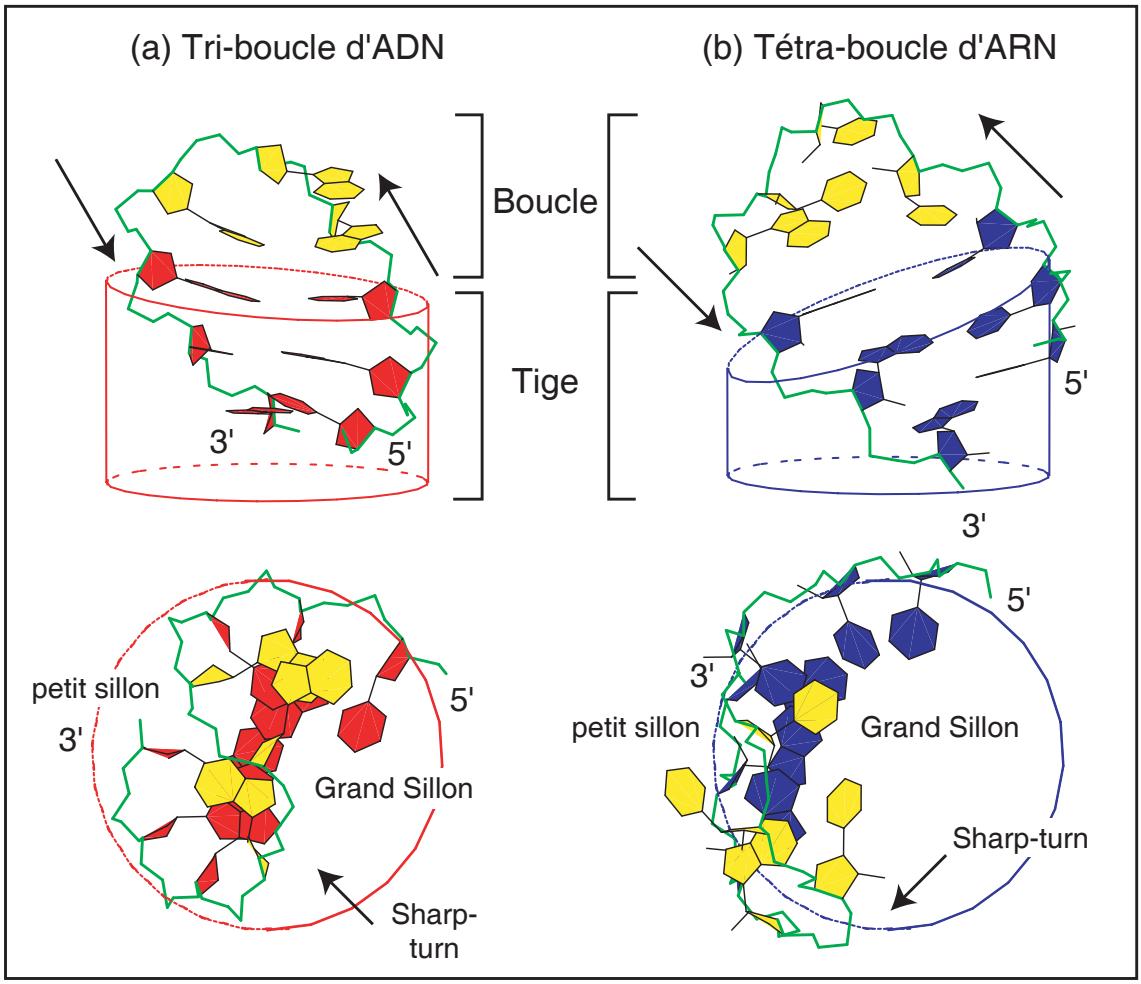


Fig. I.9 : ***Structure tridimensionnelle des motifs en épingle à cheveux :*** *Représentation, (a), de la première conformation du fichier PDB de la tri-boucle d'ADN 1BJH-AAA et, (b), de la première conformation du fichier PDB de la tétra-boucle d'ARN 1AUD-UUCG. En haut représentation de la structure vue de côté depuis le grand sillon, et en bas vue le long de l'axe de la double hélice. Les bases de la boucle sont représentées en jaune. Les bases de la tige sont représentées en rouge pour la molécule d'ADN et en bleu pour la molécule d'ARN. Sur les figures les cylindres associés aux doubles hélices donnent une idée des différences de trajectoire entre les boucles d'ADN (au-dessus du cylindre) et d'ARN (sur le côté du cylindre).*
N.B. : Cette figure est indispensable pour introduire l'objet de nos études; comme nous le verrons plus loin, elle constitue également un résultat important, car ces deux molécules, très différentes, peuvent être introduites ici dans un même système de référence grâce à l'application de BCE.

du côté 3' de la boucle. D'un point de vue structurale, cette zone de fort repliement est associée à des variations importantes des angles de torsion de la chaîne sucre-phosphate des nucléotides concernés. Du point de vue de la RMN, cette zone est associée à une rupture de la continuité des attributions des pics séquentielles.

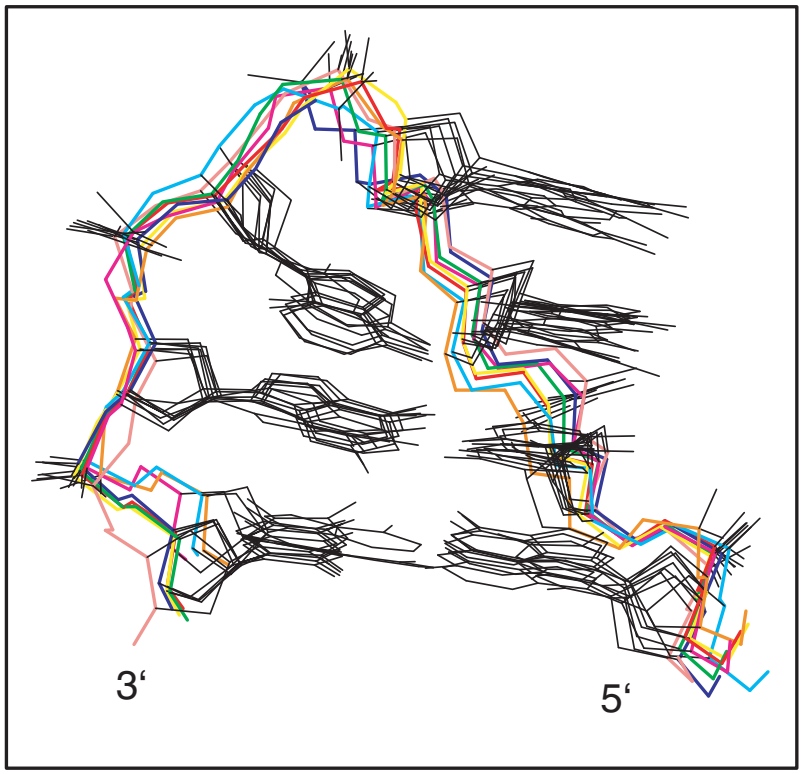


Fig. I.10 : ***Superposition des huit tri-boucles d'ADN données dans le tableau I.3 comportant des appariements dans la boucle [47, 55, 56, 62–65, 67] :*** *Représentation de la boucle et des deux derniers plateaux de paires de bases sous-jacents. Les conformations sont ajustées dans l'espace pour minimiser la distance entre atomes homologues de la chaîne sucre-phosphate. En noir, les bases; en couleur, le squelette de la chaîne sucre-phosphate. 1BJH-AAA en rouge, 1XUE-GCA en vert, 1ZHU-GCA en bleu, 1JVE-GAA en violet, 1PQT-GAA en rose, 1P0U-GAC en cyan, AGC en jaune et ATC en orange.*

I.4.3 Les plateaux de paires de bases dans les structures en épingles à cheveux

De nombreuses épingles à cheveux décrites dans la littérature présentent des appariements dans la partie en boucle. Comme dans les doubles hélices, ils sont formés par l'association de deux bases dont la géométrie quasi-coplanaire est stabilisée par une ou plusieurs liaisons hydrogène. Ils se distinguent des appariements canoniques des hélices par la nature des bases et des groupements donneurs et accepteurs de protons engagés dans les liaisons hydrogène. Ils sont couramment appelés mésappariements. Considérés comme déstabilisants dans les hélices, ils sont, dans les motifs en épingle à cheveux, considérés comme un facteur positif très important de la stabilité des boucles.

I.4.3.1 Les plateaux de paires de bases appariées et mésappariées

De façon générale, on appelle "appariements" les associations coplanaires proposées par Griffith, Watson et Crick [30], entre une Adénine et une Thymine (Uracyl dans l'ARN) ou entre une Guanine et une Cytosine. Ces associations sont stabilisées par plusieurs liaisons hydrogène entre un proton lié covalemment à une base et un doublet libre d'électrons portés par un atome electronégatif (Azote ou Oxygène) de l'autre base. Les appariements Watson-Crick sont rencontrés dans les doubles hélices A et B canoniques *(cf.* FIG. : I.11 et ANNEXE : B.1). Les liaisons hydrogène qui stabilisent ces plateaux sont bien définies et sont reportées dans le tableau I.4. À ces liaisons hydrogène canoniques il faut ajouter les liaisons de type CH$\cdots$O reportées dans l'étude de certaines structures par dynamique moléculaire [38]. Ces liaisons interviennent dans les plateaux Watson-Crick A-T, entre le proton H2 de la thymine et l'oxygène O2 de l'adénine. Bien que de moindre énergie que les liaisons hydrogène canoniques, elles semblent participer à la stabilisation de ces appariements [38, 71–73].

Par ailleurs, il existe d'autres appariements dit "non-canoniques" qui interviennent également entre les bases A et T(U) ou C et G, tels que les appariements Reverse Watson-Crick, (Reverse-) Hoogsteen et (Reverse-) Wooble. Ils se différencient des premiers par la géométrie de l'orientation des bases et par la nature des liaisons hydrogène qui les stabilisent (*cf.* ANNEXE : B.1).

	Plateaux A-T	Plateaux G-C
Liaisons Hydrogène	ADE[HN6]-[O4]THY	GUA[O6]-[HN4]CYT
	ADE[N1]-[H3]THY	GUA[H1]-[N3]CYT
	ADE[H2]-[O2]THY	GUA[HN2]-[O2]CYT

TAB. I.4 : ***Liaisons hydrogène des appariements canoniques Watson-Crick :*** *Description des protons et des accepteurs de protons dans les plateaux de paire de bases Watson-Crick.*

En contraste avec les appariements Watson-Crick, on appelle "mésappariement" tous les autres appariements. Ces mésappariements adoptent dans les hélices des géométries très différentes selon les bases impliquées et l'environnement nucléotidique. La présence de ces mésappariements est en général associée à des déformations partielles de la chaîne sucre-phosphate de l'hélice, et à une déstabilisation de la structure moléculaire (les températures de fusion des doubles brins sont généralement plus basses) [71, 74–78]. De nombreuses familles ont été

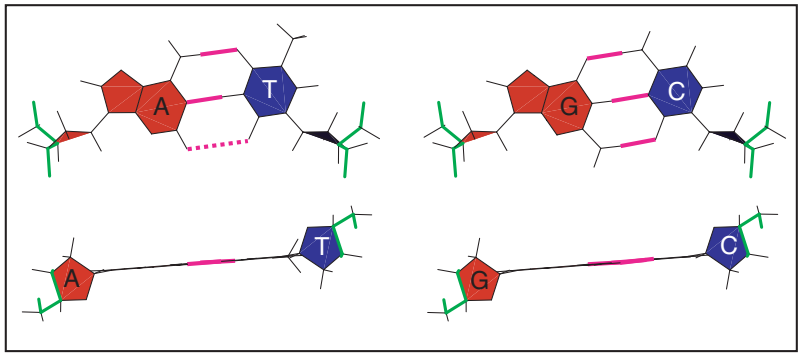


Fig. I.11 : *Appariements Watson-Crick canoniques :* *Vue gauche, appariement A-T; Vue droite, appariement GC dans une double hélice d'ADN-B; en haut, appariement vu le long de l'axe de l'hélice; en bas, appariement vu de côté du petit sillon; en vert, la chaîne sucre-phosphate; en rouge, les bases puriques; en bleu, les bases pyrimidiques; en violet, les liaisons hydrogène.*

décrites (*cf.* Annexe : B.2, B.3 et B.4), dont certaines se rencontrent dans les boucles des épingles à cheveux [79].

I.4.3.2 Caractérisation des empilements

Dans les doubles hélices d'ADN-B et d'ARN-A, les plateaux formés par les paires de bases (més-)appariées s'empilent les uns sur les autres. Dans la suite de notre étude nous avons besoin de deux paramètres simplifiés pour décrire les empilements :

- la distance moyenne d'empilement séparant les deux plateaux, correspond à l'épaisseur moyenne d'un plateau de paire de bases et ne doit pas être confondue avec le déplacement en z ("Rise", *cf.* Annexe : A.2) qui est une projection de cette distance sur l'axe de l'hélice,
- l'angle $\eta_{plat.}$ caractérisant l'inclinaison totale entre deux plateaux empilés. Il est défini comme l'angle entre les deux vecteurs normaux aux plans des deux plateaux de paires de bases consécutifs. Il peut être décomposé dans le formalisme de la convention de Cambridge en une somme de contributions en Roulis ("Roll") qui est l'angle de rotation autour de l'axe y et en basculement ("Tilt", *cf.* Annexe : A.2) qui est l'angle de rotation autour de l'axe x.

Nous avons calculé cette distance et cet angle entre les plateaux adjacents pour huit structures de référence :

- Six structures [23–28] en hélices obtenues à partir de données de diffraction de fibres, et
- Deux structures obtenues à partir de minimisations dans un champ de force moléculaire (PARM94 du programme AMBER) [4, 80, 81].

Les valeurs calculées de ces distances sont reportées dans le tableau I.5 ci-après.

N° Arnott	**Réf.**	**Plateau**	**Distances**	$\eta_{plat.}$
ARN-A				
-	[4, 80, 81]	r(G-C)$_2$	3,335 Å	9,4°
20	[23, 27]	r(A-U)$_2$	3,287 Å	8,7°
ADN-A				
1	[23, 24]	d(G-C)$_2$	3,300 Å	12,3°
3	[23]		3.385 Å	15,0°
ADN-B				
-	[4, 80, 81]	d(G-C)$_2$	3,361 Å	2,9°
4	[23, 25]		3,379 Å	1,6°
6	[26]		3,457 Å	1,3°
46	[28]		3,364 Å	1,9°

Tab. I.5 : ***Distances moyennes séparant les plateaux de paires de bases consécutifs :*** *Type : Type de l'hélice; N° Arnott : Numéro d'identification dans la classification Arnott [23]; Réf. : Références bibliographiques de la structure; Plateau : Type de plateau testé; Distance : Distances séparant les deux plateaux de paires de bases consécutives. Elles sont calculées entre les points médians des atomes N_i[C6]-[C8]N_{-i} de chaque plateau; $\eta_{plat.}$: Angle entre les 2 plans définis par les plateaux consécutifs. Ils sont calculés comme l'ArcCosinus entre les vecteurs normaux des plans moyens des deux plateaux consécutifs.*

La distance moyenne d'empilement est calculée à partir des positions des atomes formant les cycles puriques et pyrimidiques de chaque base. Elle donne une idée de l'épaisseur de chaque cycle (*i.e.* du diamètre de van der Waals des atomes d'une paire de bases) et donc de la distance idéale moyenne entre deux cycles qui s'empilent. Cette distance entre les deux plateaux ne peut être plus courte : les contraintes stériques l'interdisent, car les atomes se chevaucheraient. La valeur communément admise pour cette distance, indépendamment de la nature de la chaîne (deoxy- ou ribonucléique), est de 3,34 Å, ce qui est proche des valeurs observées dans nos hélices de référence. Nous retiendrons cette valeur comme la distance minimale entre deux plateaux de paires bases. Elle est importante pour la résolution du problème posé au chapitre II.5 car elle va conditionner l'étape de redressement d'empilement des bases lors de l'exploration des conformations de bases appariées dans les tri-boucles

d'ADN (les atomes qui font partie de la base empilée ne doivent être ni trop loin, ni trop proche des atomes du dernier plateau de la tige (*cf.* PART : II.5.4).

Dans les hélices B les plateaux sont tous approximativement orthogonaux à l'axe de l'hélice. Il en résulte que les angles entre les plans des plateaux empilés $\eta_{plat.}$ sont faibles (entre 1,3° et 2,9°), puisque les plateaux sont globalement parallèles dans ces hélices. Dans les hélices A, les valeurs observées sont beaucoup plus importantes (entre 8,7° et 15°). Ceci s'explique par la géométrie de cette hélice où les plans des plateaux s'enroulent autour de l'axe de l'hélice en introduisant nécessairement un angle $\eta_{plat.}$ non nul entre les plans des plateaux empilés.

L'angle d'inclinaison moyen $\eta_{plat.}$ entre plateaux consécutifs est une constante dans les hélices régulières. La direction moyenne de n'importe quel plateau d'une hélice peut donc être calculée en connaissant seulement la direction de l'axe de l'hélice, l'orientation du plateau précédent et la valeur de $\eta_{plat.}$. Dans le cadre de la formation des mésappariements dans les tri-boucles d'ADN, cet angle permettra de déterminer l'orientation idéale du plateau de paire de bases mésappariées dans la boucle (*cf.* PART : II.5.2).

I.4.3.3 Les appariements dans les boucles des épingles à cheveux

Historiquement les parties simple-brins en boucle étaient considérées comme peu ou pas structurées par rapport aux double-brins en hélice. Cette vision a évolué avec la découverte, notamment dans l'ARN, de motifs en tige-boucles hyperstables (UUCG et GNRA [36, 82]). Leur résolution structurale a montré que ces boucles étaient en fait très ordonnées, notamment grâce à la présence d'appariements stabilisant la structure de la partie en boucle. Par la suite, des boucles à trois ou quatre nucléotides contenant des appariements dans la boucle ont également été découvertes dans l'ADN.

Dans les boucles à trois ou quatre nucléotides des épingles à cheveux d'acides nucléiques, les appariements, s'ils existent, interviennent entre les bases extrémales de la boucle. Dans ces structures, les bases centrales ont tendance à s'empiler sur ce plateau mésapparié de la boucle. Elles forment une structure compacte qui minimise l'exposition des bases au solvant. L'hyperstabilité de ces structures s'explique notamment par la présence du mésappariement dans la boucle avec :

- la formation de liaisons hydrogène entre les bases mésappariées,

- les interactions d'empilement et hydrophobe entre :
  - le plateau de paire de bases mésappariées de la boucle et le dernier plateau de paire de bases de la tige,
  - les bases centrales de la boucle et le plateau de paire de bases mésappariées de la boucle.

La grande stabilité de certaines épingles à cheveux est indispensable à leurs fonctions biologiques. La compréhension du mode de structuration des mésappariements est donc essentielle pour comprendre la stabilité de ces motifs.

Dans les structures d'ARN étudiées, un appariement non Watson-Crick intervient entre la première et la dernière base de la boucle UUCG. La base G est en conformation SYN et une liaison hydrogène entre URA[O2] et GUA[HN2B] stabilise l'appariement. À ce jour, huit structures en tri-boucles d'ADN comportant un mésappariement dans la boucle sont accessibles dans les bases de données structurales ou reconstructibles à partir des données publiées. Ces huit structures [47, 55, 56, 62–65, 67] publiées par différents auteurs présentent des mésappariements purine-purine (A···A et 4 x G···A) ou purine-pyrimidine (2 x A···C ou G···C). La nature de la base centrale de la boucle peut changer, ainsi que la séquence de l'hélice sous-jacente. Pourtant, bien que de séquences et de mésappariements différents, toutes ces molécules présentent des structures globales très similaires.

I.4.3.4 Caractéristiques communes des géométries des appariements dans les tri-boucles d'ADN

L'observation approfondie des huit mésappariements met en évidence de nombreuses similitudes (*cf.* FIG. : I.12) dans la géométrie des bases de la boucle. On observe notamment que :

- Les bases en 5' de la boucle sont des purines (*i.e.* Adénine ou Guanine).
- Les conformations des bases sont toutes ANTI/ANTI (*i.e.* la valeur de l'angle χ de la liaison glycosidique est proche de la valeur observée dans les hélices), à l'exception de la tri-boucle 1P0U-GAC qui présente un mésappariement ANTI/SYN (*i.e.* la base en 3' de la boucle est tournée de près de 180° autour de la liaison glycosidique).

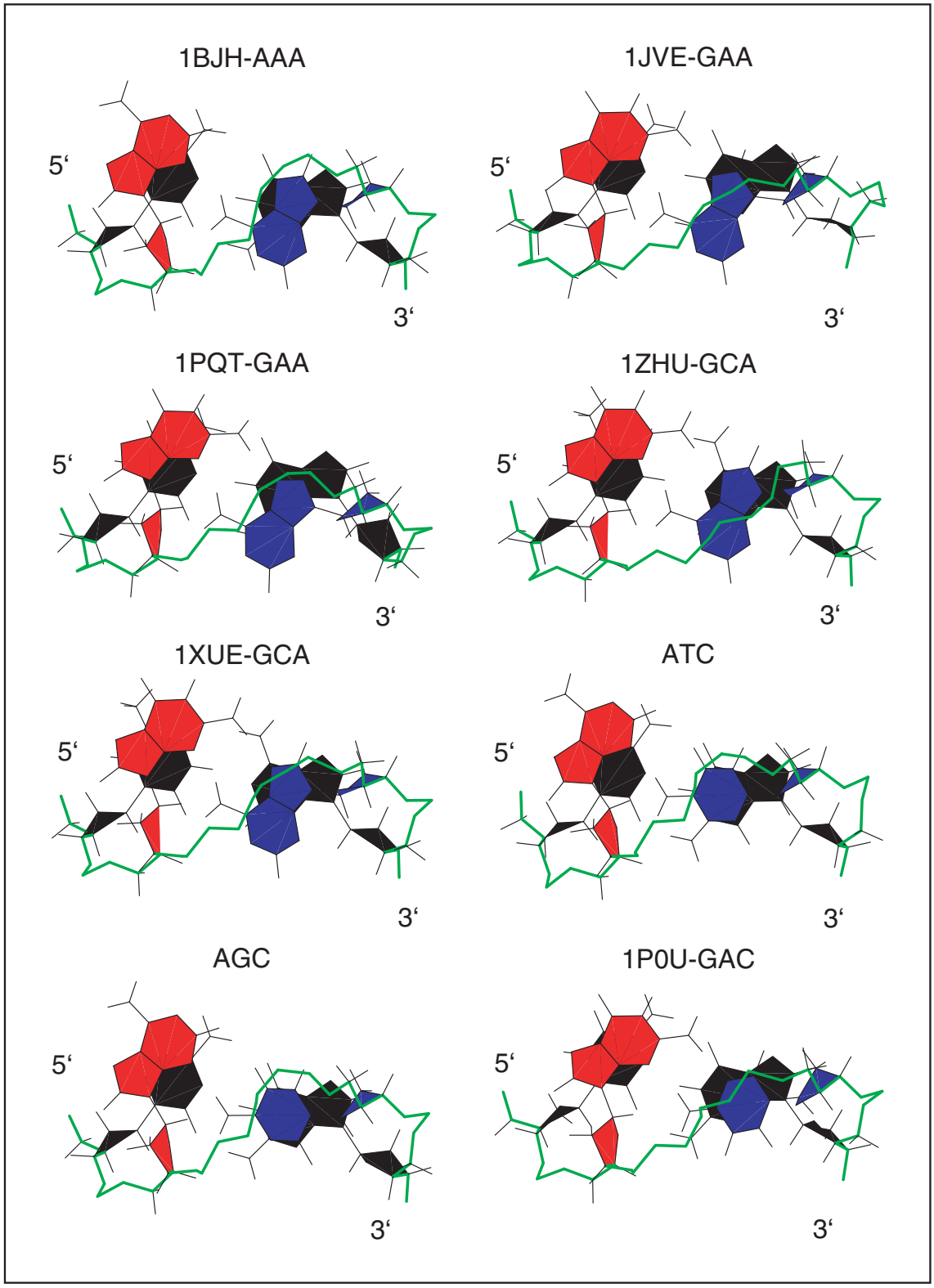


Fig. I.12 : ***Représentation des appariements des huit tri-boucles d'ADN*** *[47, 55, 56, 62–65, 67] : Vus de haut (selon l'axe de l'hélice) et de côté depuis le petit sillon. En rouge, la base en 5' de la boucle; en bleu : la base en 3' de la boucle; en noir, les bases du dernier plateau de paire de bases de la tige; en vert, le squelette de la chaîne sucre-phosphate.*

- Les bases en 5' des boucles ont tendance à se placer du côté du grand sillon, alors que les bases en 3' ont plutôt tendance à se placer au côté du petit sillon.
- Les côtés des bases qui présentent les groupements donneurs et accepteurs de protons sont identiques. Tous ces appariements sont de type "side by side sheared" ce qui correspond à :
  - **pour la base en 5' de la boucle, ce sont les atomes du côté de la liaison glycosidique** qui sont engagés dans les liaisons hydrogène : l'azote N3 de la purine est systématiquement accepteur de proton et si la base est une guanine, l'hydrogène HN2 est donneur de proton,
  - **pour la base en 3' de la boucle, ce sont les atomes opposés à la liaison glycosidique** qui sont engagés dans les liaisons hydrogène : si la base est une adénine l'HN6 est donneur de proton, si c'est une cytosine c'est l'HN4. Dans le cas des mésappariements G···A, le N7 de l'adénine est accepteur.

Conformations des bases	**Structures concernées**	**Liaisons hydogène**
A···A		
Anti-Anti	1BJH	N3···HN6A
G···A		
Anti-Anti	1XUE/1ZHU/1PQT/1JVE	HN2B···N7
		N3···HN6A
A···C		
Anti-Anti	ATC/AGC	N3···HN4A
G···C		
Anti-Syn	1P0U	N3···HN4B
		HN2B···N3

Tab. I.6 : ***Tableau des protons et des groupements accepteurs de proton engagés dans les liaisons hydrogène des appariements des tri-boucles d'ADN :*** *La première base de l'appariement correspond à la base extrémale en 5' de la boucle, et la seconde à la base en 3' de la boucle. Les groupements chimique engagés dans les liaisons hydrogène suivent le même ordre,* i.e. *le premier groupement est porté par la base en 5' et le second par la base en 3'.*

Les différents appariements caractérisés dans les structures publiées sont reportés dans la figure I.13 et le tableau I.6. Tous ces appariements sont de type "sheared" à l'exception près de l'appariement G···C. Cela veut dire que dans ces appariements, les groupements donneurs et accepteurs de protons situés sur les faces Watson-crick

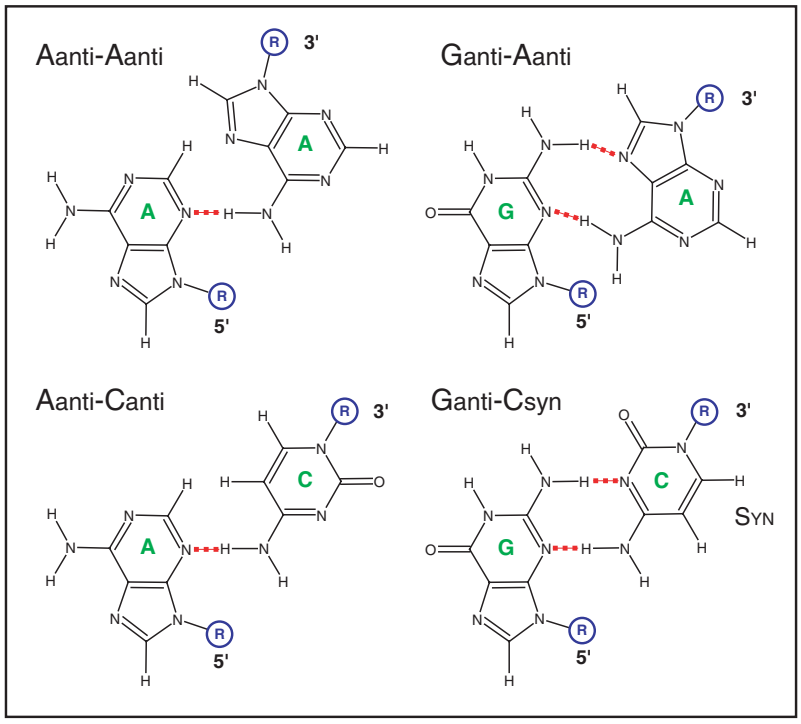


FIG. I.13 : *Schéma des trois familles d'appariements rencontrés dans les tri-boucles d'ADN sélectionnées :* *Représentation des appariements $A\cdots A$, $G\cdots A$, $A\cdots C$ et $G\cdots C$. Les trois premiers appariements présentent deux bases en conformation* ANTI. *Le dernier, l'appariement $G\cdots C$, présente une Guanine en conformation* ANTI *et une Cytosine en conformation* SYN.

ne sont pas impliqués dans la formation des liaisons hydrogène. À la place, ce sont les groupements "côté petit sillon" de la purine en 5' de la boucle qui interagissent avec les groupements "côté grand sillon" de la purine ou de la pyrimidine en 3' de la boucle. Dans le cas de l'appariement $G\cdots C$, c'est toujours le côté "petit sillon" de la guanine qui forme les liaisons hydrogène, par contre c'est le côté Watson-Crick de la cytosine qui est engagé. Pour réaliser ce dernier appariement, la base de la cytosine est tournée de près de 180° autour de la liaison glycosidique et se touve ainsi en conformation SYN, alors que toutes les autre bases sont en conformation ANTI.

Les appariements de type "sheared" sont fréquemment rencontrés dans les boucles internes. Dans les doubles hélices, seules des répétitions en tandem sont observées. La raison pourrait en être la forte déformation de la chaîne sucre-phosphate induite dans la double hélice par un seul mésappariement, alors qu'en tandem de suprenantes propriétés d'empilements rendent ces appariements quasiment aussi stables que les appariements Watson-Crick [79].

I.5 Conclusion

Les structures en épingle à cheveux sont des structures complexes. Pour les modéliser, les approches généralement utilisées font appel à des manipulations-déformations de la structure dérivée d'une description atomique du polymère dans un espace de coordonnées cartésiennes ou dans un espace de coordonnées internes (angles de torsion). Ces approches apparaissent parfois trop détaillées pour manipuler efficacement des macromolécules polymériques. Nous verrons, dans le chapitre suivant, qu'une alternative de modélisation est possible. Celle-ci, s'appuie sur le caractère de chaîne de polymère linéaire des macromolécules biologiques. Elle permet de distinguer les différentes échelles de structuration au moyen d'un approche hiérarchique et multi-échelle de modélisation.

Pour décrire ces structures, des paramètres quantitatifs dérivés de l'étude des structures en double hélice sont le plus souvent utilisées. Ces paramètres se révèlent peu adaptés à la description des conformations des parties en boucle. Nous verrons dans le prochain chapitre, que l'utilisation de notre nouvelle approche permet d'utiliser directement le nombre limité de degrés de liberté nécessaires à la déformation des molécules, comme paramètres quantitatifs de description.

Chapitre II

L'approche "Biopolymer Chain Elasticity" (BCE)

BCE est une nouvelle approche de modélisation moléculaire fondée sur une représentation continue des chaînes de biopolymères. Elle postule, dans une première approximation, que la trajectoire du squelette des polymères se comporte comme une barre mince, flexible, continue et inextensible [4, 80]. La trajectoire de cette chaîne est calculable à partir de la théorie de l'élasticité [4, 80]. Afin de passer du modèle continu de la trajectoire du polymère à une modélisation de tous les atomes de la molécule, BCE utilise une approche hiérarchique multi-échelle de modélisation qui se décompose en plusieurs étapes :

(1) Construction de la structure de départ BCE_{ori}

(1.1) génération de structures atomiques régulières en hélice de type "Arnott" [23–28] dans un espace cartésien :

(1.1.a) un double brin pour la tige de l'épingle à cheveux,

(1.1.b) un simple brin pour les nucléotides de la boucle,

(1.2) ajustement de fils (*i.e.* courbes) en hélice sur les trois trajectoires des chaînes sucre-phosphates des hélices générées en (1.1.a) et (1.1.b),

(1.3) calcul des coordonnées des deux points et des deux vecteurs tangents aux extrémités des fils en hélice de la partie tige. Ils définissent les paramètres d'encastrement ou conditions limites qui servent à calculer la trajectoire du fil de la boucle en (1.4),

(1.4) calcul de la trajectoire de la ligne élastique de la boucle, au moyen de la théorie de l'élasticité des barres minces,

(1.5) repliement de la structure simple-brin en hélice (1.1.b) sur la courbe élastique de la boucle calculée en (1.1.4) pour obtenir la conformation BCE_{ori},

(2) optimisation de la conformation de la boucle par rotation des bases autour du fil pour obtenir la conformation BCE_{opt},

(3) obtention de la conformation finale BCE_{min} par minimisation d'énergie.

Les étapes (2) et (3) sont constituées de sous-étapes qui seront détaillées plus loin.

Dans ce chapitre nous allons aborder les différents points de cette méthode.

- Après avoir présenté les principes de cette approche de modélisation hiérarchique et multi-échelle, nous rappellerons [4,80] comment sont calculées les différentes courbes utilisées pour modéliser la trajectoire de la chaîne d'acide nucléique des épingles à cheveux. Ensuite nous décrirons le passage du formalisme continu des courbes au formalisme discret de description atomique de la molécule. Nous finirons par la présentation des différents degrés de liberté utilisés pour optimiser le positionnement des atomes.

- Nous présenterons ensuite la nouvelle opération qui assure la continuité des repères de Frenet aux points de raboutage, d'une part entre les courbes de la tige et celle de la boucle, et d'autre part entre les ensembles de blocs que constituent les nucléotides. Mise en œuvre par la rotation des blocs d'atomes, cette étape permet de tenir compte et de répartir la torsion physique dont l'influence est assez faible pour le calcul de la trajectoire élastique.

- Nous introduirons un nouveau degré de liberté développé pour aller vers la prédiction des appariements dans les boucles.

- Nous finirons par décrire en quoi cet ensemble d'outils géométriques de manipulation des polymères permet de mettre en place les bases d'un champ de force mésoscopique.

II.1 BCE une approche de modélisation multi-échelle et hiérarchique pour passer du formalisme continu des courbes au modèle atomique

II.1.1 Une approche de modélisation multi-échelle

Les macromolécules biologiques sont synthétisées à partir de la répétition d'un nombre limité de motifs chimiques simples (acides nucléiques, acides aminés, sucres, ...) appelés monomères, sous-unités ou résidus. À l'échelle globale de la molécule, les acides nucléiques peuvent donc être considérés comme des chaînes linéaires déformées de polymères. Bien que fondamental, cet aspect "chaîne" n'est utilisé dans aucune approche de modélisation moléculaire de tous les atomes en biologie. Les approches de modélisation phénoménologiques dérivées de la physique généralement utilisées pour modéliser les molécules en biologie décrivent et manipulent les molécules directement à l'échelle atomique. D'un point de vue pratique, une conséquence directe de ces formalismes (*cf.* PART. : I.2.3) est le grand nombre de paramètres à manipuler pour déformer le polymère (3N degrés de liberté ou N est le nombre d'atomes). De ce point de vue, ces méthodes peuvent sembler très lourdes pour modéliser les structures globales des polymères biologiques car elles sont très détaillées.

Corrélativement, dans la plupart des logiciels de visualisation moléculaire, les acides nucléiques sont souvent représentés sous forme simplifiée. Le très grand nombre d'atomes nuisant à la compréhension de la structure globale de la molécule, celle-ci est représentée au moyen de fils ou de rubans. Le fil ou le ruban est calculé *a posteriori* par ajustement d'une courbe polynomiale (ou spline) sur la trajectoire existante de la chaîne sucre-phosphate. Ce mode de représentation, pertinent du fait de la nature polymérique des acides nucléiques, permet d'identifier rapidement les structures secondaires comme les hélices et les boucles dans la structure complexe de la molécule.

À l'image des logiciels de visualisation qui proposent des modes de représentation adaptés en fonction de l'échelle d'étude de la molécule (des rubans pour représenter les structures secondaires à l'échelle globale de la molécule ou des barres pour figurer

les liaisons atomiques à l'échelle locale des atomes), notre approche se propose de déformer la molécule au moyen de méthodes différentes selon les échelles : c'est une approche multi-échelle.

À l'échelle globale, la trajectoire du squelette de la molécule est représentée par des courbes calculées avec la théorie de l'élasticité. Ces courbes constituent un support théorique qui justifie les courbes polynômiales utilisées habituellement à des fins purement graphiques par les logiciels de visualisation. À l'échelle intermédiaire des résidus, la molécule est manipulée comme un ensemble de blocs rigides d'atomes (résidus ou parties de résidus) pouvant être déplacés autour de la courbe de la trajectoire globale ou autour de liaisons spécifiques (angle de torsion glycosidique). À l'échelle locale des atomes, la molécule est optimisée dans un champ de force (AMBER) qui minimisera son énergie totale par déplacement des atomes.

Cette approche permet de réduire le nombre de degrés de liberté utilisés pour déformer la structure car elle utilise des outils de déformation adaptés à chaque échelle de modélisation. La compréhension du mode de structuration et la description de ces molécules complexes sont ainsi simplifiées par l'emploi d'un petit nombre de paramètres quantitatifs ou de degrés de liberté.

II.1.2 Une approche de modélisation hiérarchique

Les approches de modélisation sont très généralement fondées sur l'utilisation de champs de force et définissent les conformations accessibles par des calculs d'énergie. Ces termes d'énergie conditionnent les déformations moléculaires et prennent en compte l'ensemble des interactions moléculaires suivant l'échelle moléculaire dans une expression générale (*cf.* Eq. : I.2.3.2). Ainsi, les interactions fortes telles que les forces de liaisons covalentes ou d'angles de liaisons qui sont importantes à l'échelle locale atomique, sont écrites dans la même fonction que les interactions faibles de type liaisons hydrogène ou que les interactions hydrophobes qui sont importantes à l'échelle globale de la molécule. Dans un tel système, la seule hiérarchie qui existe est celle des échelles d'ordre de grandeur. Les déformations globales de la molécule sont donc systématiquement soumises aux règles locales de géométrie des atomes qui prévalent du fait des ordres de grandeurs très élevés de leurs énergies. Dans cette logique d'une fonction d'énergie unique pour toutes les échelles, il est souvent nécessaire, pour déformer globalement la molécule, d'introduire des déformations

à l'échelle locale avec des contraintes élevées qui doivent ensuite être relachées progressivement.

Avec une approche multi-échelle et hiérarchique, ce genre de contraintes disparaît. Dans un premier temps la trajectoire globale de la molécule est mise en place sans tenir compte des contraintes locales de géométrie des liaisons atomiques. L'utilisation de courbes solutions de la théorie de l'élasticité des barres minces, ou courbes BCE, nous affranchit en effet du lourd formalisme atomique très détaillé et permet de nous focaliser sur les déformations à l'échelle globale. Une fois la forme globale de la molécule établie, les positions relatives des bases sont mises en place et optimisées pour produire une structure proche d'un minimum énergétique. Puis lors d'une dernière étape, seules les positions atomiques locales des atomes sont corrigées. Le fait de soumettre dans un champ de force une structure ainsi proche d'un minimum énergétique global a pour effet de conserver au cours de la minimisation finale la structure globale de la molécule précédemment mise en place.

L'ordre hiérarchique des étapes de modélisation (globale $\rightarrow$ intermédiaire $\rightarrow$ locale) permet donc de mettre en place d'abord les formes des structures conditionnées par des interactions de plus faibles énergies (interactions faibles à l'échelle de l'énergie d'une liaison covalente ou de l'énergie d'ionisation d'un atome), pour aller progressivement vers l'affinement des interactions de plus fortes énergies (jusqu'aux liaisons covalentes). Cette approche est possible parce que le repliement de la structure moléculaire sur le fil élastique déforme faiblement les longueurs et les angles des liaisons atomiques entre les blocs attachés au fil. L'utilisation d'une hiérarchie d'échelle moléculaire explicite, à la place d'une hiérarchie d'échelle énergétique implicite, facilite ainsi la mise en place de la structure globale.

II.1.3 Des courbes, des blocs et des atomes décrivent la molécule aux différentes échelles moléculaires

L'approche de modélisation BCE est hiérarchique et multi-échelle. Elle est hiérarchique car elle procède par étapes successives en mettant en place d'abord la structure globale et seulement ensuite les structures locales, pour finir par les positions atomiques. Elle est multi-échelle car au cours de chaque étape de modélisation elle manipule la molécule à une échelle différente.

À l'échelle globale, la structure, appelée BCE_{ori}, est manipulée comme un fil élastique. La trajectoire de la chaîne sucre-phosphate est définie au moyen d'une ou

Étape **Échelle**	**(1)** **Globale**	**(2)** **Intermédiaire**	**(3)** **Atomique**
Nom de la structure			
	BCE_{ori}	BCE_{opt}	BCE_{min}
Objets manipulés			
	Déformation de courbes de courbes mathématiques	Déplacement de blocs rigides	Déplacement des coordonnées atomiques
Nombre de d.d.l.			
	5 par courbe 0 dans le repère de modélisation de BCE	3 n (n nucléotides)	3 N (N atomes)

Tab. II.1 : ***Hiérarchie des étapes de modélisation de l'approche BCE.***

de plusieurs courbes mathématiques (3 dans le cas des motifs en épingles à cheveux). Le nombre de degrés de liberté (d.d.l.) est alors limité aux paramètres nécessaires pour définir les courbes :

- pour les courbes en hélice : d'une façon générale il faut connaître le rayon, le pas et la phase de la courbe. En fait, comme nous utilisons systématiquement, dans le cadre de BCE, des hélices canoniques dont les paramètres d'hélicité ne varient pas et ne sont pas modifiés, le nombre de degrés de liberté associés à la modélisation des courbes de la tige des épingles à cheveux d'acides nucléiques se réduit à zéro : le nombre de d.d.l. = 0.
- pour les courbes élastiques : il faut connaître la longueur de la chaîne, la distance entre les deux extrémités de la courbe et trois angles pour orienter les tangentes aux extrémités de la courbe : nombre de d.d.l. = 5. Comme nous le verrons, ces paramètres ne sont pas des variables. Ce sont des constantes définies par la géométrie de l'hélice de la tige et de la séquence de la boucle dans le cas des épingles à chevuex d'acides nucléiques. Comme ce sont des constantes on peut considérer que : le nombre de d.d.l. = 0.

À l'échelle intermédiaire, la structure, appelée BCE_{opt}, est manipulée comme un ensemble de blocs rigides déplaçables par rotation. Trois degrés de liberté sont utilisés pour faire tourner les blocs associés à chaque nucléotide : la rotation d'angle Ω du nucléoside autour de la tangente à la courbe à laquelle il est rattaché, la rotation d'angle Θ de redressement d'empilement qui sera utilisée pour empiler les

paires de bases de la boucle sur le dernier plateau de paires de bases de la tige en hélice et la rotation d'angle χ de la base autour de liaison glycosidique, (= 3n ddl, n étant le nombre de nucléotides de la boucle).

À l'échelle atomique, la structure, appelée BCE_{min}, est manipulée comme un ensemble d'atomes placés dans un champ de forces. Afin de corriger les déformations des longueurs et angles de liaisons introduites par les manipulations précédentes la molécule est soumise à une courte minimisation. Le nombre de degrés de liberté utilisés est de 3N-6 d.d.l., N étant le nombre d'atomes de la molécule, dont chaque coordonnée (x,y,z) peut être modifiée pour rétablir une conformation atomique correcte.

La notion de hiérarchie se retrouve dans le nombre de degrés de liberté introduit à chaque étape. Plus l'échelle devient locale, plus la description de la molécule devient précise et plus le nombre de degrés de liberté augmente avec la complexité de l'objet manipulé. L'idée dans BCE est donc de commencer par décrire et manipuler la molécule "simplement" à l'échelle du biopolymère, puis d'aller pas à pas vers plus de complexité jusqu'à une description de "tous les atomes de la molécule".

II.2 Modélisation de la structure globale de la molécule : Trajectoire de la chaîne sucre-phosphate et courbes mathématiques associées

Comme nous l'avons abordé dans le paragraphe précédent, la manipulation de la molécule à l'échelle globale s'opère par la manipulation-déformation de courbes BCE. Nous avons précisé qu'il est possible d'utiliser plusieurs courbes différentes pour une même molécule sans en préciser la raison. En fait, le découpage d'un motif en plusieurs courbes est nécessaire pour tenir compte de la complexité de l'ensemble sans avoir à définir des courbes trop complexes. Dans le chapitre introductif, nous avons décrit le motif en épingle à cheveux comme composé d'une partie en double hélice dénommée la tige et d'une partie liant les deux extrémités de la tige appelée la boucle. Afin de modéliser la structure d'un tel motif dans BCE, il est pratique d'utiliser ce découpage qui donne lieu à deux courbes en hélice pour chaque brin de la tige et une courbe différente pour la partie en boucle.

Pour la partie en double hélice le choix du type de courbes à utiliser est évident. Pour la trajectoire de la boucle, la courbe associée doit être définie à partir des directions calculées aux extrémités des courbes associées aux brins de la tige au moyen de la théorie de l'élasticité. Après une brève description de la théorie de l'élasticité et de la façon dont elle est utilisée dans l'approche BCE, nous spécifierons le repère cartésien orthonormé de notre espace de modélisation (*i.e.* repère du laboratoire qui sert notamment à définir l'axe et le cylindre de la double hélice) puis la façon dont les trajectoires de la tige et de la boucle sont déterminées, en suivant l'ordre des étapes de modélisation.

II.2.1 Elasticité, flexion des barres minces et calcul de la trajectoire de la boucle des épingles à cheveux

II.2.1.1 Théorie de l'élasticité et calcul de la trajectoire d'une barre mince

La théorie de l'élasticité constitue une branche classique des mathématiques et de la physique. Elle permet de calculer les trajectoires d'une barre mince soumise à des contraintes [4,80]. Dans le formalisme classique, les contraintes sont exprimées en terme de couples et de forces appliquées aux extrémités d'une barre (*cf.* FIG. : II.1.a). La trajectoire calculée par cette théorie est une trajectoire qui minimise les énergies des contraintes physiques à l'intérieur de la barre soumise à déformation.

D'un point de vue géométrique, il est possible de ne pas considérer les contraintes de couples et de forces, mais seulement les paramètres d'encastrement imposés aux extrémités de la barre. Ces paramètres se réduisent aux deux points d'encastrement de la barre et aux deux tangentes qui sont imposées aux extrémités. Il est intéressant de remarquer que pour un même jeu de paramètres d'encastrement, la trajectoire prédite par la théorie de l'élasticité est indépendante de la nature du matériau si les barres sont de longueurs identiques. Ainsi, que l'on considère un fil d'acier ou un fil de nylon, bien que de rigidités intrinsèques différentes, les trajectoires données par l'élasticité sont identiques si les paramètres d'encastrement sont identiques. Les couples et les forces à appliquer aux extrémités ne sont évidemment pas les mêmes (tordre un fil d'acier demande plus d'énergie que tordre un fil de nylon de même section), mais la trajectoire décrite par les deux fils est identique.

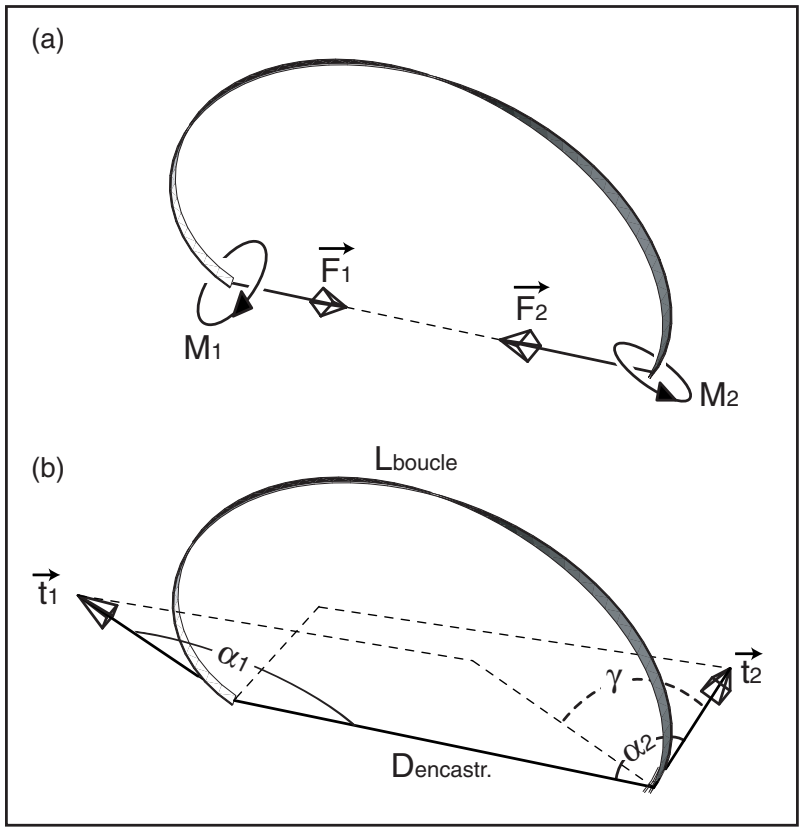


Fig. II.1 : ***Trajectoire élastique et paramètres d'encastrement :*** *Représentation des paramètres nécessaires au calcul de la trajectoire élastique. (a) le formalisme physique avec les couples et les forces appliquées aux extrémités. (b) le formalisme géométrique développé pour BCE avec les points et les tangentes imposées aux extrémités.*

II.2.1.2 Application à la modélisation de la trajectoire de la boucle des épingles à cheveux d'acides nucléiques

Dans le cadre de l'approche BCE, nous postulons en première approximation que la chaîne sucre-phosphate se comporte comme une barre mince et que la trajectoire d'énergie minimale des acides nucléiques peut être approximée par la trajectoire d'énergie minimale déterminée par la théorie de l'élasticité. La trajectoire étant indépendante de la nature du matériau, il est possible, au moyen de cette théorie, de prédire la trajectoire du squelette si l'on connaît seulement les paramètres d'encastrement de la boucle sur la double hélice et la longueur de la boucle. Pour les acides nucléiques, les valeurs des forces et des couples qui s'appliquent aux extrémités de la barre associée à la boucle sont inconnus à l'heure actuelle. C. Pakleza et J. Cognet ont appliqué et développé à cet effet un nouveau formalisme [4, 80] et le logiciel S-Mol pour déterminer la trajectoire d'un fil élastique à partir des

seules contraintes géométriques d'encastrement (*cf.* FIG. : II.1.b). Ainsi, avec ce logiciel , il est aisé de déterminer la trajectoire tridimensionnelle élastique d'un fil en connaissant les points et les tangentes de ses extrémités.

La définition de ces paramètres d'encastrement est primordiale car elle est la première étape dans la hiérarchie de notre approche et elle conditionne la validité de l'ensemble des étapes suivantes de modélisation. Le choix des deux points, et des deux tangentes des extrémités du fil associé à la partie en boucle découle de l'observation des structures déjà résolues par RMN ou cristallographie [4, 80, 81]. L'étude de ces structures montre qu'il n'y pas de rupture brusque de trajectoire de la chaîne entre les parties en double hélice et les parties simple-brin en boucle. Les conditions d'encastrement sont apparemment données par les positions et les directions des extrémités des brins en hélice constituant la tige. Pour les définir, nous générons une hélice moléculaire régulière d'ADN ou d'ARN correspondant à la tige. Sur les trajectoires en hélices des chaînes sucre-phosphates de la tige nous ajustons des courbes mathématiques helicoïdales. Les points et les tangentes aux extrémités de ces courbes définissent les paramètres d'encastrement de la boucle.

La construction de la partie en tige est donc une étape fondamentale. Elle met en place les fondations de la construction de l'édifice moléculaire de la boucle de l'épingle à cheveux. Les paramètres d'encastrement définis à ses extrémités sont utilisés pour calculer la trajectoire élastique associée à la chaîne sucre-phosphate de la boucle. Les conformations accessibles aux bases de la boucle dépendent donc directement de la géométrie de l'hélice de la tige.

II.2.2 Modélisation de la trajectoire de la Tige

Même si nous ne prenons pas en compte les modifications fines de la conformation des hélices de la tige nous avons vu que sa géométrie est déterminante pour les étapes suivantes de notre protocole. Cette géométrie est caractérisée par la conformation de l'hélice sur laquelle les courbes de l'approche BCE sont ajustées.

II.2.2.1 Caractéristiques des doubles hélices moléculaires utilisées pour modéliser la tige

Nous choisissons de modéliser les hélices des tiges des épingles à cheveux à partir d'hélices régulières en conformation standard d'ADN-B ou d'ARN-A [23]. Ces

hélices sont caractérisées par des angles de torsion qui sont invariants le long de la chaîne sucre-phosphate quelque soit le type et la position du nucléotide considéré (*cf.* TAB. : II.2). Des jeux caractéristiques différents donnent les formes A et B d'acides nucléiques.

	α	β	γ	δ	ϵ	ζ	χ
ARN-A	-83,5°	175,8°	68,4°	78,2°	-166,0°	-69,4°	-164,9°
ADN-B	-71,4°	179,4°	59,5°	129,4°	-133,0°	-101,6°	-102,7°

TAB. II.2 : ***Tableau des angles de torsion des hélices canoniques de forme A et B utilisées dans BCE.***

Afin de faciliter la comparaison et de rationaliser la modélisation de toutes nos structures en épingles à cheveux, nous choisissons de toujours positionner les hélices de la même façon dans le repère cartésien du laboratoire (*cf.* FIG. : II.2.a) :

- l'axe (Oz) du repère du laboratoire est confondu avec l'axe de la double hélice de la tige,
- le repère du laboratoire est choisi suivant la convention de Cambridge [1, 69] pour le dernier plateau de paire de bases de la tige en double hélice (*i.e.* le plateau jouxtant la partie en boucle). Le repère est orthonormé et direct. Il est défini par un axe (Oz) précédemment défini et orienté suivant la direction 5'→3' du brin I, un axe (Oy) colinéaire au vecteur liant les C1' des deux bases du plateau et orienté depuis le brin II vers le brin I, et un axe (Ox) orthogonal aux deux précédents. (Ox) pointe vers le grand sillon et correspond à l'axe de pseudo-dyadicité des doubles hélices d'ADN ou d'ARN [1].

Les hélices utilisées pour modéliser les épingles à cheveux sont de l'un ou de l'autre type choisi (conformation B pour l'ADN ou A pour l'ARN).

II.2.2.2 Définition des courbes mathématiques associées aux squelettes moléculaires de la tige et définition des paramètres d'encastrement

À partir des hélices moléculaires ainsi générées, il est possible de calculer, pour chaque brin, une courbe hélicoïdale mathématique qui passe au mieux (au plus près)

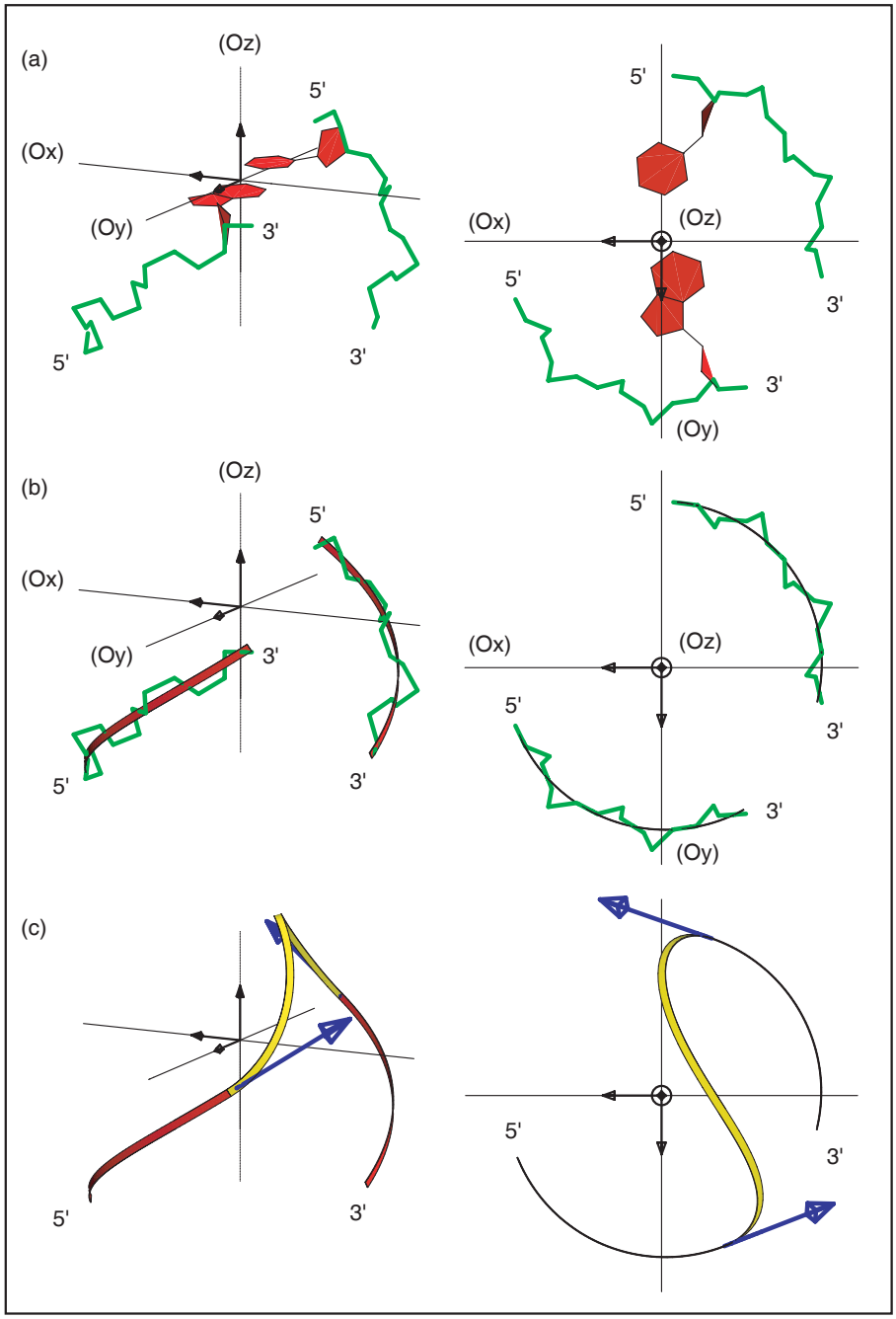


Fig. II.2 : ***Les courbes de l'approche BCE :*** *(a) Le repère du laboratoire est défini en utilisant les conventions de Cambridge [1] pour le dernier plateau de paire de bases de la double-hélice de la tige, comme indiqué dans le texte. En vert, la chaîne sucre-phosphate de la double hélice de la tige; en rouge, le dernier plateau de paire de bases de la tige; en noir, les axes du repère. (b) en rouge, la trajectoire des courbes en hélice associés à la tige. (c) en bleu, les points et les tangentes qui servent de paramètres d'encastrement pour le calcul de la courbe BCE; en jaune, la trajectoire de la courbe BCE associée à la boucle.*

des atomes de la chaîne sucre-phosphate. Elle permet de définir une trajectoire globale associée au squelette de la molécule. Ces courbes sont définies de façon paramétrique comme des fonctions de l'abscisse curviligne. Les coordonnées $\overrightarrow{r}$ des points de la courbe hélicoidale sont données par les équations suivantes :

$$\overrightarrow{r}(s) = \begin{cases} x(s) &= R\,cos(s/c + \varphi) \\ y(s) &= R\,sin(s/c + \varphi) \\ z(s) &= (p/2\,\pi)\,(s/c) \end{cases}$$

où

R, est le rayon de l'hélice,

φ, est l'angle de déphasage de hélice,

p, est le pas de l'hélice (*i.e.* la hauteur d'un tour d'hélice), et

c $= \sqrt{R^2 + (p/2\pi)^2}$ tel que 2πc est la longueur d'un tour d'hélice.

L'équation de cette courbe est définie par un ajustement numérique des paramètres R, φ et p qui minimise la distance à la courbe des atomes C5', C4', C3', O3' et P du squelette nucléotidique du brin de l'hélice considéré. Pour l'ADN et l'ARN les valeurs calculées de R sont respectivement de 8,35 Å et 9,35 Å et pour p elles sont de 33.74 Å/*tour* et 30.85Å/*tour*. Les courbes associées aux squelettes de la tige (*cf.* FIG. : II.2.b) servent à définir les paramètres permettant de calculer la trajectoire BCE de la boucle.

II.2.3 Modélisation de la trajectoire de la boucle avec l'élasticité

Les paramètres d'encastrement choisis sont les points et les tangentes calculées aux extrémités des fils associés aux deux hélices de la tige (*cf.* PART. : II.2.2). La longueur de la boucle est celle d'une courbe en hélice ajustée sur un simple brin en hélice de séquence identique à la séquence de la boucle. Nous postulons que la longueur de la chaîne sucre-phosphate d'un simple brin en boucle de séquence donnée est proche de celle d'une séquence identique structurée en double hélice.

Nous utilisons la théorie de l'élasticité des barres minces, flexibles et inextensibles pour modéliser la trajectoire des simple brins d'acides nucléiques. Il n'est pas très surprenant que cette approche fonctionne puisque le squelette des acides nucléiques est à 90% de son extension maximale dans une hélice d'ADN [83, 84].

La solution donnée par l'élasticité (*cf.* FIG. : II.2.c) donne une courbe qui adopte naturellement une forme en "S" similaire aux trajectoires des chaînes sucre-phosphates des structures en épingles à cheveux résolues par d'autres approches. Comme observé dans les structures expérimentales la trajectoire de la boucle :

- s'inscrit naturellement, sans rupture, dans la continuité de la trajectoire de la partie en hélice car les tangentes aux extrémités des fils en hélice et celles du fil élastique de la boucle sont identiques, et
- semble présenter une zone de plus forte déformation pouvant s'apparenter à la notion de "sharp-turn".

II.3 Modélisation des blocs d'atomes à partir de la trajectoire

Une fois les courbes définies, la modélisation de la structure complète de l'épingle à cheveux requiert le calcul des positions de tous les atomes. Dans le cadre de l'approche BCE, il faut donc passer du formalisme de description de la molécule de l'échelle globale des courbes à la description de la molécule à l'échelle locale des atomes. Pour ce faire, nous utilisons d'abord une opération de géométrie différentielle qui nous permet de replier un simple brin en hélice d'acide nucléique sur la courbe élastique associée à la boucle. Cette étape utilise la trajectoire de la chaîne sucre-phosphate calculée au moyen de la théorie de l'élasticité. Ensuite, au moyen de différentes opérations de rotations, nous ajustons l'orientation des nucléotides en fonction de divers critères géométriques ou énergétiques.

II.3.1 Repliement d'une hélice sur la courbe de la boucle calculée par élasticité

La structure atomique de l'épingle à cheveux doit comprendre les positions de tous les atomes de la tige comme de la boucle. La structure de la tige est générée dès la

première étape de l'approche BCE avec tous les atomes, il ne reste donc plus qu'a exprimer les positions des atomes de la boucle. Ces atomes sont positionnés autour de la courbe BCE qui donne la trajectoire globale de la chaîne sucre-phosphate de la boucle. Pour positionner ces atomes, nous procédons en plusieurs étapes. Un simple brin en hélice régulière est généré avec la séquence de la boucle. Une courbe en hélice est ajustée sur le squelette de ce simple brin. Les atomes sont regroupés en blocs. Les coordonnées cartésiennes des atomes des différents blocs sont exprimées dans différents repères. Ces repères correspondent à des trièdres de Frenet associés à certaines abscisses curvilignes de la courbe en hélice. De même longueur que la courbe en hélice, des repères homologues (*i.e.* aux mêmes abscisses curvilignes) sont calculés sur la courbe de la boucle donnée par la solution de l'élasticité. Ces repères permettent de ré-exprimer les coordonnées des atomes dans l'espace cartésien, non plus autour du fil en hélice, mais autour du fil élastique de la boucle. Cette opération de changement de repère à l'aide de géométrie différentielle permet de replier un simple brin en hélice sur une courbe BCE.

En résumé, l'idée est de "faire suivre" l'ensemble des atomes et blocs d'atomes attachés initialement au fil en hélice du squelette en hélice du simple brin d'ADN ou d'ARN par des déplacements appropriés pour garder la même position sur le fil calculé par la théorie de l'élasticité.

II.3.1.1 Définition des différents blocs d'atomes

L'opération de géométrie différentielle qui permet de replier un simple brin en hélice sur une courbe BCE provoque des déformations des positions relatives des atomes. De façon logique et afin de limiter les déformations des liaisons atomiques, des blocs rigides d'atomes sont définis (*cf.* FIG. : II.3.a et Tab. : II.3).

À l'image d'un collier de perles, la molécule est donc décrite à une échelle intermédiaire comme un ensemble séquentiel de blocs s'enfilant le long du squelette de la molécule (*cf.* FIG. : II.3.b.1). Ces blocs servent à déplacer les atomes par groupes (*cf.* FIG. : II.3.b.2). A l'intérieur d'un bloc les positions relatives des atomes sont figées. Ainsi, la géométrie des liaisons atomiques entre atomes d'un même bloc est toujours conservée. Pour déformer une molécule, seules les positions relatives des blocs peuvent être modifiées. En conséquence, seules les liaisons entre atomes de groupes différents sont susceptibles d'être affectées lors des déformations, notamment lors des étapes d'optimisation du placement des atomes.

Atomes Pivots	Atomes du bloc	
O5'	*O5' ;*	
C5'	*C5' ;*	H5A' ; H5B'
C4'	*C4' ;*	H4'
C3'	*C3' ;*	TOUS LES AUTRES
		C1' ; C2' ; O1' ; O2' ;
		H1' ; H1A' ; H1B' ; H2A' ; H2B' ; H3' ; HO2' ;
		N1 ; N2 ; C2 ; O2 ; N3 ; N4 ; C4 ; O4 ; C5 ;
		N6 ; C6 ; O6 ; N7 ; C7 ; C8 ; N9 ;
		H1 ; H2 ; HN2A ; HN2B ; H3 ;
		HN3A ; HN3B ; HN4A ; HN4B ; H5 ; H6 ;
		HN6A ; HN6B ; H7A ; H7B ; H7C ; H8
O3'	*O3' ;*	
P	*P ;*	OPA ; OPB

TAB. II.3 : ***Définition des atomes pivots et des blocs d'atomes associés.***

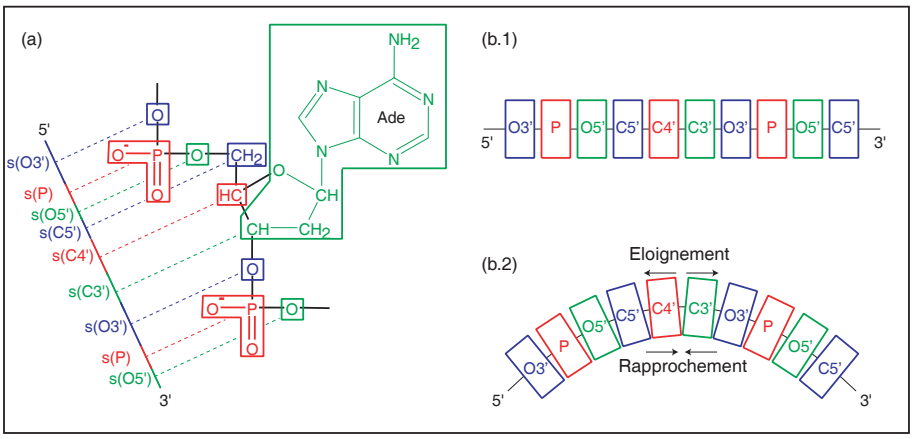


FIG. II.3 : ***Regroupement des atomes en blocs rigides :*** *Chaque nucléotide est décomposé en six blocs correspondant aux atomes pivots P (en rouge), O5' (en vert), C5' (en bleu), C4' (en violet), C3' (en marron) et O3' (en noir). La projection de l'atome pivot du bloc sur la courbe donne l'abscisse curviligne* s(Atome_Pivot) *du trièdre de Frenet qui défini l'origine du repère local du bloc.*

II.3.1.2 Expression des coordonnées atomiques dans les repères locaux des courbes et repliement de l'hélice sur la courbe élastique de la boucle

Pour replier le simple brin en hélice sur la courbe élastique de la boucle, nous faisons appel à une opération élémentaire de géométrie différentielle qui s'articule en plusieurs étapes :

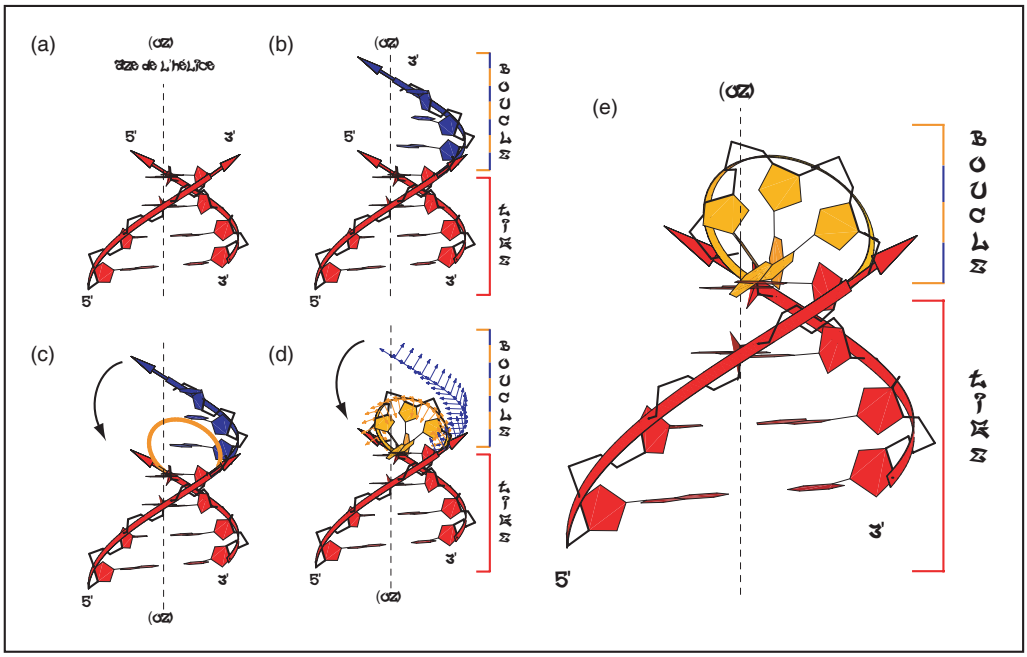


FIG. II.4 : ***Repliement de l'hélice sur la courbe élastique de la boucle :*** *Premières étapes de la modélisation d'une tri-boucle d'ADN. En rouge sont représentés les atomes et les courbes de la tige en hélice. En bleu, la courbe en hélice, les atomes et les trièdres de Frenet associés au simple brin de la boucle avant repliement. En orange, la courbe élastique, les atomes et les trièdres de Frenet associés au simple brin de la boucle après le repliement. (a) structure en hélice de la tige, (b) boucle de l'épingle à cheveux sous sa forme hélicoïdale de départ, (c) trajectoire de la boucle calculée par la théorie de l'élasticité appliquée à la flexion des barres minces, (d) repliement de l'hélice de la boucle sur la courbe élastique au moyen du formalisme des trièdres de Frenet, (e) structure BCE_{ori} obtenue par le repliement élastique d'une hélice d'acides nucléiques.*

- Dans un premier temps, des trièdres de Frenet sont calculés en certains points de la courbe en hélice (*cf.* FIG : II.3.a & II.4.(b & c & d)). Les axes de ces trièdres sont donnés par la normale, la binormale et la tangente en ces points. L'origine des repères est donnée par la projection des atomes pivots de chaque bloc sur la courbe en l'hélice. Ils forment un ensemble de repères orthonormés directs appelés repères locaux.

- Dans un deuxième temps les coordonnées des atomes de chaque bloc sont exprimées dans ces repères locaux.

- Des repères homologues sont calculés sur la courbe BCE (*cf.* FIG : II.4.c & d). Cette courbe est de même longueur que la courbe en hélice, et les repères locaux sont calculés aux mêmes abscisses curvilignes que celles des repères de la courbe en hélice. Les atomes exprimés dans les repères locaux du fil

en hélice peuvent être exprimés identiquement dans les repères locaux du fil élastique. La position des repères locaux du fil élastique étant connue dans le repère cartésien absolu, les coordonnées des atomes peuvent être exprimées à nouveau dans ce repère.

	Res.	P-O5'	O5'-C5'	C5'-C4'	C4'-C3'	C3'-O3'	O3'-P
Différence	1	0,091Å	0,001Å	-0,085Å	-0,061Å	-0,007Å	0,053Å
en longueur de	2	0,115Å	0,002Å	-0,099Å	-0,070Å	-0,006Å	0,054Å
valence	3	0,086Å	0,001Å	-0,098Å	-0,062Å	-0,008Å	-0,091Å
		O3'-P-O5'	P-O5'-C5'	O5'-C5'-C4'	C5'-C4'-C3'	C4'-C3'-O3'	C3'-O3'-P
Différences	1	-5,4°	-4,1°	6,3°	4,1°	4,5°	-10,1°
en angle de	2	-5,7°	-13,0°	13,4°	4,9°	6,3°	-7,7°
valence	3	-4,8°	-5,7°	4,3°	3,7°	-0,5°	3,1°
		α	β	γ	δ	ϵ	ζ
Différences	1	-6,3°	-9,2°	-12,1°	4,2°	-21,3°	-5,0°
en angle de	2	-15,2°	3,5°	-23,9°	3,7°	-21,6°	-1,9°
torsion	3	-7,1°	-9,9°	-10,3°	13,5°	-26,8°	14,0°

Tab. II.4 : ***Modification des angles de torsion, angles et longueurs de liaisons de la chaîne sucre-phosphate lors du repliement élastique :*** *Les différences sont calculées entre une hélice canonique (*cf. *Part. : II.2.2) de séquence $d(A_3)$ et la même hélice repliée sur la courbe BCE (BCE_{ori}). Pour les trois nucléotides que compte la séquence repliée, les différences de longueur de liaisons sont exprimées en Angstroms et les différences d'angle de valence et de torsion sont en degrés.*

Cette opération de géométrie différentielle équivaut à replier les atomes et blocs d'une hélice régulière sur la courbe donnée par l'élasticité. Cette opération modifie les positions relatives de deux blocs adjacents (*cf.* Fig. : II.3.b.1) entrainant des modifications des géométries atomiques des liaisons entre les blocs. Ces modifications restent faibles. Les variations induites des longueurs de valence sont comprises entre 0,001Å et 0,115Å, les angles de valence s'écartent de leur valeurs canoniques de 0,5° à 13,4° en valeur absolue et les angles de torsion varient de -26,9° à+14,0° pour les cas extrèmes (*cf.* Tab. : II.4). Au total, les valeurs des angles de torsion, des angles et longueurs de valence des atomes de la chaîne sucre-phosphate sont globalement conservées relativement à celles de l'hélice de départ. Par ailleurs, les liaisons n'appartenant pas au squelette de la molécule restent identiques car elles font partie des blocs rigides.

II.3.1.3 L'optimisation de la structure repliée sur la courbe BCE est nécessaire

La structure obtenue (*cf.* FIG : II.4.e) par le repliement de l'hélice sur la courbe BCE de la boucle définit la structure de départ BCE_{ori}. La trajectoire de la chaîne sucre-phosphate de cette boucle a donc été globalement optimisée au moyen de l'élasticité, par contre, les orientations des bases de la boucle ne sont pas correctes. En effet, à l'issue de l'opération de repliement, les bases pointent le long de l'axe de la double hélice et télescopent le dernier plateau de paires de bases de la tige. Afin de résoudre ces encombrements stériques et d'optimiser la structure en fonction des données structurales dérivées de l'expérience, nous devons modifier l'orientation des nucléotides. Pour cela, BCE utilise différents degrés de liberté à l'échelle des blocs. Ces degrés de liberté sont définis comme des rotations de blocs rigides. Pour préserver la trajectoire de la chaîne sucre-phosphate, les différents axes de rotation passent sytématiquement par la courbe BCE.

II.3.2 Rotation des nucléosides autour de la tangente au fil élastique et rotation des bases autour de la liaison glycosidique

Après l'étape de calcul de la trajectoire optimale de la chaîne sucre-phosphate qui conduit à la conformation BCE_{ori}, vient l'étape d'optimisation de l'orientation des nucléotides qui conduit à la conformation BCE_{opt}. Pour ce faire, deux degrés de liberté, Ω et χ, ont initialement été définis lors de l'élaboration de l'approche BCE [4, 80, 81]. Ils sont utilisés pour modéliser les structures en tri-boucles comportant trois thymines dans la boucle et dans l'étude présentée au chapitre III de ce manuscript.

II.3.2.1 Rotation d'angle Ω

Le premier degré de liberté est la rotation d'angle Ω. Ce degré de liberté est propre à l'approche BCE. Il découle du formalisme d'expression des coordonnées par blocs dans les repères locaux du fil. Chaque bloc est associé à un repère constitué des trois axes du trièdre de Frenet : la normale, la binormale et la tangente à la courbe. Il est simple de définir des rotations autour de chacun de ces axes. La rotation que nous utilisons est la rotation du bloc (*i.e.* des coordonnées des atomes du bloc)

autour de l'axe tangent au fil (*cf.* FIG. : II.5). Le sens de la rotation définit des angles positifs lorsque les atomes de la boucle sont tournés dans un mouvement qui forme une hélice droite en progressant le long du fil dans le sens 5'→3' de la chaîne sucre-phosphate. Cela correspond à un mouvement depuis leur position d'origine, BCE_{ori}, vers le grand sillon de la tige.

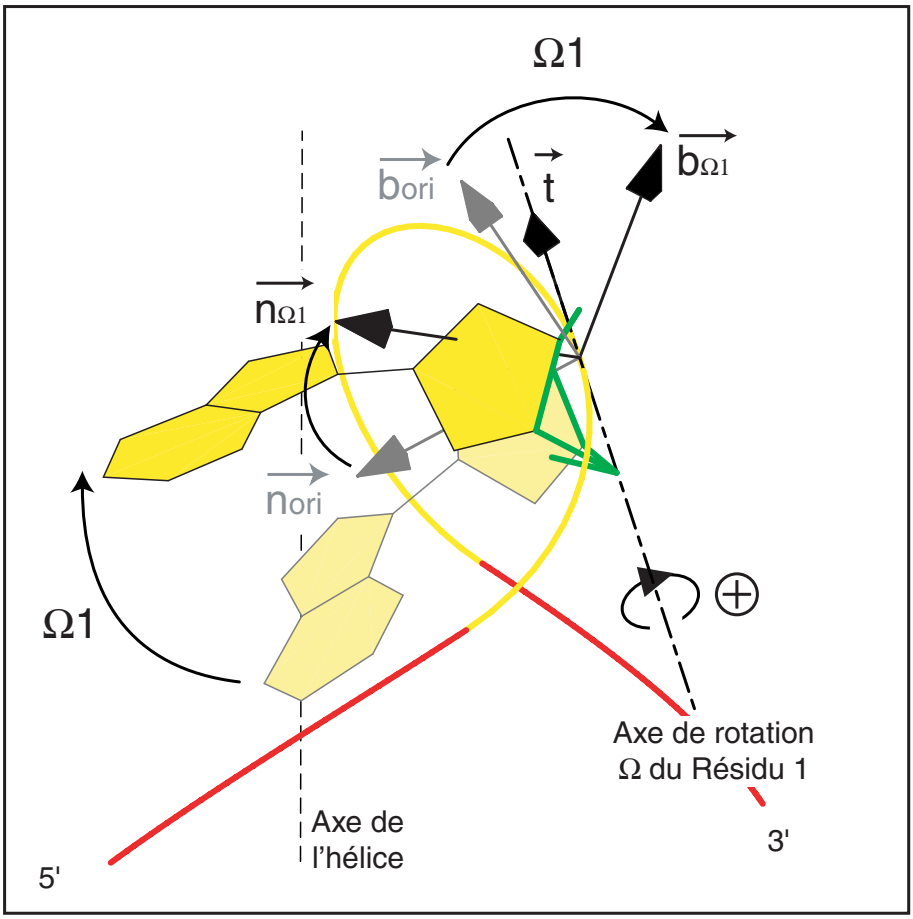


FIG. II.5 : ***Rotation de la base autour de la tangente au fil élastique :*** *En jaune et rouge la trajectoire des courbes associées aux parties en tige et en boucle de l'épingle à cheveux. En jaune clair, représentation du premier nucléotide d'une tri-boucle d'ADN. En jaune foncé, représentation du même nucléotide tourné par la rotation d'angle Ω. En pointillé, représentation de l'axe de rotation autour de la tangente au fil pour le premier nucléotide de la tri-boucle. En noir et gris, représentation des axes du trièdre de Frenet associé au premier nucléotide de la tri-boucle, respectivement avant et après rotation d'angle Ω.* $\overrightarrow{n_{ori}}$ *et* $\overrightarrow{n_{\Omega 1}}$ *les vecteurs normaux du trièdre avant et après rotation.* $\overrightarrow{b_{ori}}$ *et* $\overrightarrow{b_{\Omega 1}}$ *les vecteurs bi-normaux du trièdre avant et après rotation.* $\overrightarrow{t}$ *le vecteur tangent inchangé par la rotation d'angle Ω.*

La rotation conjointe du même angle Ω des blocs C3' et C4' permet de tourner les atomes du sucre et de la base autour du fil élastique. Comme les déplacements des

atomes sont d'autant plus importants qu'ils sont éloignés de l'axe de rotation (*i.e.* du fil) ce degré de liberté permet d'imprimer, moyennant des angles relativement faibles, de forts déplacements absolus des atomes des bases éloignés de l'axe de rotation, tout en modifiant peu la position des atomes de la chaîne sucre-phosphate. En effet, l'axe étant tangent au fil, les positions des atomes proches du fil sont peu déplacés. Le fil joue ainsi le rôle d'une armature qui permet de modifier la conformation des molécules tout en préservant la forme globale de sa trajectoire.

II.3.2.2 Rotation d'angle χ

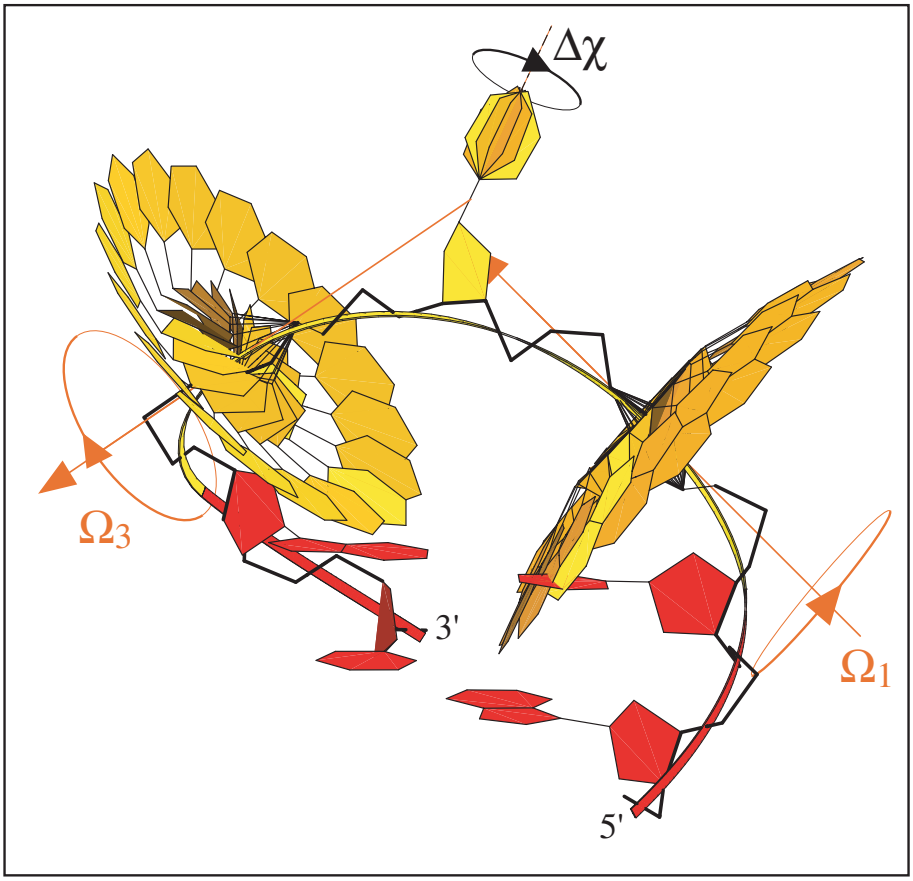


Fig. II.6 : Ω *et* χ *les deux premiers degrés de liberté de l'approche BCE :* *Représentation de l'espace conformationnel accessible aux bases d'une tri-boucle d'ADN avec les degrés de liberté Ω et χ de BCE. La base centrale de la boucle décrit l'espace parcouru en faisant varier le paramètre χ et les deux bases extrémales en 5' et en 3' de la boucle décrivent l'espace parcouru en faisant varier le paramètre Ω. Les bases de la tige sont représentées en rouge, et les bases de la boucle en jaune. Les axes de la rotation d'angle Ω sont représentés en orange.*

Le second degré de liberté est la rotation d'angle χ. Il s'agit d'un degré de liberté

naturel de rotation de la base autour de la liaison glycosidique (C1'-N9 pour les purines et C1'-N1 pour les pyrimidines). Cette rotation modifie la position relative des atomes de la base par rapport aux atomes du sucre (*cf.* FIG. : II.6). Elle n'implique aucun atome de la chaîne sucre-phosphate et ne modifie donc pas la trajectoire globale de la molécule. Les positions des atomes du sucre, les longueurs et les angles de valences des liaisons ne sont pas modifiées par cette rotation.

II.3.2.3 La structure BCE_{opt} et l'optimisation de l'orientation des nucléotides de la boucle

Les rotations d'angle Ω et χ sont utilisées pour optimiser l'orientation des nucléotides et obtenir la conformation BCE_{opt}. Deux angles suffisent donc pour définir la conformation d'un nucléotide. Ainsi, pour une tétra-boucle d'ADN ou d'ARN faut-il définir quatre couples de valeurs (Ω_i, χ_i), i étant la position du nucléotide dans la séquence de la boucle.

Les valeurs des angles Ω et χ associés à chaque nucléotide sont déterminées à partir des données structurales dont le modélisateur souhaite tenir compte. Ainsi, l'étude des tri-boucles d'ADN -TTT- [4,80] qui est le premier exemple d'application de l'approche BCE, a-t-elle été menée d'après des données de distance RMN. Pour l'étude des boucles d'ADN et d'ARN comportant des appariements dans la boucle [81] présentée au chapitre III, nous avons utilisé les positions cartésiennes des fichiers de structure PDB. Enfin, pour l'étude de la formation des appariements dans les tri-boucles d'ADN, nous utiliserons des critères de recherche de conformations optimisant la formation de liaisons hydrogène comme nous le verrons au chapitre IV.

II.4 Torsion des blocs d'atomes autour de la trajectoire

II.4.1 Conservation qualitative des angles de torsion de la structure de départ

Une des particularités des angles de torsion de la chaîne sucre-phosphate observés dans les boucles des épingles à cheveux est leur similarité avec les angles de torsion

des parties en hélice. Pour tenir compte de cette donnée structurale, l'idée de l'approche BCE est de partir d'une structure en hélice qui présente des angles de torsion canoniques et de la déformer pour converger vers la conformation en boucle. Afin de satisfaire les données structurales d'angles de torsion, il faut que toutes les opérations de déformation de la structure de départ préservent au mieux ces angles de torsion. De ce point de vue, l'étape de repliement de l'hélice sur la trajectoire BCE respecte cette règle. L'opération de géométrie différentielle utilisée permet en effet de déformer de façon importante l'hélice de départ en répartissant les déformations sur l'ensemble du brin : localement les déformations sont faibles et donc les valeurs des angles de torsion sont peu affectées (*cf.* PART. : II.3.1). Si au cours de cette étape les angles de torsions obtenus sont corrects, leur conservation soulève des difficultés lors :

- du raboutage des portions de la molécule modélisées à partir de courbes différentes (hélice - courbe élastique - hélice), et
- des rotations indépendantes d'angle Ω_i de nucléotide i autour du fil élastique.

Dans ces deux cas la difficulté provient du fait que la molécule est manipulée par morceaux contrairement à l'étape de repliement sur la courbe élastique où le brin est considéré comme une unité déformable soumis à une déformation globale. Chaque nucléotide est tourné individuellement par des rotations locales réalisées dans des référentiels différents (les trièdres de Frenet associés à chaque blocs). Dans le cas du raboutage des portions de molécules de courbes différentes, les molécules que l'on assemble (deux molécules pour la tige et une pour la boucle) pour former l'épingle à cheveux ont été manipulées dans des référentiels différents et sont assemblées comme des objets séparés dans un référentiel commun du repère du laboratoire. Dans les deux cas, les opérations impliquent des solutions de continuité entre les différentes parties de la même molécule. Ces solutions de continuité doivent être formulées. À l'échelle de la chaîne sucre-phosphate, elles évitent des déformations des angles de torsion. La cause en est le traitement inhomogène de parties contigües de la molécule qui provoque une rupture de la continuité des déformations à l'échelle des angles de torsion.

Afin de résoudre ces déformations inacceptables car non observées des angles de torsion de l'épingle à cheveux, nous dotons notre approche d'un outil permettant de manipuler la molécule à cette échelle. Cet outil permet de répartir continuement

sur tous les atomes (*i.e.* tous les angles de torsion) l'effet d'une déformation due aux manipulations par morceaux de la molécule. Il résout à l'échelle des angles de torsion les deux problèmes décrits précédemment : l'effet du raboutage des différentes portions de la molécule et l'effet de la rotation en Ω des blocs C3' et C4' qui correspondent à la partie centrale du nucléotide.

II.4.2 Le "raboutage" en torsion des différentes portions de la structure en épingle à cheveux

La manipulation des molécules à l'échelle globale est réalisée au moyen de courbes BCE. À cette échelle, la continuité de la molécule est assurée par la continuité des tangentes aux extrémités des différentes courbes. Pour passer de l'échelle globale à l'échelle de description atomique, nous utilisons les trièdres de Frenet, moyennant une opération de géométrie différentielle pour exprimer la position des atomes relativement à la trajectoire du fil de la boucle. Cette condition de continuité des tangentes des différentes courbes, donnée par les paramètres d'encastrement de la barre n'est pas suffisante. En effet, sans conditions supplémentaires, la continuité de la molécule à l'échelle atomique serait perdue au niveau du raboutage. Il faut donc assurer non seulement la continuité des tangentes mais aussi la continuité de l'ensemble des systèmes de repères locaux dans lesquels sont exprimées les coordonnées des atomes. Il faut donc assurer la continuité des normales et des binormales des repères en plus de la continuité des tangentes.

L'expression des trièdres de Frenet aux points de raboutage entre les courbes en hélice et la courbe donnée par la théorie de l'élasticité montre que les normales et les binormales ne sont pas continues (*cf.* FIG. : II.7). Les normales associées à la courbe en hélice sont en effet tournées d'un angle $\Omega_{5'}$ et $\Omega_{3'}$ par rapport aux normales associées au fil de la boucle. Ces angles sont des constantes qui ne dépendent pas de la composition de la séquence. Ils ne dépendent que de la géométrie des courbes, et donc de la nature de la chaîne (*i.e.* ADN ou ARN) et de la longueur du fil de la boucle (*i.e.* de la longueur de la séquence de la boucle). Les valeurs de ces angles sont donc constantes pour toutes les tri-boucles d'ADN, et de la même façon elles sont constantes pour toutes les tétra-boucles d'ADN ou d'ARN (*cf.* TAB. : II.5).

Afin de rétablir la continuité des repères locaux aux points de raboutage des courbes nous devons introduire une torsion physique le long du fil élastique, *i.e.* nous

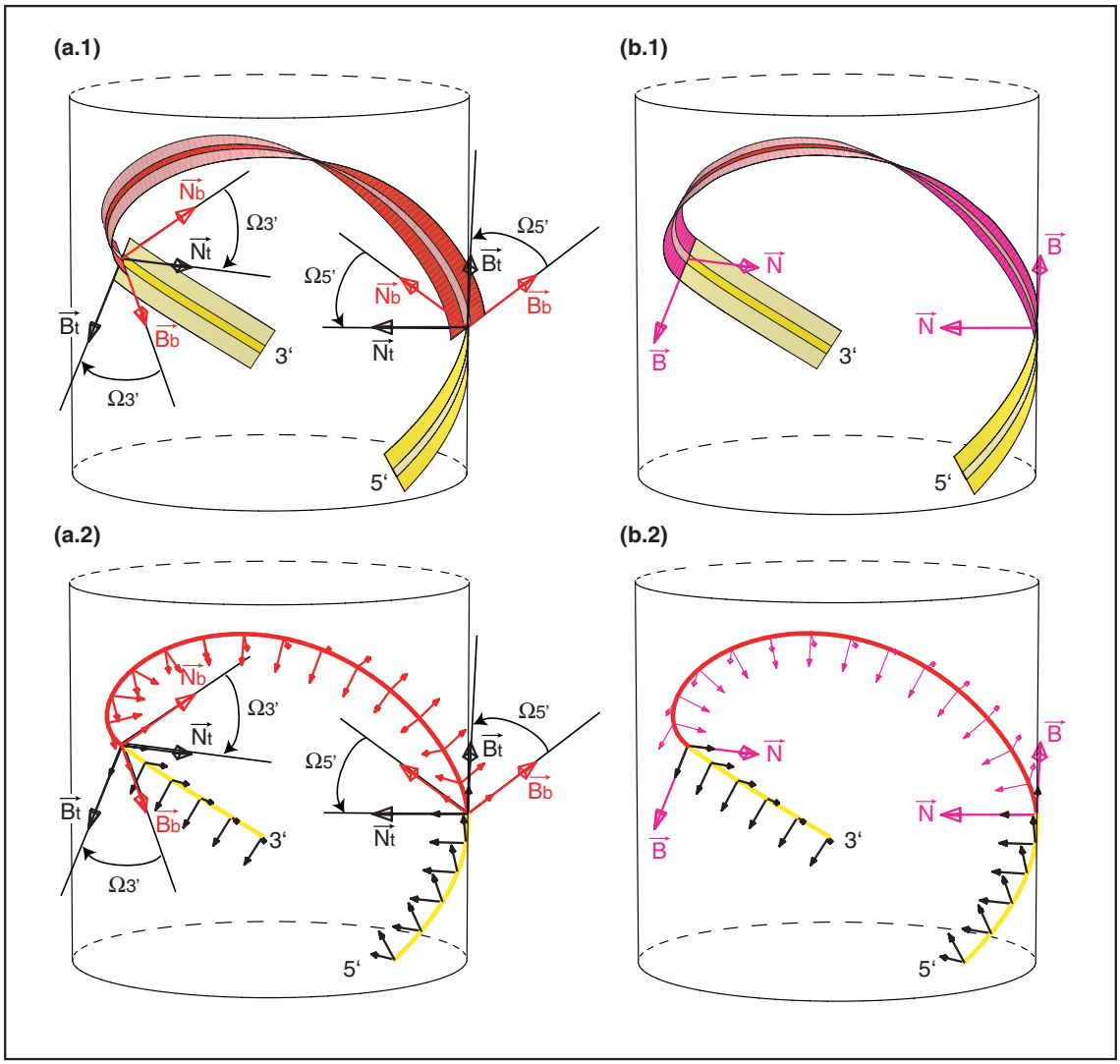


FIG. II.7 : *Continuité des normales et binormales des trièdres des repères locaux :* *Représentation des normales et des binormales (a) avant rotation et (b) après rotations des trièdres de Frenet. (1) En haut les binormales sont représentées continuement sous forme d'un ruban. (2) En bas les normales et binormales sont représentées de façon discrète sous forme de vecteurs. Aux points de raboutage des courbes, les normales et binormales des trièdres de Frenet de chaque courbe sont représentées par de grands vecteurs. Après rotation, les trièdres tournés de la partie boucle sont confondus avec les trièdres non-tournés de l'extrémité de la courbe en hélice.*

devons tourner tous les repères de Frenet de la partie en boucle (les parties en hélice sont fixes) de $\Omega_{5'}$ du côté 5' et de $\Omega_{5'}+(\Omega_{3'}-\Omega_{5'})=\Omega_{3'}$ du côté 3'. Pour assurer la continuité des normales et binormales entre les deux extrémités, il faut donc répartir continuement le long de la courbe l'effet de ces rotations aux extrémités, soit la rotation totale d'angle $(\Omega_{3'}- \Omega_{5'})$. Au final, les repères locaux de la partie en boucle sont tournés d'un angle égal à la rotation du repère à une extrémité $(\Omega_{5'})$ et

Nature de la chaîne	**Longueur de la boucle**	$\Omega_{5'}$	$\Omega_{3'}$	$\Delta = \Omega_{5'} - \Omega_{3'}$
ADN	Tri-boucles	43.7 °	-43,8 °	87,5 °
	Tétra-boucles	65,6 °	-65,0 °	130,6 °
ARN	Tétra-boucles	17,7 °	-101,5 °	119,2 °

TAB. II.5 : ***Valeurs des rotations autour de la tangente à appliquer pour assurer la continuité des rubans aux points de raboutages en 5' et 3'.***

d'une contribution proportionnelle à la fraction de l'abscisse curviligne qui sépare l'origine du repère local des deux extrémités de la courbe et à la rotation totale $\Omega_{3'}$-$\Omega_{5'}$ (*cf.* FIG. : II.4.2.1).

$$\begin{aligned} \Omega_{bloc} &= \Omega_{5'} + \frac{(s_{bloc} - s_{5'})}{(s_{3'} - s_{5'})} (\Omega_{3'} - \Omega_{5'}) \\ \Leftrightarrow \Omega_{bloc} &= \Omega_{5'} + \frac{(s_{bloc} - s_{5'})}{L_{boucle}} (\Omega_{3'} - \Omega_{5'}) \end{aligned} \qquad \text{(II.4.2.1)}$$

où

L_{boucle} est la longueur de la boucle.

En fait, cette opération est nécessaire car nous avons négligé en première approximation l'effet de la torsion physique de la barre mince associée à la partie en boucle puisque les couples et les forces nécessaires pour déformer une chaîne d'acides nucléiques nous sont inconnus. En fait, l'effet de la torsion physique est double :

- Elle modifie la torsion géométrique de la barre, donnant une trajectoire tridimensionnelle légèrement différente. Cependant, C. PAKLEZA [4] a montré que, pour un fil de longueur constante et des paramètres d'encastrement identiques, l'introduction de torsion physique modifie peu la forme donnée par la théorie de l'élasticité des barres minces sans torsion. La torsion géométrique de la barre peut donc être correctement approximée en négligeant la torsion physique.
- D'autre part la torsion physique se traduit par des rotations locales infinitésimales des portions solides de la barre autour de l'axe de la barre. Transposé aux chaînes d'acides nucléiques, cela se traduit par la rotation des blocs autour de la tangente à la trajectoire prédite par l'élasticité. Pour

tenir compte de cet effet, nous utilisons les trièdres de Frenet à la fois pour repérer les atomes et comme des repères associés physiquement à la barre. Les rotations des blocs sont donc introduites en tournant les repères locaux autour de l'axe du fil élastique afin de tenir compte de l'effet de la torsion physique sur la rotation des blocs.

Cette dernière opération découle de l'hypothèse d'isotropie de la barre mince manipulée avec la théorie de l'élasticité qui implique que toute torsion physique de la barre se répartit linéairement en fonction de l'abscisse curviligne entre les deux extrémités [85, 86]. Cette propriété qui s'applique au cas du raboutage des courbes élastiques, doit donc s'appliquer de façon équivalente entre deux points quelconques de la courbe. Pour cette raison, nous appliquons le même procédé entre deux nucléotides de la boucle lorsqu'ils sont tournés par la rotation d'angle Ω.

II.4.3 Dans le cas des rotations Ω_i des nucléotides de la boucle

Lors de l'optimisation de l'orientation des bases de la boucle, les blocs C3' et C4' de chaque nucléotide N_i sont tournés conjointement d'un angle Ω_i autour de la tangente au fil élastique. Pour de fortes valeurs de Ω_i cette opération peut modifier fortement les angles de torsions impliquant les atomes C3' et C4' si les atomes des blocs contigus ne sont pas également tournés. Pour préserver les angles de torsion il faut donc tourner les blocs voisins dans le même sens que les blocs C3' et C4'. Il n'est cependant pas possible de tourner les blocs O3' et C5' de la même valeur Ω_i que les blocs du sucre. En effet, cela reporterait le problème aux blocs voisins des blocs O3' et C3' et propagerait l'opération jusqu'aux blocs des sucres des nucléotides N_{i-1} et N_{i+1} rendant l'opération insoluble.

Comme nous l'avons expliqué plus haut, afin d'amortir l'impact des rotations des sucres sur les angles de torsion voisins nous devons introduire une opération de répartition linéaire des rotations des blocs entre chaque sucre, similaire à celle du raboutage des différentes courbes. L'opération consiste à tourner chaque bloc compris entre deux sucres d'un angle Ω_{bloc}, de façon à répartir l'effet des rotations des sucres sur les angles de torsions, à tous les angles de torsion compris entre les deux sucres. L'angle Ω_{bloc} dont est tourné chaque bloc intermédiaire est calculé en fonction de l'abscisse curviligne du bloc et en fonction des valeurs d'angle Ω_i et Ω_{i+1} dont sont tournés chaque sucre encadrant le bloc (*cf.* Eq. : II.4.3.2).

$$\Omega_{bloc} = \Omega_i + \frac{(s_{bloc} - s_i)}{(s_{i+1} - s_i)} (\Omega_{i+1} - \Omega_i) \qquad \text{(II.4.3.2)}$$

L'opération ainsi définie permet de répartir l'impact de la rotation des sucres et des bases sur tous les angles de torsion compris entre les deux sucres. Ici, les blocs associés aux atomes pivots C3' et C4' des sucres sont considérés comme des points de référence de contrainte extérieur qui impose une torsion. On remarque que la répartition de la torsion préserve automatiquement au mieux les angles de torsion de la structure de départ. Les rotations des blocs et des atomes dans l'espace

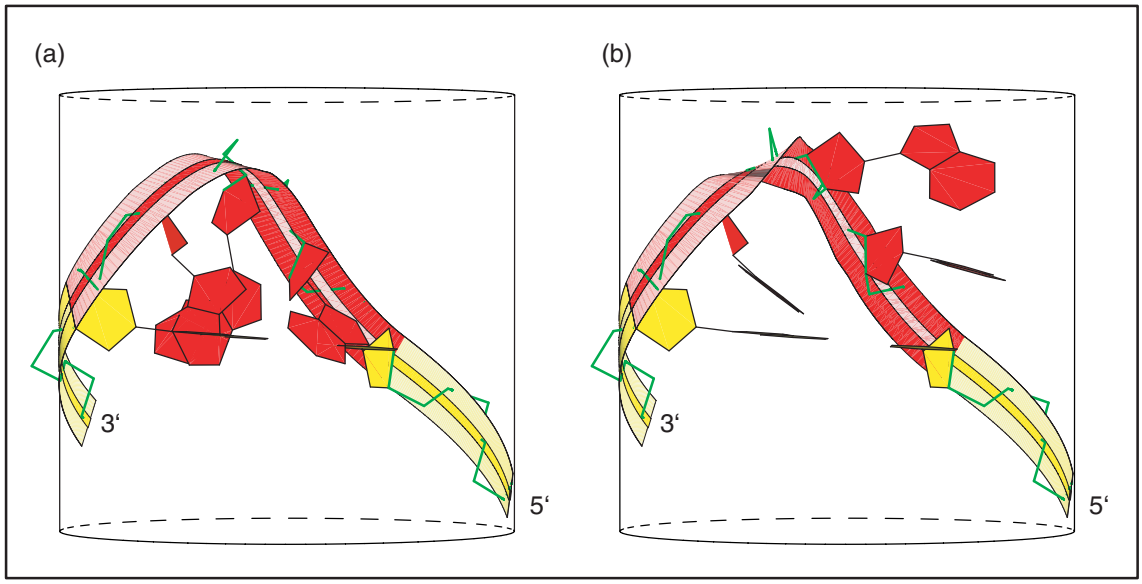


Fig. II.8 : ***Répartition de la torsion entre les sucres tournés autour du fil :*** *Représentation d'une tri-boucle d'ADN de séquence -AAA-. En jaune est représenté le ruban et le dernier plateau de paire de bases de la tige. En rouge, le ruban et les 3 nucléotides de la boucle. (a) La structure BCE_{ori} a subi le raboutage des rubans, mais les rotations des nucléotides ne sont pas encore optimisées. (b) La structure BCE_{opt} où les rotations des nucléotides i sont optimisées par rotation d'angle Ω_i. Cette rotation des sucres entraine, par répartition, la rotation de tous les blocs intermédiaires, linéairement, suivant l'abscisse curviligne du bloc.*

peuvent, selon le point de vue adopté, être considérées comme des rotations des positions atomiques dans les repères locaux où sont exprimées leur coordonnées, ou être considérées comme des rotations de ces repères dans le repère absolu, avec des coordonnées atomiques fixes. Prenons le deuxième point de vue pour illustrer graphiquement l'opération. Dans la figure II.8, les trois bases d'une tri-boucle sont tournées, et l'opération de répartition des rotations des blocs est appliquée à l'ensemble des repères locaux concernés. Dans cette figure, les binormales des repères

définis le long de la courbe associée à la partie en boucle forme le ruban. La rotation de chaque repère par l'angle Ω_{bloc} défini précédemment se traduit par une rotation du ruban. Par rapport aux représentations traditionnelles en ruban des molécules cette représentation ajoute de l'information, puisqu'en plus de la trajectoire, le ruban traduit également les déformations qu'a subies l'hélice de départ pour donner la conformation obtenue. La déformation peut être appréciée visuellement en observant la torsion du ruban autour de la ligne centrale.

II.5 Flexion des nucléotides à partir de la position donnée par la trajectoire

II.5.1 Rotation de redressement d'empilement : un nouveau degré de liberté de l'approche BCE

Les deux degrés de liberté Ω et χ permettent de déplacer les sucres et les bases de la boucle. Il est donc possible dans certains domaines de valeurs de lever les encombrements stériques entre les atomes des bases de la tige et ceux des bases de la boucle. Cependant, les deux angles de rotations ne permettent pas de décrire toutes les rotations possibles. Pour ce faire il faut ajouter une troisième rotation pour compléter les rotations d'un corps solide. En effet, comme le montre la figure II.6 qui décrit l'espace conformationnel accessible en fonction du degré de liberté Ω, pour des valeurs faibles de Ω les bases extrémales de la boucle (*i.e.* les bases en 5' et 3' de la séquence de la boucle) sont toujours en conflit avec les bases adjacentes de la tige. Dans ces conformations, la base N(Ω) de la boucle reste fortement inclinée. Pour lever ces contraintes stériques en tenant compte de l'épaisseur des bases et de leur placement au dessus du dernier plateau de paire de bases de la tige, nous introduisons ici un nouveau degré de liberté de rotation appelé "rotation de redressement d'empilement des bases" d'angle Θ_{empil}, qui peut être aussi comprise plus généralement comme une rotation de flexion du fil pour placer la base dans un plan particulier.

Leonhard Euler a montré comment décrire le positionnement d'un corps solide au moyen de trois rotations. Ici, il s'agit de définir des rotations analogues autour d'axes si possible perpendiculaires, en tenant compte de la géométrie imposée des nucléotides. L'axe de rotation doit donc être choisi perpendiculairement aux deux

directions importantes pour cette optimisation : la direction idéale finale que doit adopter le plan de la base empilée, et la direction de la liaison glycosidique autour de laquelle s'opère déjà la rotation d'angle χ. Quelque soit la conformation de départ d'une base (*i.e.* quelque soit l'angle Ω dont il a été tourné), il s'agit de redresser la base $B(\Omega)$ de façon que :

- le plan moyen de la base redressée soit globalement parallèle au plan moyen du plateau de paire de bases adjacent,
- la distance séparant le plan de la base redressée du plan moyen du plateau de paire de bases adjacent soit proche de 3,34 Å.

A priori il existe dans l'espace cartésien une infinité d'axes de rotation qui redressent un nucléotide $N_i(\Omega)$ pour placer sa base parallèlement à une direction donnée. Cependant il est possible de définir un axe optimal pour le faire. Nous allons poser des conditions pour définir un axe unique de redressement d'empilement.

Comme pour la rotation d'angle Ω et la déformation élastique, la rotation de redressement d'empilement doit respecter la règle de préservation des angles de torsion de la structure de départ. Et comme pour les degrés de liberté précédemment définis dans BCE, le redressement d'empilement est effectué en déplaçant des blocs rigides d'atomes. Ici le bloc rigide tourné est constitué de l'ensemble des atomes des cycles du sucre et de la base. Les déformations de bord à craindre portent donc sur les longueurs (C5'-C4' et C3'-O3') et les angles de valence ($O5'\widehat{C5'}C4'$, $C5'\widehat{C4'}C3'$, $C4'\widehat{C3'}O3'$ et $C3'\widehat{O3'}P$) et les angles de torsion (β, γ, δ, ε et ξ) liant le bloc rigide au reste de la chaîne sucre-phosphate. Afin de minimiser l'impact de cette rotation sur la structure initiale de la molécule nous posons donc deux conditions pour définir l'axe de rotation :

Condition 1 : l'axe de rotation est choisi pour minimiser l'impact de la rotation sur la déformation des longueurs et angles de valence et des angles de torsion aux jonctions entre le bloc rigide tourné et la chaîne sucre-phosphate fixe.

Condition 2 : l'axe de rotation est choisi pour optimiser le redressement de la liaison glycosidique, et donc pour que l'angle de redressement d'empilement Θ_{empil} soit le plus faible possible pour un nucléoside $N_i(\Omega)$ donné.

Dans les sections suivantes, nous allons détailler la mise en place de ce nouveau degré de liberté. Nous expliquerons dans un premier temps comment est déterminée la direction idéale d'empilement. Nous définirons ensuite l'axe de rotation en fonction de la position initiale de la base à redresser et de la position finale souhaitée. Nous expliquerons comment calculer la valeur de l'angle Θ_{empil} pour assurer une distance proche de 3,34 Å entre les atomes de la base tournée et ceux du dernier plateau de paire de bases de la tige sur laquelle la base doit s'empiler. Finalement, nous montrerons que la combinaison des degrés de liberté de rotation Ω, χ et Θ_{empil} permet de parcourir un espace conformationnel où les bases extrémales de la boucle sont empilées sur le dernier plateau de la tige.

II.5.2 La direction idéale de redressement du plateau

Pour mettre en place l'opération de redressement d'empilement nous avons besoin de connaître la direction finale que doit adopter le plan moyen de la base de la boucle à empiler. Cette direction idéale est définie comme la direction qu'adopterait le plateau de paire de bases suivant si on prolongeait l'hélice régulière de la tige d'un plateau de paire de bases supplémentaire. Elle peut s'exprimer au moyen d'un vecteur unique, le vecteur normal à ce plan. La direction moyenne d'un plateau de paire de bases de la tige est définie comme la direction du plan moyen P_{Tige} associé aux atomes de ce plateau que l'on peut exprimer avec le vecteur $\overrightarrow{Vn_T}$ normal au plan P_{Tige}.

Dans les hélices régulières en forme B, qui forment les tiges des épingles à cheveux d'ADN dans BCE, les deux plateaux consécutifs T_{i-1} et T_i font un angle constant proche de zéro mais non nul. Cet angle peut être mesuré de façon locale et relative par l'angle η_{plat} (*cf.* PART. : I.4.3.1) qui a été défini comme l'angle entre les vecteurs normaux des deux plateaux consécutifs $\overrightarrow{Vn_{T_{i-1}}}$ et $\overrightarrow{Vn_{T_i}}$ (*cf.* FIG. : II.10). Comme l'hélice est régulière, cette inclinaison relative constante entre les deux plateaux peut également être évaluée de façon globale et absolue dans le repère du laboratoire par les angles $\eta_{hel.}$ et η_{twist} :

- L'angle η_{hel} est l'angle défini entre le vecteur normal $\overrightarrow{Vn_{T_i}}$ de chaque plateau et l'axe de l'hélice (Oz). Cet angle η_{hel} est une constante dans les hélices régulières.

- L'angle η_{twist} est défini comme l'angle formé entre les projections des deux vecteurs $\overrightarrow{Vn_{T_{i-1}}}$ et $\overrightarrow{Vn_{T_i}}$ sur le plan (Oxy). L'angle η_{twist} est une constante qui lie n'importe quels vecteurs $\overrightarrow{Vn_{T_{i-1}}}$ et $\overrightarrow{Vn_{T_i}}$ normaux à deux plans consécutifs dans les hélices régulières.

Ces deux angles permettent de calculer l'inclinaison globale η_{hel} et l'orientation η_{twist} de n'importe quel plateau T_{i+1} en connaissant les plateaux T_{i-1} et T_i dans une hélice régulière, par l'opération suivante :

Soit :

$$\overrightarrow{Vn_{T_{i-1}}} = \begin{pmatrix} Xn_{T_{i-1}} \\ Yn_{T_{i-1}} \\ Zn_{T_{i-1}} \end{pmatrix} \; et \; \overrightarrow{Vn_{T_i}} = \begin{pmatrix} Xn_{T_i} \\ Yn_{T_i} \\ Zn_{T_i} \end{pmatrix}$$

On a, selon z :

$$\eta_{hel} = \arccos\left(\frac{\overrightarrow{Vn_{T_i}}}{\left\|\overrightarrow{Vn_{T_i}}\right\|}.\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}\right) = \arccos\left(\frac{Zn_{T_i}}{\sqrt{Xn_{T_i}^2 + Yn_{T_i}^2 + Zn_{T_i}^2}}\right)$$

et en projetant sur le plan (xOy) :

$$\eta_{twist} = 2\arctan\left(\frac{Yn_{T_i}}{Xn_{T_i}}\right) - \arctan\left(\frac{Yn_{T_{i-1}}}{Xn_{T_{i-1}}}\right)$$

d'où :

$$\overrightarrow{Vn_{T_{i+1}}} = \begin{pmatrix} \cos(\eta_{twist})\sin(\eta_{hel}) \\ \sin(\eta_{twist})\sin(\eta_{hel}) \\ \cos(\eta_{hel}) \end{pmatrix} \qquad \text{(II.5.2.3)}$$

Pour calculer la direction du plan d'empilement des bases mésappariées de la boucle nous procédons donc à une approximation hélicoïdale. C'est-à-dire que nous considérons que la direction empilée idéale que doit adopter la base B_i de la boucle empilée sur le plateau T_{-1}, est proche de la direction qu'aurait adopté un plateau T_0

s'empilant au dessus de T_{-1} dans une hélice. On définit donc la direction normale au plan idéal d'empilement comme étant :

$$\overrightarrow{Vn_{B_{empil}}} = \overrightarrow{Vn_{T_0}} \tag{II.5.2.4}$$

Ce vecteur est donc défini à partir de deux angles de coordonnées sphériques, η_{hel} et η_{twist}, exprimés dans le repère absolu du laboratoire (le repère de Cambridge du dernier plateau de paire de bases de la tige). Cette définition a un sens puisque l'axe de l'hélice régulière de la tige est confondu avec l'axe Oz du repère du laboratoire. Il est possible d'établir des correspondances entre ces angles absolus et les angles définis dans la convention de Cambridge. Ainsi, l'angle η_{hel} peut être décomposé en terme d'angles "d'inclinaison" ("tip" : rotation de la paire de bases autour de Oy) et de "dévers" ("inclination" : rotation de la paire de bases autour de Ox) de la convention de Cambridge (*cf.* ANNEXE : A.2), et l'angle η_{twist} est une expression absolue de l'angle de "torsion" ("twist" : rotation d'une paire de bases par rapport à la suivante autour de l'axe Oz) de cette même convention.

II.5.3 Définition de l'axe de rotation de redressement d'empilement

Pour définir l'axe de rotation de redressement d'empilement, nous devons définir l'origine O_{empil} et le vecteur directeur $\overrightarrow{V_{empil}}$ de l'axe de rotation.

II.5.3.1 Définition du point O_{empil}

Naturellement, une rotation déplace les objets tournés d'autant plus que ceux-ci sont éloignés de l'axe de rotation. Pour satisfaire les conditions 1 et 2 (*cf. supra*) l'axe de rotation doit passer par le fil élastique de la boucle, à proximité du sucre redressé. Ainsi choisi, l'axe préserve (pour de faibles rotations) la position des atomes de la chaîne sucre-phosphate qui sont proches du fil et autorise des mouvements de grande amplitude des atomes de la base qui en sont plus éloignés.

Comme l'axe de redressement passe par le fil élastique, il est possible de définir le point O_{empil} comme le point d'intersection entre l'axe et le fil. Le long du fil, ce point est repéré par son abscisse curviligne $s(O_{empil})$. De façon à satisfaire la condition

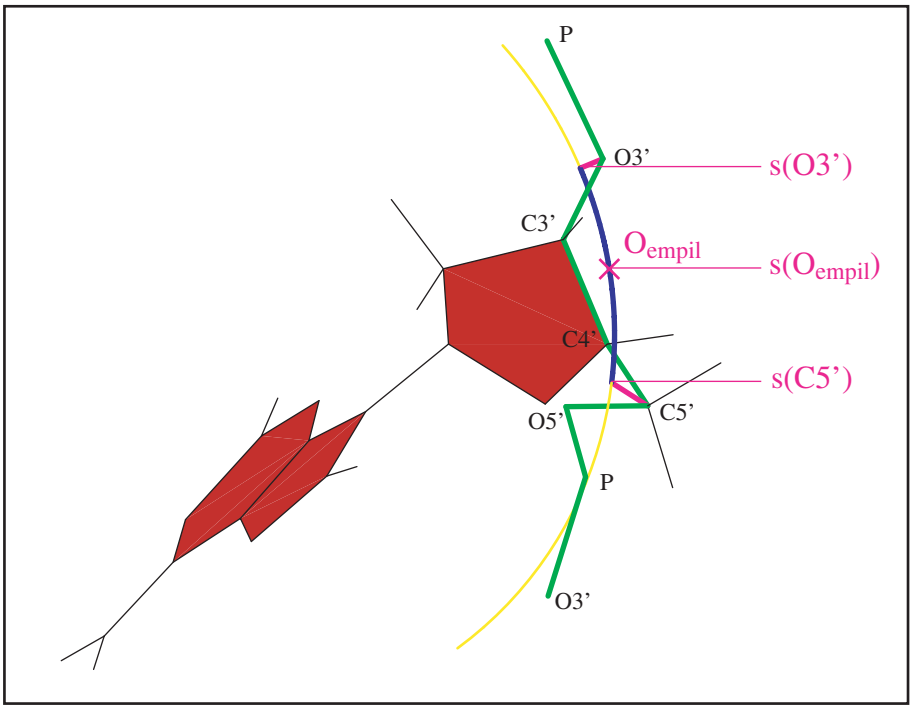


FIG. II.9 : *Origine de l'axe de redressement d'empilement :* *Représentation d'un nucléotide (en rouge), de sa chaîne sucre-phosphate (en vert) et du fil élastique qui lui est associé (en jaune). En violet : la projection des atomes C5' et O3' sur le fil élastique aux points d'abscisses curvilignes $s(C5'_{proj})$ et $s(O3'_{proj})$; en bleu : la portion du fil comprise entre les abscisses curvilignes $s(C5'_{proj})$ et $s(O3'_{proj})$; en noir : le point O_{empil} origine de l'axe de rotation de redressement.*

1 (*cf.* PART. :II.5) nous choisissons $s(O_{empil})$ pour équilibrer les déformations de chaque côté du bloc rigide tourné (le sucre). $s(O_{empil})$ est donc définie par la formule suivante :

$$s(O_{empil}) = \frac{s(C5'_{proj}) + s(O3'_{proj})}{2} \qquad \text{(II.5.3.5)}$$

où

$s(C5'_{proj})$ et $s(O3'_{proj})$ sont les abscisses curvilignes respectives de la projection orthogonale sur le fil élastique des points associés aux coordonnées des atomes C5' et O3', comme représenté dans la figure II.9.

II.5.3.2 Définition du vecteur directeur $\overrightarrow{V_{empil}}$ de l'axe de redressement

L'axe de rotation de redressement d'empilement est défini et optimisé pour empiler la base d'un nucléoside de la boucle sur le dernier plateau de paire de bases de la tige. Pour être optimal il doit tenir compte de la position initiale du nucléoside $N_i(\Omega)$ et de l'orientation finale que l'on souhaite pour la base $B_i(\Omega,\Theta_{empil})$ du nucléotide $N_i(\Omega,\Theta_{empil})$ après redressement. D'un point de vue géométrique, il faut donc qu'après redressement d'empilement, le vecteur $\overrightarrow{Vn_{B_i}}$ normal à la direction du plan de la base B_i soit parallèle au vecteur $\overrightarrow{Vn_{Bempil}}$ normal à la direction idéale d'empilement. *A priori*, il pourrait dépendre de la valeur de l'angle χ_i qui modifie l'orientation relative de la base. Or cela ne doit pas être le cas. En effet, *pour que la base B_i soit parallèle à un plan donné, il faut que la liaison glycosidique (C1'N1 pour les pyrimidines ou C1'N9 pour les purines) du nucléotide N_i soit contenue dans ce plan.* Si cette condition est remplie, alors, une simple rotation d'angle χ_i autour de la liaison glycosidique suffit à rendre parallèle la base B_i à ce plan.

D'autre part, la direction de l'axe de redressement d'empilement doit être choisie pour minimiser les déformations de bord (*cf.* Condition 1 & 2 *supra*) et permettre le redressement voulu moyennant un angle de rotation $\Theta_{empil}(\Omega)_i$ minimal. Cet axe doit donc être défini :

- orthogonalement à la direction de départ de la liaison glycosidique $\overrightarrow{V_{Glyco(\Omega)_i}}$, et
- orthogonalement à la direction finale de la liaison glycosidique $\overrightarrow{V_{Glyco(\Omega,\Theta_{empil})_i}}$,
- dans un plan parallèle à celui de la base $B_i(\Omega,\Theta_{empil})$, donc orthogonalement au vecteur $\overrightarrow{Vn_{B_{empil}}}$.

Géométriquement on a donc :

$$\overrightarrow{V_{empil}} \perp \overrightarrow{V_{Glyco(\Omega)_i}} \quad \text{(II.5.3.6)}$$

$$\overrightarrow{V_{empil}} \perp \overrightarrow{V_{Glyco(\Omega,\Theta_{empil})_i}} \quad \text{(II.5.3.7)}$$

$$\overrightarrow{V_{empil}} \perp \overrightarrow{Vn_{B_{empil}}} \quad \text{(II.5.3.8)}$$

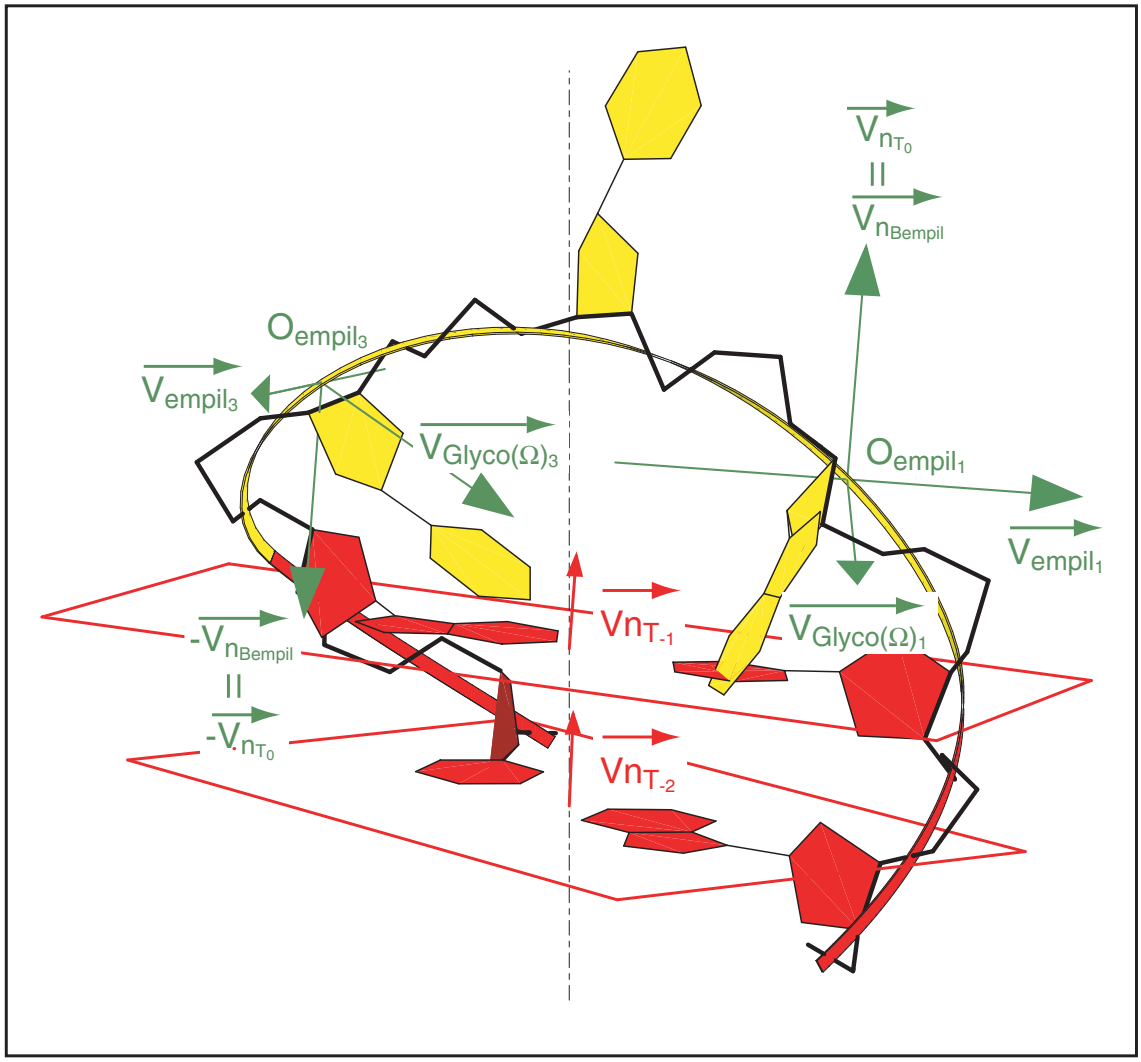


Fig. II.10 : ***Défintion de l'axe de rotation de redressement d'empilement :*** *En rouge : les plans de paire de base des deux derniers plateaux de la tige en hélice et les vecteurs normaux définissant leur direction moyenne; En jaune : les nucléosides de la boucle; En vert : l'axe de rotation de redressement et les vecteurs qui servent à le définir : les vecteurs* $\overrightarrow{V_{Glyco(\Omega,\Theta_{empil})_i}}$ *associés aux liaisons glycosidiques et les vecteurs normaux au plan idéal d'empilement.*

La relation II.5.3.7 ne peut servir à la définition du vecteur $\overrightarrow{V_{empil}}$, car elle comporte un paramètre inconnu, le vecteur $\overrightarrow{V_{Glyco(\Omega,\Theta_{empil})_i}}$. En effet, celui-ci est un des éléments que doit déterminer la rotation de redressement d'empilement. Par contre les deux relations II.5.3.7 et II.5.3.8 suffisent à déterminer mathématiquement le vecteur $\overrightarrow{V_{empil}}$ par la relation :

$$\overrightarrow{V_{empil}} = \overrightarrow{Vn_{B_{empil}}} \wedge \overrightarrow{V_{Glyco(\Omega)_i}} \qquad \text{(II.5.3.9)}$$

En résumé l'axe de rotation de redressement d'empilement associé à un nucléoside $N_i(\Omega)$ est défini par le couple $(O_{empil},\overrightarrow{V_{empil}})$ tel que :

$$\begin{cases} O_{empil},\ tel\ que\ s(O_{empil}) = & \dfrac{s(C5'_{proj}) + s(O3'_{proj})}{2} \\ \overrightarrow{V_{empil}} = & \overrightarrow{Vn_{B_{empil}}} \wedge \overrightarrow{V_{Glyco(\Omega)_i}} \end{cases} \qquad \text{(II.5.3.10)}$$

II.5.4 Définition de l'angle $\Theta(\Omega)_i$ de redressement d'empilement

L'axe de rotation de redressement d'empilement $(O_{empil},\overrightarrow{V_{empil}})$ est défini pour assurer le redressement de la liaison glycosidique du nucléotide $N_i(\Omega)$ dans une direction globalement parallèle au dernier plateau de paire de bases de la tige T_{-1} (direction donnée par le vecteur normal $\overrightarrow{Vn_{B_{empil}}}$). Il est donc théoriquement possible de définir un angle Θ_{plan} qui permet de placer parfaitement la base $B_i(\Omega,\Theta_{plan})$ dans un plan orthogonal au vecteur $\overrightarrow{Vn_{B_{empil}}}$. Pourtant, lorsque l'on applique cette transformation, la structure obtenue n'est pas bonne. En effet, :

- le mésappariement obtenu est quasiment coplanaire ce qui est contraire aux observations faites sur les structures déjà résolues qui présentent des inclinaisons non-négligeables des bases mésappariées (*cf.* PART. : I.4.3.4)
- la distance moyenne entre le plateau T_{-1} et le plateau moyen défini par les bases redressées est supérieure à 3,34 Å, ce qui conduit les atomes des bases redressées à être à la fois trop proches des atomes de la base centrale de la boucle et trop éloignés des atomes du plateau T_{-1},

Toutes ces constatations laissent penser que la base redressée de l'angle Θ_{plan} est trop tournée et qu'en fait la base $B_i(\Omega)$ devrait être seulement redressée d'un angle $\Theta_{exclu.}$ inférieur à Θ_{plan}.

Pour déterminer l'angle $\Theta_{exclu.}$ nous postulons que, comme dans les hélices régulières, les atomes de deux plateaux empilés doivent être séparés d'une distance moyenne de 3,34 Å (*cf.* PART. : I.4.3.1) du dernier plateau de paire de bases de la tige. L'angle $\Theta_{exclu.}$ est alors défini comme l'angle de rotation de redressement d'empilement suffisant pour placer les atomes de la base redressée à une distance moyenne de 3,34 Å du plan $P_{T_{-1}}$.

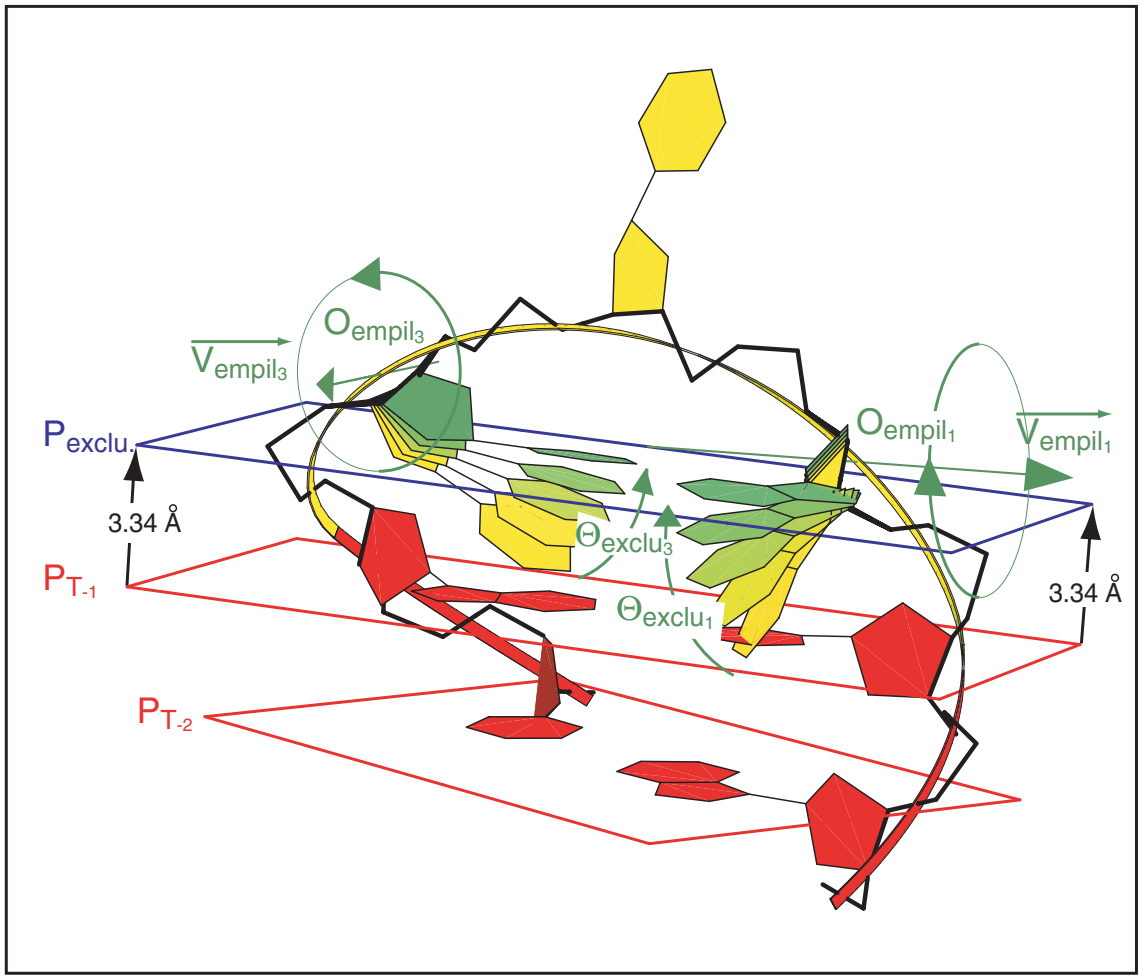


FIG. II.11 : *Angle de redressement d'empilement :* *Définition de l'angle de redressement d'empilement $\Theta_{exclu.}$ assurant une distance minimale de 3,34 Å entre les atomes du dernier plateau de paire de bases de la tige et l'atome de la base redressée le plus éloigné de l'axe de rotation de redressement.*

Pour le définir nous introduisons un plan $P_{exclu.}$ qui joue le rôle d'une limite d'exclusion au-dessous de laquelle les atomes des bases redressées ne doivent pas se trouver. Ce plan est défini parallèlement au plan $P_{T_{-1}}$, à une distance de 3,34 Å de celui-ci. Le plan $P_{exclu.}$ est alors défini par le couple ($G_{exclu.}$, $\overrightarrow{Vn_{T_{-1}}}$), avec :

$$G_{exclu} = G_{T_{-1}} + \overrightarrow{Vn_{T_{-1}}}$$

où $G_{T_{-1}}$ étant le centre géométrique des atomes des bases formant le plateau T_{-1}.

Nous postulons que si l'atome le plus éloigné de l'axe de rotation, soit l'HN6B pour les adénines, l'O4 pour les thymines, l'HN4B pour les cytosines et l'O6 pour

les guanines, se trouve après rotation de redressement d'empilement dans le plan $P_{exclu.}$, alors les atomes de la base redressée seront à une distance proche de 3,34 Å du plateau T_{-1}. La définition mathématique de l'angle $\Theta_{exclu.}$ est donc la suivante :

Étant donnés :

- $M_i(\Omega)$ le point de la base $B_i(\Omega)$ qui est le plus éloigné de l'axe de rotation $(O_{rot},\overrightarrow{V_{empil}})_i$,
- $M_i(\Omega,\Theta_{exclu.})$ le point $M_i(\Omega)$ tourné par la rotation de redressement d'empilement $\Theta_{exclu.}$ de la base $B_i(\Omega,\Theta_{exclu.})$.

l'angle $\Theta_{exclu.}$ est l'angle qui permet de placer le point $M_i(\Omega,\Theta_{exclu.})$ dans le plan $P_{exclu.}$. Il vérifie :

$$\overrightarrow{G_{exclu}M_i(\Omega,\Theta_{exclu.})}.\overrightarrow{Vn_{T_{-1}}} = 0$$

Cette équation permet de définir un système d'équations qui admet deux solutions (la base pouvant tourner dans un sens ou dans l'autre). La solution retenue est celle d'angle minimal.

II.6 Obtention de la structure à l'échelle atomique

Toutes les étapes précédentes de modélisation interviennent à une échelle moléculaire globale et mésoscopique. Afin de rejoindre les autres approches de modélisation, nous terminons notre protocole de modélisation à l'échelle atomique. L'étude des conformations à ce niveau de détail utilise les approches classiques de mécanique et dynamique moléculaire.

II.6.1 Obtention de la structure atomique BCE_{min} par minimisation d'énergie des structures BCE_{opt}

La minimisation d'énergie permet de relaxer les structures BCE_{opt} pour obtenir les structures BCE_{min}. Cette étape est pensée comme la première étape de modélisation

à l'échelle atomique de l'approche BCE. En effet, notre approche de modélisation est hiérarchique. Elle s'emploie à modéliser dans un premier temps les structures globales de plus faible énergie caractéristique des fonctions des bio-molécules, pour, dans une deuxième temps, prendre en compte les faibles déplacements locaux de forte énergie.

Cette minimisation est nécessaire pour obtenir une conformation respectant le principe d'énergie minimale que suivent toutes les structures moléculaires. En effet, la manipulation des structures par blocs d'atomes introduit nécessairement des violations de géométries à l'échelle des longueurs et angles de valence, même si les déformations sont conçues à chaque étape pour éviter ce type de déformation. Cette minimisation est très courte, puisque la structure soumise à la minimisation d'énergie est déjà très proche d'un minimum énergétique global. Cette proximité du minimum global est effectivement observée dans cette approche de modélisation au travers de l'accord avec les structures dérivées de l'expérimentation.

La conséquence de cette relaxation finale est de rétablir les géométries à l'échelle des liaisons atomiques (longueurs et angles de valence). Elle ne modifie que très peu la structure globale mise en place dès la structure BCE_{opt} comme nous le verrons au chapitre III. Une raison pour cela est que les déformations introduites autour de l'abscisse curviligne s sont souvent de signe opposé à celles introduites à l'abscisse curviligne s+Δs (*cf.* FIG. : III.2). Cette propriété facilite les raffinements d'énergie .

II.6.2 Dynamique moléculaire en solvant aqueux explicite des structures BCE_{min}

Les seules données cartésiennes d'une structure moléculaire ne donnent qu'une vision partielle de la molécule. À température ambiante, elle peut correspondre selon les cas à un état moyen ou à un état limite de moindre énergie. Elle omet cependant un aspect fondamental dont dépend une activité biologique : sa dynamique. Cette dynamique, associée au phénomène de flexibilité de la molécule, est importante pour comprendre les modes d'interaction de la molécule avec des ligands, pour comprendre les processus de réparation, de réarrangement des ADN, de réactivité catalytique des ARN, etc.

Différentes revues montrent en détail que les simulations par ordinateur permettent maintenant d'explorer finement les conformations dynamiques des acides nucléiques

[87–89]. Ceci est principalement dû à la possiblité de conduire des simulations de dynamique moléculaire de l'ordre de la nano-seconde ou plus, en solvant explicite, grâce au progrès des capacités de calcul des ordinateurs, à l'amélioration des champs de forces et à l'optimisation des algorithmes de calcul P.M.E. (Particle Mesh Ewald). Des simulations de dynamique moléculaire ont récemment mis en évidence la flexibilité différentielle de l'ADN et de l'ARN en double hélice [90,91]. À partir de ces programmes, de nouveaux protocoles de dynamique moléculaire ont été développés [92] et ont montré "qu'un niveau remarquable de résolution dynamique et atomique peut être atteint" [87].

Nous avons d'abord utilisé ce protocole dans une étude portant sur la flexibilité différentielle des dimères de thymines en simple-brin. Nous avons ensuite effectué des simulations sur la structure 1BJH-AAA construite avec l'approche BCE. Ces simulations de plusieurs nano-secondes ont mis en évidence la bonne stabilité structurelle des structures élaborées avec notre approche. Les conformations explorées par la dynamique restent en effet proches de la structure BCE. Celle-ci est donc bien proche d'un minimum énergétique global.

II.7 BCE utilise implicitement un champ de force mésoscopique

II.7.1 Déformation minimale et énergie minimale

Le critère fondamental de validité d'une structure est sa stabilité thermodynamique. C'est notamment ce principe de recherche des structures de moindre énergie qui est utilisé dans l'exploration des espaces conformationnels. Fondamentalement, l'approche BCE s'appuie sur un principe de déformation de moindre énergie du fil élastique. L'approche BCE fait donc partie d'une démarche de Physique fondamentale. À l'échelle globale des courbes, le principe d'énergie minimale est assuré par la théorie de l'élasticité. Son originalité est de montrer qu'elle s'applique à l'ensemble de la chaîne sucre-phosphate d'une molécule en épingle à cheveux d'ADN ou d'ARN.

Dans le cas du fil flexible, isotrope et inextensible et dans les conditons limites d'une barre encastrée à ses deux extrémités, il est possible de montrer [80,86,93,94]

que le problème physique de la déformation de la barre peut être formulé de façon adimensionnée. Dans ce cas particulier, la solution du problème posé par l'approche BCE ne dépend pas des constantes physiques de rigidité du fil. Il est donc possible de calculer la trajectoire dont la déformation est d'énergie minimale sans connaître la constante de rigidité du fil. Sous cette forme ou dans ce cas particulier, BCE applique un principe d'énergie minimale de déformation du fil qui peut s'exprimer sous forme uniquement géométrique.

À l'échelle de déformation intermédiaire de déplacement des résidus, toute déformation a un coût énergétique qui doit être pris en compte. Les structures sont considérées comme d'autant moins favorables que les déformations sont importantes. Toutes les déformations introduites seront donc choisies pour être minimales.

À l'échelle atomique de déformation moléculaire, ce même principe d'énergie minimale est le fondement de la modélisation moléculaire par minimisation d'énergie.

II.7.2 Approximation géométrique des sphères dures atomiques pour l'empilement des bases dans la boucle

Afin de prendre en compte les volumes des atomes lors de l'opération de redressement d'empilement, l'angle de redressement est calculé pour assurer une distance moyenne de 3,34 Å entre les coordonnées des atomes de la base redressée et les coordonnées des atomes du dernier plateau de paire de bases de la tige. Cette distance, couramment utilisée pour décrire les distances entre plateaux consécutifs dans une hélice, permet d'assurer une exclusion stérique entre les atomes de la base de la boucle et ceux des bases de la tige.

II.7.3 Géométrie et évaluation d'une liaison hydrogène sur une structure ponctuelle

II.7.3.1 Définition de la liaison hydrogène

La liaison hydrogène est une interaction de nature électrostatique et quantique. Elle se met en place lorsqu'un groupement donneur de proton est proche d'un doublet

libre d'électrons porté par un autre atome. Le groupement donneur de proton (Don.-H) est en général constitué d'un atome d'hydrogène (H) lié covalemment à un atome électronégatif, le donneur de proton (Don.=Oxygène ou Azote). D'un point de vue électrostatique, cette liaison est due à la polarisation des groupes donneurs, proton et accepteur. Entre un proton et un atome électronégatif la distribution électronique de la liaison covalente est fortement déplacée vers l'atome donneur D. La charge partielle positive qui apparaît sur le proton peut alors engager une interaction de type électrostatique avec un doublet libre d'électrons porté par un atome accepteur Acc..

L'établissement de liaisons hydrogène est le critère principal de formation d'un mésappariement entre deux bases. Au nombre de deux dans les plateaux A-T(U) et de trois dans les plateaux G-C, elles peuvent également s'établir entre le proton 2'OH des riboses et des atomes d'azote ou d'oxygène d'autres bases lors de repliements tertiaires dans l'ARN par exemple. Dans les (més-)appariements, elles interviennent en général entre des groupements donneurs de protons amines, imines ou hydroxyles, et des doublets libres d'électrons portés par des atomes d'oxygène ou d'azote. D'autres types de liaisons hydrogène ont été documentés par certains auteurs. Elles interviendraient entre des groupements donneurs de type C-H et les accepteurs classiques (N ou O) et stabiliseraient notamment le plateau A-T par une troisième liaison hydrogène de type ADE[H2]-[O2]THY [38, 72, 73].

Dans les approches de modélisation classiques fondées sur l'utilisation de champs de forces, comme dans les logiciels AMBER ou CHARMm, il n'y a pas de terme explicite dans la fonction de potentiel pour modéliser les interactions de liaison hydrogène. D'abord traitées comme une somme de contributions de type électrostatique, van der Waals et d'un terme en 10-12, $v_{HB}(r) = \frac{A}{r^{12}} - \frac{B}{r^{10}}$ (avec r la distance séparant le proton de l'atome portant le doublet libre d'électrons), elles sont désormais traitées uniquement comme une contribution électrostatique et de van der Waals. Les charges partielles définies dans le champ de force pour les atomes pouvant engager des liaisons hydrogène sont un peu plus fortes que les charges habituellement rencontrées pour ces types d'atomes. La directionnalité des liaisons hydrogène n'est pas traitée explicitement. Ce sont les encombrements stériques (terme de van der Waals) qui conditionnent les positions accessibles par le proton et donc les orientations admises.

Ainsi les liaisons hydrogène sont traitées comme une somme de termes répulsifs et attractifs de type Lennard-Jones et un terme attractif coulombien. Les déviations

éventuelles à la géométrie des liaisons hydrogène n'est pas directement prise en compte. En effet, pour évaluer rigoureusement la qualité d'une liaison hydrogène, une fonction qui ne dépend que de la distance est insuffisante. La liaison hydrogène présente en effet une géométrie particulière caractérisée par la distance séparant le proton de l'accepteur portant le doublet libre, mais aussi par :

- l'alignement de la liaison Don.-H avec l'atome accepteur,
- l'alignement du proton avec la direction de l'orbitale portant le doublet libre (DL).

Dans cet esprit d'autres potentiels ont été développés pour prendre en compte la directionnalité des liaisons hydrogène [32] :

$$v_{HB} = \left(\frac{A}{r_{H\cdots Acc}^{12}} - \frac{C}{r_{H\cdots Acc}^{10}} \right) cos^2\left(\theta_{Don-H\cdots Acc}\right) cos^4\left(\theta_{H\cdots Acc-DL}\right)$$

avec :

$r_{H\cdots Acc}$, la distance entre le proton et l'atome accepteur portant le doublet libre,

$\theta_{Don-H\cdots Acc}$, l'angle formé par l'atome donneur, le proton et l'atome accepteur et

$\theta_{H\cdots Acc-DL}$, l'angle formé par le proton, l'atome accepteur et la direction de l'orbitale portant le doublet libre d'électrons.

II.7.3.2 Évaluation de la présence des liaisons hydrogène dans BCE

Dans BCE nous ne disposons pas de champ de force. Les calculs de potentiels d'interaction de liaison hydrogène nous sont donc pour le moment interdits. Dans AMBER, la géométrie de la liaison hydrogène est généralement prise en compte indirectement, et notamment dans une paire de bases, par les termes de Lennard-Jones et d'interaction électrostatique (*cf.* PART. : I.2.3.1). Pour évaluer la présence de liaisons hydrogène nous nous limiterons à mettre en place une fonction d'évaluation purement géométrique. Néanmoins, cette fonction sera élaborée dans le souci de simuler au mieux les forces impliquées dans la formation des interactions de liaison hydrogène habituellement rencontrées dans une paire de bases.

Notre approche veut tenir compte de tous les paramètres géométriques qui participent à l'établissement de la liaison hydrogène. Aussi la fonction de score de liaison hydrogène S_H que nous utilisons prend en compte l'éloignement du proton de l'atome accepteur, mais aussi l'alignement de la liaison hydrogène (H-.) avec l'atome donneur de proton et avec la direction de l'orbitale portant les doublets libres d'électrons accepteur. Cette fonction est établie comme un produit de fonctions de probabilité de liaison hydrogène. Chaque terme du produit est une exponentielle d'une fonction d'écart quadratique centrée sur la valeur idéale d'alignement ou de distance des groupements donneurs et accepteurs de protons.

II.7.3.3 Terme d'éloignement du proton et de l'atome accepteur

La distance idéale séparant le proton de l'atome accepteur est de 1,9 Å. Au dessous de cette distance les volumes propres des deux atomes provoquent des gènes stériques et l'on considère que l'interaction est de moins en moins favorable jusqu'à 1,5 Å, et interdite au delà de cette distance seuil. Au dessus de cette distance, l'interaction électrostatique perd de sa force et au delà de la distance seuil de 3,5 Å, on considère que l'interaction entre les deux atomes ne peut plus être qualifiée de liaison hydrogène.

Pour tenir compte de toutes ces données, nous avons défini la terme d'éloignement $S_{H_{el.}}$ de la fonction de score de liaison hydrogène de la façon suivante :

$$\forall r_{H-Acc.}, \begin{cases} r_{H-Acc.} \in [r_{seuil\,inf.},\ r_{seuil\,sup.}] & \Rightarrow S_{H_{el.}} = e^{-\tau\,(r_{H-Acc.} - r_0)^2} \\ r_{H-Acc.} \in [0,\ r_{seuil\,inf.}[\ \cup\ ]r_{seuil\,sup.},\ +\infty[& \Rightarrow S_{H_{el.}} = 0 \end{cases}$$

Avec,

$r_{H-Acc.}$, la distance entre le proton et l'atome accepteur,

r_0 =1,9 Å la distance de référence entre le proton et l'atome accepteur,

$r_{seuil\,sup.}$ =3,5 Å la distance au delà de laquelle il n'y a plus de liaison hydrogène,

$r_{seuil\,inf.}$ =1,5 Å la distance minimale permettant l'établissement d'une liaison hydrogène,

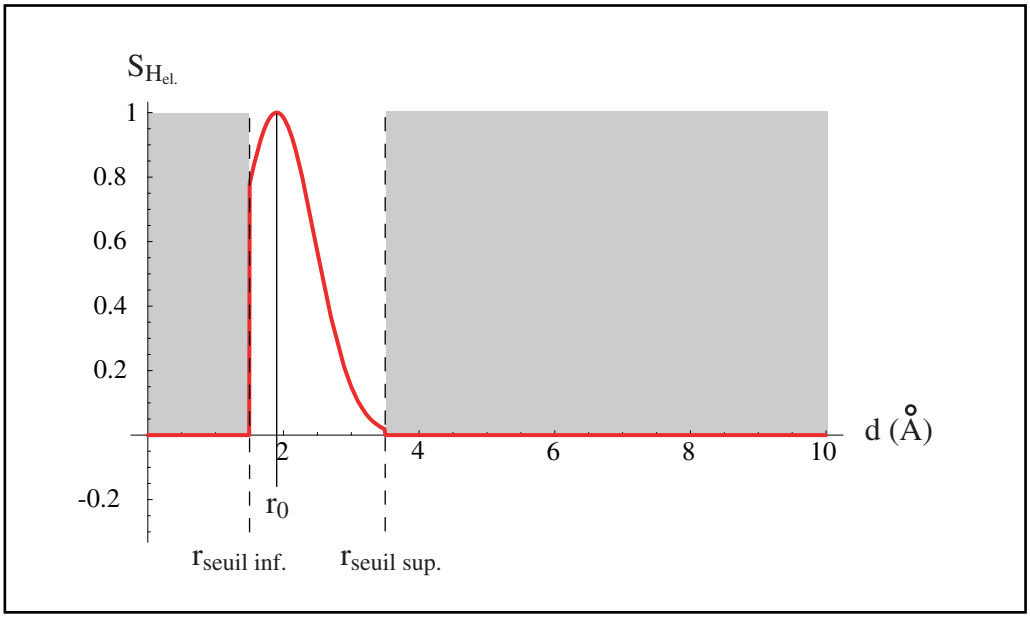


Fig. II.12 : ***Fonction de score d'éloignement des liaisons hydrogène :*** *En rouge : la fonction de score d'éloignement de liaison hydrogène; en noir continu : la distance de référence entre le proton et l'atome accepteur de liaisons hydrogène; en noir pointilé : les distances seuils au delà desquelles les liaisons hydrogène ne se forment plus.*

$\tau = \frac{\pi}{2}$, un facteur choisi pour que la fonction de score adopte une forme proche de la forme attendue qualitativement pour la formation des liaisons hydrogène de longueurs comprises entre 1,5 Å et 3,5 Å. Notamment, il permet d'atteindre un score quasiment nul lorsque l'élongation de la liaison approche 3,5 Å et pénalise les liaisons courtes jusqu'à 1,5 Å sans les exclure.

Cette fonction de score a la forme d'une Gaussienne tronquée, dont le maximum est centré sur la distance de référence (*cf.* Fig. : II.12), et dont l'aire sous la courbe est de 1,07. Ces propriétés donnent à cette fonction de score un caractère de densité de probabilité.

II.7.3.4 Termes d'orientation des liaisons hydrogène

Les termes d'orientation choisis sont au nombre de deux (*cf.* Fig. : II.13). Le premier concerne l'alignement de la liaison hydrogène H-Acc. avec la liaison covalente Don.-H et le second concerne l'alignement de la liaison hydrogène H-Acc avec la direction de l'orbitale de l'atome accepteur présentant le doublet libre

d'électrons. Ces deux alignements sont évalués respectivement par les angles $\theta_{Don.}$ et $\theta_{Acc.}$, qui dans une configuration idéale de liaison hydrogène prennent la valeur de π radians. On considère qu'un écart supérieur à $\pm\frac{\pi}{3}$rad. à ces valeurs de référence rend impossible la formation de la liaison.

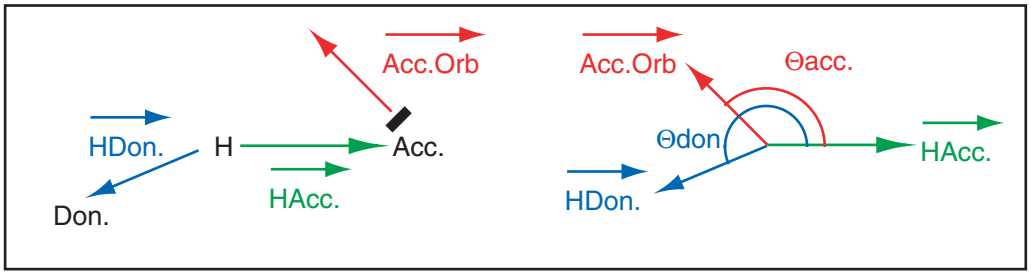


Fig. II.13 : ***Directionalité des liaisons hydrogène intervenant entre les bases des acides nucléiques.***

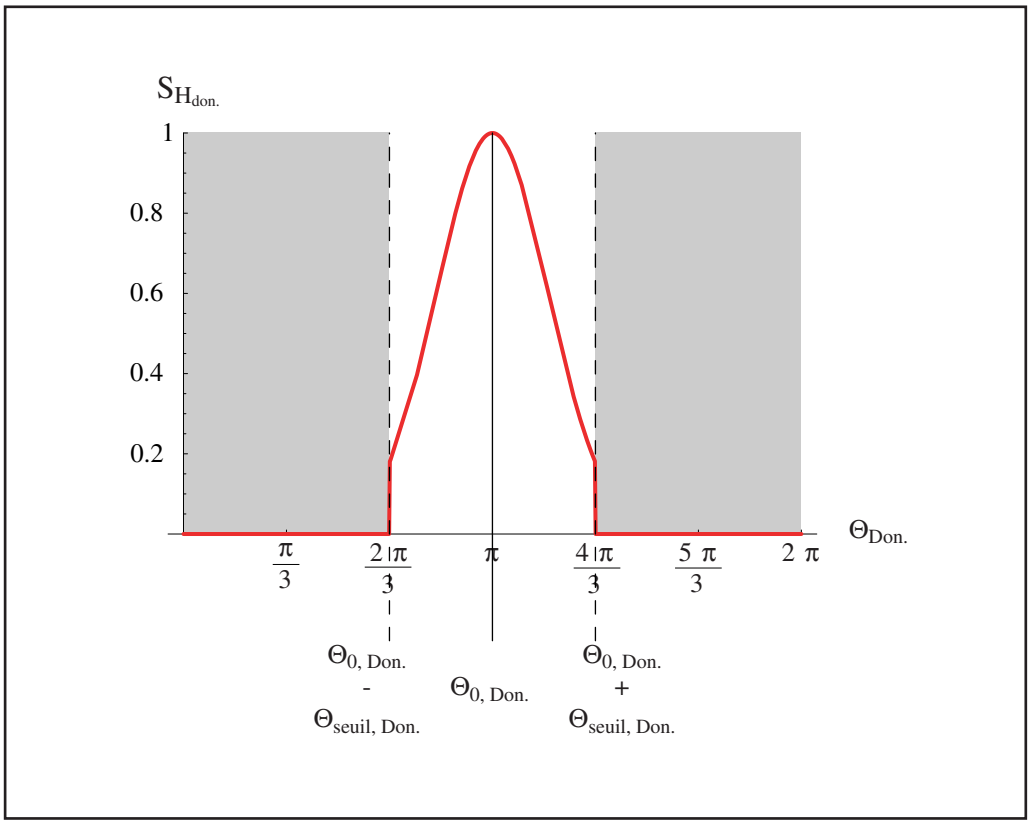


Fig. II.14 : ***Fonction de score d'orientation des liaisons hydrogène :*** *En rouge : la fonction de score d'orientation des liaisons hydrogène; en noir continu : la valeur de référence de l'angle mesurant l'orientation; en noir pointille : les angles limites au delà desquelles on considère que les liaisons ne peuvent plus se former.*

Afin de pondérer la fonction de score de liaison hydrogène par les écarts aux alignements idéaux, nous introduisons donc les deux termes d'alignement $S_{H_{Don.}}$ et $S_{H_{Acc.}}$ suivants :

$$\forall \theta_{Don.}, \begin{cases} \theta_{Don.} \in [\theta_{0,\,Don.} - \theta_{seuil,\,Don.}, \theta_{0,\,Don.} + \theta_{seuil,\,Don.}] & \Rightarrow S_{H_{Don.}} = e^{-\tau\,(\theta_{Don.} - \theta_{0,Don.})^2} \\ \theta_{Don.} \notin [\theta_{0,\,Don.} - \theta_{seuil,\,Don.}, \theta_{0,\,Don.} + \theta_{seuil,\,Don.}] & \Rightarrow S_{Don.} = 0 \end{cases}$$

$$\forall \theta_{Acc.}, \begin{cases} \theta_{Acc.} \in [\theta_{0,\,Acc.} - \theta_{seuil,\,Acc.}, \theta_{0,\,Acc.} + \theta_{seuil,\,Acc.}] & \Rightarrow S_{H_{Acc.}} = e^{-\tau\,(\theta_{Acc.} - \theta_{0,Acc.})^2} \\ \theta_{Acc.} \notin [\theta_{0,\,Acc.} - \theta_{seuil,\,Acc.}, \theta_{0,\,Acc.} + \theta_{seuil,\,Acc.}] & \Rightarrow S_{H_{Acc.}} = 0 \end{cases}$$

avec,

$\theta_{Don.}$, l'angle formé par la liaison hydrogène et la liaison donneur-accepteur. On a : $\theta_{Don.}=\left(\widehat{\overrightarrow{H\,Don.}, \overrightarrow{H\,Acc.}}\right)=(\widehat{Don.\,H\,Acc.})$,

$\theta_{Acc.}$, l'angle formé par la liaison hydrogène et la direction de l'orbitale des doublets libres d'électrons de l'atome accepteur. On a : $\theta_{Acc.}=\left(\widehat{\overrightarrow{DL}, \overrightarrow{H\,Acc.}}\right)$, avec $\overrightarrow{DL}$ le vecteur associé à la direction de l'orbitale acceptrice,

$\theta_{0,Don.}$ $=\theta_{0,Acc.}=\pi$ rad. : l'angle de référence pour $\theta_{Don.}$ et $\theta_{Acc.}$ dans une liaison hydrogène idéale,

$\theta_{seuil,\,Don.}=\theta_{seuil,\,Acc.}=\frac{\pi}{3}$rad. : la déviation maximale tolérée aux angles de références $\theta_{0,Don.}$ et $\theta_{0,Acc.}$.

$\tau=\frac{\pi}{2}$, un facteur choisi pour que la fonction de score adopte une forme proche de la forme attendue qualitativement pour la formation des liaisons hydrogène avec un écart toléré à l'alignement compris entre -$\frac{\pi}{3}$rad. et +$\frac{\pi}{3}$rad. Notamment, il permet d'atteindre un score faible lorsque la déviation à l'alignement s'approche de $\frac{\pi}{3}$rad.

Cette fonction de score a la forme d'une Gaussienne tronquée, dont le maximum est centré sur l'alignement idéal (*cf.* FIG. : II.14), et dont l'aire sous la courbe est de 1.32. Ces propriétés donnent à cette fonction de score un caractère de densité de probabilité.

II.7.3.5 Fonction d'évaluation de score de liaison hydrogène

La fonction de score de liaison hydrogène S_H est le produit du terme d'éloignement et des deux termes d'alignement. Sa forme complète est la suivante :

$$S_H = S_{H_{el.}} \cdot S_{H_{Don.}} \cdot S_{H_{Acc.}}$$

$$\begin{cases} \left[\begin{array}{l} \forall \left(r_{H-Acc.},\, \theta_{Don.},\, \theta_{Acc.}\right), \left| \begin{array}{rcll} r_{H-Acc.} & \in & [r_{seuil\,inf.},\, r_{seuil\,sup.}] & et \\ \theta_{Don.} & \in & [\theta_{0,\,Don.} - \theta_{seuil,\,Don.}, \theta_{0,\,Don.} + \theta_{seuil,\,Don.}] & et \\ \theta_{Acc.} & \in & [\theta_{0,\,Acc.} - \theta_{seuil,\,Acc.}, \theta_{0,\,Acc.} + \theta_{seuil,\,Acc.}] & \end{array} \right. \\ alors, \\ S_H = e^{-\tau \left(r_{H-Acc.} - r_0\right)^2} \cdot e^{-\tau \left(\theta_{don.} - \theta_{0,don.}\right)^2} \cdot e^{-\tau \left(\theta_{acc.} - \theta_{0,acc.}\right)^2} \end{array} \right. \\ \left[\begin{array}{l} \forall \left(r_{H-Acc.},\, \theta_{Don.},\, \theta_{Acc.}\right), \left| \begin{array}{rcll} r_{H-Acc.} & \in & [0,\, r_{seuil\,inf.}[\cup]r_{seuil\,sup.},\, +\infty[& ou \\ \theta_{Don.} & \notin & [\theta_{0,\,Don.} - \theta_{seuil,\,Don.}, \theta_{0,\,Don.} + \theta_{seuil,\,Don.}] & ou \\ \theta_{Acc.} & \notin & [\theta_{0,\,Acc.} - \theta_{seuil,\,Acc.}, \theta_{0,\,Acc.} + \theta_{seuil,\,Acc.}] & \end{array} \right. \\ alors, \\ S_H = 0 \end{array} \right. \end{cases}$$

Cette fonction permet de calculer un score de liaison hydrogène pour un groupement donneur de proton (Don.-H) et un groupe accepteur de proton (Acc.) donné, dans une conformation spatiale donnée. C'est une fonction d'évaluation ponctuelle d'une liaison hydrogène pour une molécule dans une conformation donnée. Elle sera donc adaptée à l'évaluation d'une liaison hydrogène potentielle pour un couple [$N_1(\Omega_i,\Theta_i)$-$N_3(\Omega_j,\Theta_j)$] unique.

Produit de trois fonctions dont les valeurs renvoyées sont comprises entre 0 et 1, cette fonction renvoie une valeur elle aussi comprise entre 0 et 1. Plus la liaison hydrogène est proche de la liaison hydrogène idéale, plus le score sera proche de 1. Si un des paramètres d'alignemement ou d'éloignement n'est pas compris entre les valeurs seuils qui lui sont associées, alors la fonction de score renverra la valeur 0, synonyme d'absence de liaison hydrogène.

Chaque contribution $S_{H_{el.}}$, $S_{H_{Don.}}$ et $S_{H_{Acc.}}$ est comparable puisque les fonctions de score sont toutes comprises entre 0 et 1 et que l'aire sous chacune de ces courbes est

proche de 1. Ces propriétés en font des fonctions qui peuvent être interprétées de façon approximative comme des densités de probabilité.

Chapitre III

Les épingles à cheveux d'ADN et d'ARN

Nous cherchons ici à comprendre l'architecture trimensionnelle des motifs en épingle à cheveux d'un ensemble de boucles à l'aide de l'approche BCE. Nous avons retenu huit structures en épingle à cheveux d'ADN et d'ARN sur la base de critères de disponibilité des structures dans la Protein Data Bank, pour la qualité de leur détermination ainsi que pour leurs documentations bibliographiques parmi les plus complètes. Il s'agit de quatre structures d'ADN - les trois tri-boucles 1BJH-AAA, 1XUE-GCA, 1ZHU-GCA, la tétra-boucle 1AC7-GTTA - et les quatre tétra-boucles d'ARN 1AUD-UUCG, 1B36-UUCG, 1C0O-UUCG, 1HLX-UUCG présentées dans le chapitre I. Toutes ces structures présentent des conformations des bases de la boucle où les deux bases extrémales sont dans des géométries quasi-coplanaires appariées ($N_1 \cdots N_3$ dans les tri-boucles et $N_1 \cdots N_4$ dans les tétra-boucles).

Le propos de cette étude est de rechercher s'il est possible de retrouver la conformation globale établie au moyen d'autres approches de modélisation à l'aide du simple concept de chaîne rigide de biopolymères (BCE). Les données expérimentales utilisées pour modéliser ces conformations proviennent de fichiers de coordonnées de la PDB. Notre travail consistera donc à chercher à reproduire au mieux les conformations proposées dans ces fichiers au moyen du nombre très restreint de d.d.l. à notre disposition dans BCE. Par ce travail, nous cherchons à savoir si notre approche est capable de prédire la trajectoire globale de la molécule au moyen de la théorie de l'élasticité quelque soit la nature de la chaîne sucre-phosphate (acide ribo- ou désoxyribonucléique), et le nombre de nucléotides dans la chaîne d'ADN.

Dans le cas où cette hypothèse serait satisfaite, nous chercherons également à savoir si les degrés de liberté Ω et χ permettent de placer adéquatement les bases et les sucres des nucléotides.

Dans une première partie, après avoir présenté les difficultés qui découlent de l'utilisation des fichiers PDB comme source de données expérimentales, nous détaillerons le protocole utilisé pour modéliser toutes ces structures. Ensuite, nous présenterons les résultats obtenus concernant la modélisation de la trajectoire de la chaîne sucre-phosphate. Nous établirons en quoi notre approche permet de traiter aussi bien des molécules d'ADN et d'ARN. Nous comparerons les structures obtenues avec BCE aux structures proposées par les auteurs originaux à différentes échelles. À l'échelle globale de la trajectoire de la molécule, nous étudierons la conformation de la chaîne sucre-phosphate sur l'ensemble de la molécule et sur la partie en boucle. De la même façon, à l'échelle intermédiaire du placement des nucléotides, nous comparerons les structures BCE et les structures publiées. Nous discuterons notamment de l'aptitude et de la suffisance des d.d.l. Ω et χ développés dans BCE pour modéliser les conformations des épingles à cheveux comportant des appariements dans la boucle.

III.1 Position du problème : les trajectoires de tri- et tétra-boucles d'ADN et d'ARN prédites par la théorie de l'élasticité

III.1.1 Les conditions d'encastrements des boucles sur une tige d'ADN et d'ARN

Bien que de structures secondaires comparables, les épingles à cheveux d'ADN et d'ARN présentent en trois dimensions des structures très différentes. La tige de ces épingles à cheveux se structure comme des hélices B dans le cas des molécules d'ADN et comme des hélices A pour les molécules d'ARN. Les hélices régulières utilisées pour modéliser les tiges des épingles à cheveux d'ADN et d'ARN reprennent ce caractère.

Les hélices des épingles à cheveux d'ADN et d'ARN présentent donc des géométries différentes, ainsi que les fils en hélice ajustés sur les structures atomiques (rayon et

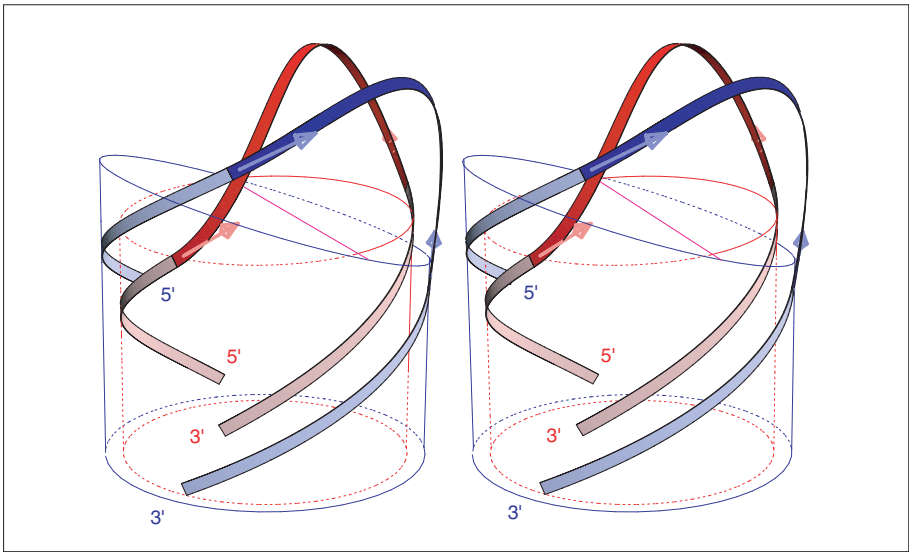


Fig. III.1 : ***Vue stéréoscopique des trajectoires attendues des courbes des épingles à cheveux d'ADN et d'ARN*** *depuis le petit sillon d'après la théorie de l'élasticité des barres minces pour des tétraboucles d'acides nucléiques. En rouge, les ruban associés aux fils de l'hélice et de la boucle d'une structure d'ADN et en bleu ceux d'une structure d'ARN. Les rayons des cylindres et des sections des cylindres sont ceux des squelettes des chaînes sucre-phosphates et sont à la même échelle. Les sections supérieures des cylindres représentent le plan moyen du dernier plateau de paire de bases de la tige.*

pas des hélices régulières associés aux fils). La conséquence directe dans le cadre de l'approche BCE est que les paramètres d'encastrement définis aux extrémités de ces fils, utilisés pour prédire la trajectoire de la boucle avec la théorie de l'élasticité, sont très différents pour l'ADN et l'ARN. Dans le cas des épingles à cheveux d'ADN et d'ARN, les tangentes aux extrémités des hélices de la tige, adoptent une symétrie pseudo-dyadique par rapport à l'axe (Ox) du dernier plateau de paire de bases de la tige, et pointent vers le haut tangentiellement au cylindre associé à la double hélice de la tige. Les trajectoires prédites par la théorie de l'élasticité sont donc très différentes comme on peut l'observer avec la figure III.1.

La trajectoire de la boucle prédite par la théorie de l'élasticité pour les boucles d'ADN passe globalement au dessus du cylindre associé à la double hélice de la tige en suivant une forme de S ou de "yin-yang" quand la trajectoire est projetée sur un plan perpendiculaire à l'axe de l'hélice. Pour les structures d'ARN, cette trajectoire passe sur le côté du cylindre, dans une trajectoire tangentielle à celui-ci. Ces différentes trajectoires s'expliquent par la différence de géométrie des hélices. Dans l'ADN, les plans des plateaux de paires de bases de la tige sont globalement orthogonaux à l'axe de l'hélice. Les tangentes aux trajectoires des squelettes de

la tige adoptent donc une symétrie pseudo-dyadique dans le plan quasi orthogonal à l'axe de l'hélice. La trajectoire de la boucle passe donc au dessus du cylindre pour joindre les deux extrémités de la tige. Dans l'ARN, les plans des plateaux de paires de bases de la tige sont inclinés par rapport à l'axe de la double hélice. Les points de tangence aux trajectoires des squelettes de la tige ne sont pas sur un plan perpendiculaire à l'axe de l'hélice. La trajectoire prédite par la théorie de l'élasticité est donc très différente, et passe sur le côté du cylindre pour joindre les deux extrémités de la tige.

III.1.2 La longueur des fils des tri- et des tétra-boucles

Qualitativement les trajectoires prédites par la théorie de l'élasticité pour les tri- et les tétra-boucles d'ADN sont très similaires. Les deux adoptent la forme en S décrite précédemment. Ceci s'explique par le fait que lors de la modélisation à l'échelle globale des épingles à cheveux d'ADN, le seul paramètre qui différencie significativement les tri- des tétra-boucles d'ADN est la longueur du fil associé à la boucle. Parallèlement les paramètres d'encastrement qui donnent les conditions aux limites pour calculer la courbe avec la théorie de l'élasticité sont identiques. En effet, les tangentes le long des squelettes sucre-phosphates et le rayon du cylindre dans lequel s'inscrit la trajectoire du squelette sont identiques

La longueur du fil dépend du nombre de nucléotides dans la séquence de la boucle et de leur conformation (C2' ou C3' endo). Nous supposons que la nature de la séquence ne modifie quasiment pas la longueur du fil, aussi peut-on considérer que les fils de toutes les tri-boucles sont de même longueur. Il en va de même pour les tétra-boucles d'ADN. Les paramètres d'encastrement du fil ne dépendent que de la géométrie de l'hélice utilisée pour modéliser la tige. Or, dans le cadre de cette étude, les tri- et tétra-boucles d'ADN sont modélisées chacunes avec une hélice régulière de même géométrie - *i.e.* la même base de données sert à générer toutes les hélices d'ADN, pour les tri- comme pour les tétra-boucles -. Dans l'approximation utilisée, les conditions d'encastrement pour les tri et les tétra-boucles d'ADN sont donc identiques.

Comme seule la longueur du fil change, et que celle-ci n'est jamais trop courte pour joindre les deux extrémités de l'hélice, les solutions données par la théorie de l'élasticité se ressemblent fortement. Il est possible de conclure que qualitativement, l'élasticité rend compte de la forme globalement identique des trajectoires de chaînes

sucre-phosphates des boucles dans les épingles à cheveux d'ADN. Ceci est dû à l'identité de la géométrie en conformation B de l'hélice qui forme la tige de la boucle.

III.2 La modélisation des structures en épingles à cheveux à partir de BCE et de la PDB

III.2.1 Nature des données structurales

Les fichiers de structure PDB sont des fichiers de coordonnées atomiques exprimées dans un repère cartésien orthonormé direct. Chaque fichier de structure contient selon les études et les auteurs, plusieurs conformations de la même molécule. Dans les fichiers PDB choisis pour cette étude, il y a entre 10 et 31 conformations différentes (*cf.* Tab. : III.1).

Fichier de structure d'ADN				
Identificateur PDB	1BJH	1XUE	1ZHU	1AC7
Nombre de conformations	16	10	10	10
Fichier de structure d'ARN				
Identificateur PDB	1AUD	1B36	1C0O	1HLX
Nombre de conformations	31	10	19	20

Tab. III.1 : ***Nombre de conformations différentes dans chacun des 8 fichiers PDB étudiés.***

Toutes ces conformations sont jugées potentiellement correctes par les auteurs. Elles correspondent à différents points de convergence lors de la modélisation par dynamique ou mécanique moléculaire. Ce sont des conformations d'énergie minimale satisfaisant les données de structures dérivées de l'expérience. Elles sont le produit de l'exploration de l'espace des conformations accessibles aux protocoles de modélisation utilisés dans chaque étude. Elles sont donc représentatives de la variabilité structurale de ces molécules et de l'imprécision ou du caractère partiel des données structurales. Afin de tenir compte de ces imprécisions lors de notre étude, nous devons prendre en compte l'ensemble des structures proposées. Se limiter à une structure par fichier introduirait un biais et une perte d'information qui ne sont pas souhaitables. Dans notre étude nous modéliserons donc chacune des conformations proposées par les auteurs, pour chaque fichier, soit 126 structures au total.

Le grand nombre de ces structures nous oblige à mettre en place un protocole de modélisation automatisé et robuste. Celui-ci doit tenir compte de la variablité de la séquence et de la nature de la chaîne sucre-phosphate (ADN ou ARN) des molécules étudiées. Il doit déplacer de façon automatique toutes ces données structurales (*cf. infra*) dans un référentiel commun pour permettre la comparaison des résultats à chaque étape de la modélisation.

III.2.2 Modélisation par déformation avec les paramètres Ω et χ et comparaison dans l'espace cartésien

Les données expérimentales des fichiers PDB sont de nature cartésienne. Ce sont les coordonnées atomiques de tous les atomes de la structure. Afin de modéliser ces conformations avec l'approche BCE, nous allons, au moyen des degrés de liberté de cette approche, chercher à optimiser le placement des atomes dans un espace cartésien en tournant les sucres et les bases avec les rotations d'angle Ω et χ pour retrouver au mieux les conformations PDB. La conformation optimale de la structure BCE_{opt} pour une conformation PDB donnée est donc déterminée par le calcul des angles Ω et χ qui minimisent, avec le calcul de RMSd, les distances entre atomes homologues de la conformation optimisées BCE_{opt} et ceux de la conformation originale PDB_{ori} (issue du fichier PDB).

Pour pouvoir calculer un RMSd qui ait un sens, nous avons vu qu'il est nécessaire de superposer préalablement les parties sur lesquelles le RMSd est calculé (*cf.* PART. : I.3.2). En effet, on ne peut comparer les coordonnées de deux molécules que si elles sont placées de façon similaire dans le repère du laboratoire. Or les structures PDB sont placées, selon les auteurs, de façon très différente dans le repère cartésien. Nous choisissons donc de déplacer toutes les structures PDB pour les placer comme les structures BCE. C'est-à-dire que l'axe de leur double hélice est confondu avec l'axe (Oz) du repère de modélisation, et le dernier plateau de paire de bases de la tige suit la convention de Cambridge.

Le positionnement des structures PDB dans le repère de modélisation est problématique car les tiges des conformations PDB ne sont pas des hélices régulières. Ce sont des hélices déformées. Le choix de l'opération de déplacement n'est donc pas une chose aisée. Pour ce faire, nous choisissons d'ajuster la structure PDB sur une structure BCE_{ori} qui, elle, est par construction correctement

positionnée et qui sert de référence. L'opération d'ajustement de toute la molécule PDB est effectuée en calculant une translation et une matrice de rotation qui minimise les distances entre atomes homologues de la structure fixe (BCE_{ori}) et de la structure déplacée (PDB). Cette matrice est appliquée à l'ensemble des atomes de la conformation PDB. La difficulté dans cette opération est donc de choisir les atomes qui vont servir à calculer les opérations de translation et de rotation. En effet, les deux structures sont très différentes et il n'est pas possible de prendre en compte tous les atomes des deux conformations. Superposer deux objets n'a de sens que s'ils sont semblables.

La structure PDB, qui est la structure témoin, est correctement structurée mais irrégulière. La structure BCE_{ori} est mal structurée, mais elle est régulière et peut être correctement placée dans un repère cartésien où l'axe (Oz) est l'axe de l'hélice. En effet, le placement des bases de la boucle de la structure BCE_{ori} n'est pas encore optimisé et pointent toutes vers l'axe de la double hélice de la tige. Les atomes de ces bases ne peuvent donc être utilisés pour optimiser le positionnement de la structure PDB, car cela reviendrait à prendre en compte des données que l'on sait être incorrectes. Le choix des atomes qui servent au calcul de superposition est alors guidé par notre objectif. Il faut placer globalement la structure de façon que les chaînes sucre-phosphates de la tige et de la boucle soient proches. Le dernier plateau de paire de bases de la tige sert à définir le repère absolu en utilisant la convention de Cambridge. Il en ressort que les atomes utilisés pour ce calcul d'ajustement sont tous les atomes de la chaîne sucre-phosphate de la partie en boucle et tous les atomes des deux derniers plateaux de paires de bases de la tige. Ce choix permet de donner un poids à peu près égal aux parties "boucle" et "tige". Si la structure PDB est déformée au niveau de la jonction tige-boucle, ce choix permettra de répartir l'erreur sur les deux portions de la molécule.

Les macromolécules comme l'ADN et l'ARN sont des objets intrinsèquement complexes, déformés et déformables. Ces caractères rendent difficiles leur étude et leur comparaison avec les structures théoriques, souvent plus régulières. Le placement des épingles à cheveux issues des fichiers PDB dans un repère commun d'analyse est donc nécessaire et utile. D'une part il permet d'unifier la procédure de construction ou de modélisation des structures théoriques BCE à partir des structures PDB dans un système de référence commun absolu, d'autre part, la comparaison de leur structure cartésienne est facilitée puisque les structures BCE et PDB sont placées de la même façon dans le repère absolu.

III.2.3 Présentation synthétique du protocole de modélisation des épingles à cheveux d'acides nucléiques avec BCE et la PDB

Les différentes étapes de la modélisation des structures en épingles à cheveux à partir de BCE et de chacune des conformations des fichiers PDB sont les suivantes :

III.2.3.1 Étape #1 : BCE$_{ori}$

Cette étape correspond à la construction d'une épingle à cheveux non-optimisée BCE$_{ori}$ à partir de la théorie de l'élasticité des barres minces. La conformation de la tige est une hélice régulière. Celle de la boucle est donnée par le repliement d'une hélice canonique sur la trajectoire prédite par la théorie de l'élasticité. Cette trajectoire est calculée à partir des conditions d'encastrement définies aux extrémités des deux brins de l'hélice régulière formant la tige. Les bases, dont l'orientation n'est pas encore optimisée, pointent vers la tige.

III.2.3.2 Étape #2 : Ajustement et placement de la conformation PDB

La structure PDB est placée dans le repère de modélisation de BCE par ajustement sur la structure BCE$_{ori}$ issue de l'étape 1. Les atomes qui servent à la superposition sont ceux que l'on considère les mieux placés dans la structure BCE$_{ori}$. Ce sont les atomes de la chaîne sucre-phosphate dont la position doit être correctement prédite par l'approche BCE et la théorie de l'élasticité, et les atomes des deux derniers plateaux de paire de bases, qui sont dans les deux structures en conformation appariée dans une hélice B ou A selon que la molécule est un ADN ou un ARN.

III.2.3.3 Étape #3 : BCE$_{opt}$

La structure BCE$_{ori}$ est optimisée pour donner la structure BCE$_{opt}$. L'optimisation consiste à rechercher la rotation Ω associée à chaque nucléotide de la boucle, et la rotation χ de chaque base de la boucle pour minimiser les distances entre atomes homologues de la structure BCE et de la structure PDB qui est prise ici comme référence.

III.2.3.4 Étape #4 : BCE_{min}

La structure BCE_{opt} est soumise à une courte minimisation d'énergie avec AMBER pour relaxer la structure et donner la structure BCE_{min}. La relaxation n'affecte que très peu la structure globale de la conformation car la structure BCE_{opt} donnée au minimiseur est très proche d'un minimum énergétique global. Elle sert à rétablir la géométrie locale des angles et des longueurs de liaison, et à tenir compte des encombrements stériques avec les potentiels de van der Waals.

L'ensemble de ce protocole est appliqué aux 126 conformations des huit épingles à cheveux étudiées. Il permet de façon "quasi-automatique" de calculer les structures et les paramètres de modélisation (Ω et χ) de chacune des structures.

III.3 Évaluations quantitatives des structures BCE obtenues

III.3.1 Quantification des déformations locales des longueurs de liaison et des angles de valence de la chaîne sucre-phosphate induites lors du repliement de l'hélice sur le fil élastique à l'étape BCE_{opt}

L'étape de repliement d'un simple brin en hélice sur la trajectoire calculée par la théorie de l'élasticité n'introduit aucune déformation de la molécule sauf au niveau des angles et longueurs de liaison entre atomes principaux de la chaîne sucre-phosphate (*cf.* PART. : II.3.1). Les courbes en pointillés de la figure III.2 montrent que les déformations induites lors du repliement et de l'optimisation (conformation BCE_{opt}) sont en général inférieures à 0,1 Å pour les longueurs de liaison, et inférieures à 10° pour les angles de valence à l'exception de la zone du "sharp-turn" dans les boucles d'ADN et en certaines positions des boucles d'ARN. Dans ces régions, les déformations induites restent respectivement inférieures à 0,25 Å et 25°. On peut remarquer que les différences longueurs de liaison oscillent entre des valeurs négatives et positives d'une part, autour de zéro, et d'autre part que ces variations sont corrélées entre les deux types de graphes. Ainsi on remarque que l'élongation d'une longueur de liaison est souvent associée à un racourcissement de la longueur

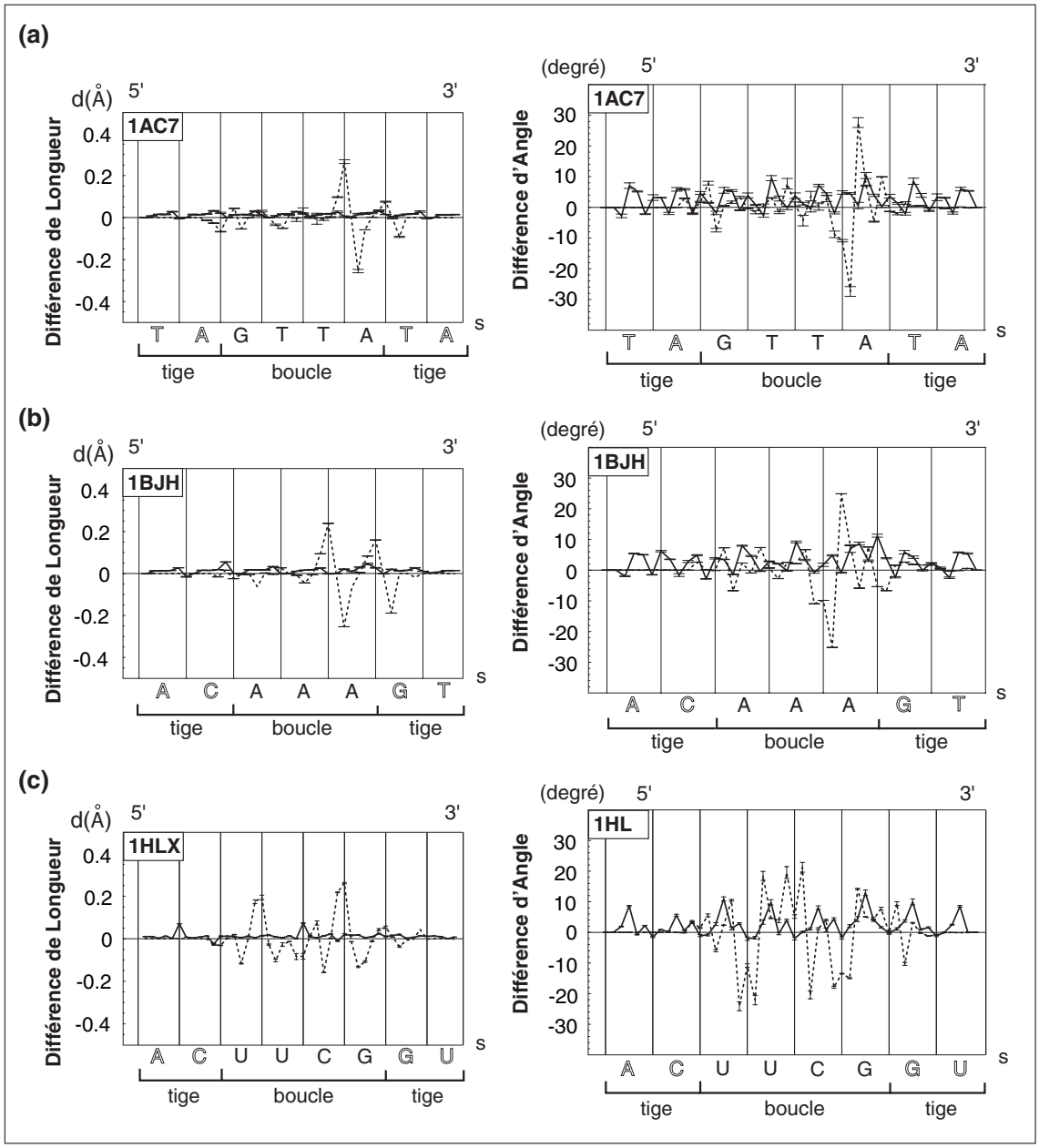


FIG. III.2 : *Écarts aux valeurs de référence des longueurs et angles de liaison de la chaîne sucre-phosphate lors du repliement de l'hélice de la boucle sur le fil élastique :* *Représentation des moyennes et des écart-types des différences de longueur de liaison en Å (à gauche) et d'angle de liaison en degrés (à droite) pour les atomes principaux de la chaîne sucre-phosphate. Les différences sont calculées entre la structure de référence en hélice régulière qui donne la norme et les boucles BCE_{opt} en traits pointillés et BCE_{min} en traits continus. Les différences sont reportées comme une fonction de l'abscisse curviligne, s. Ces abscisses curvilignes suivent l'orientation 5'→3' de la séquence, et correspondent aux projections sur le fil élastique des atomes considérés. Les trois séries de graphes sont représentatifs des trois classes de molécules étudiées : (a) la tétraboucle d'ADN 1AC7-GTTA, (b) pour les tri-boucles d'ADN, la molécule 1BJH-AAA et (c) pour les tétra-boucles d'ARN, la molécule 1HLX-UUCG.*

de la liaison suivante. Il en est de même pour les angles de valence. Les courbes en trait continu montrent que la minimisation d'énergie (conformation BCE_{min}) joue son rôle de relaxation de ces déformations induites. Les différences deviennent nulles ou presque, ce qui correspond à un retour des conformations à une géométrie canonique de type ARN-A ou ADN-B, aussi bien pour les longueurs de liaison que pour les angles de liaison.

L'approche BCE permet donc de déformer globalement les macromolécules polymériques avec de faibles déformations locales du squelette (la chaîne sucre-phosphate dans le cas des acides nucléiques). La flexion du fil en hélice et de la structure moléculaire qui lui est "attachée" sur la ligne élastique introduit alternativement, des compressions à l'intérieur de la zone courbée et des extensions à l'extérieur. Cette observation explique en partie le caractère oscillatoire des graphes de la figure III.2. Comme les déformations sont faibles et locales et qu'elles peuvent être compensées localement sans nécessiter une déformation globale, une minimisation d'énergie très courte avec AMBER restaure les angles et longueurs de liaison sans affecter de façon significative la structure globale.

III.3.2 Comparaison des chaînes sucre-phosphates BCE_{min} et PDB

Afin d'évaluer notre approche de modélisation, nous devons comparer les conformations obtenues avec BCE aux structures de références dérivées des expériences RMN et de modélisation données dans les fichiers PDB. La première hypothèse est que la théorie de l'élasticité permet de prédire la trajectoire globale de la chaîne sucre-phosphate. Cette hypothèse implique que la trajectoire de la boucle ne dépend, en première approximation, que de la géométrie des tiges en double hélice qui, selon les molécules, adopte une conformation B ou A.

Trois méthodes différentes sont utilisées pour montrer le très bon accord entre les conformations des trajectoires des chaînes sucre-phosphates des modèles théoriques BCE et des modèles PDB.

La figure III.3 donne une comparaison visuelle directe avec la superposition des chaînes sucre-phosphates des conformations PDB et des modèles BCE pour trois structures représentatives des trois classes de molécules étudiées (Tétra- et tri-boucles d'ADN et tétra-boucles d'ARN). Ces représentations montrent que les

chaînes sucre-phosphates des deux modèles oscillent en phase autour de la trajectoire du fil calculé par la théorie de l'élasticité. Nous observons un très bon acord avec les trajectoires présentées dans la figure III.1.

Identificateur PDB et séquence de la boucle	**Atomes de la chaîne sucre-phosphate**		**Atomes de la chaîne sucre-phosphate sans N_3**	
	Boucle	Tige + Boucle	Boucle	Tige + Boucle
ADN				
1AC7-GTTA	1,29 ± 0,07	1,22 ± 0,13	0,98 ± 0,06	0,98 ± 0,15
1BJH-AAA	0,91 ± 0,01	1,19 ± 0,01	0,37 ± 0,00	0,99 ± 0,01
1XUE-GCA	0,67 ± 0,00	1,22 ± 0,01	0,20 ± 0,00	1,07 ± 0,01
1ZHU-GCA	0,76 ± 0,13	1,15 ± 0,03	0,32 ± 0,05	1,05 ± 0,05
ARN				
1AUD-UUCG	1,01 ± 0,11	1,09 ± 0,23	0,92 ± 0,12	1,03 ± 0,21
1B36-UUCG	1,56 ± 0,12	1,36 ± 0,10	1,26 ± 0,13	1,05 ± 0,11
1C0O-UUCG	1,36 ± 0,04	1,24 ± 0,04	1,10 ± 0,03	0,97 ± 0,03
1HLX-UUCG	1,29 ± 0,07	1,20 ± 0,07	1,15 ± 0,05	1,10 ± 0,08

Tab. III.2 : ***Distances RMSd en Å entre atomes homologues des structures BCE_{min} et des structures PDB :*** *Les RMSd sont calculés comme la moyenne des RMSd de chacune des conformations du fichier PDB correspondant. Les valeurs reportées correspondent à la moyenne et à l'écart-type sur l'ensemble des structures du fichier PDB. Les atomes pris en compte sont les atomes principaux de la chaîne sucre-phosphate : P, O5', C5', C4',C3' et O3' dans la tige ou dans la boucle. La "Tige" est limitée aux deux derniers plateaux de paires de bases de la tige. Le nucléotide N_3 est le troisième nucluéotide dans la séquence de la boucle.*

Sur la même figure III.3, à droite, les graphes offrent une vision quantitative de cet accord. La distance au fil élastique est calculée pour les atomes des chaînes sucre-phosphates de structures BCE et PDB. La distance de 0,757 Å ± 0,46 Å représentée par un bandeau gris correspond à la distance maximale des atomes d'une hélice à un fil en hélice qui lui est ajusté. C'est la norme à l'intérieur de laquelle les atomes sont considérés parfaitement ajustés sur le fil. La distance au fil élastique des atomes des chaînes sucre-phosphates des boucles est inférieure à 1,2 Å pour les atomes de la tige et pour la plupart des atomes de la boucle, à l'exception de la zone du "sharp-turn" et dans la région de séquence UU des boucles d'ARN. Ceci montre le très bon accord des deux modèles avec la théorie de l'élasticité.

Un troisième mode de comparaison est de calculer les RMSd entre atomes homologues de la chaîne sucre-phosphate des structures BCE_{min} et PDB. Le tableau

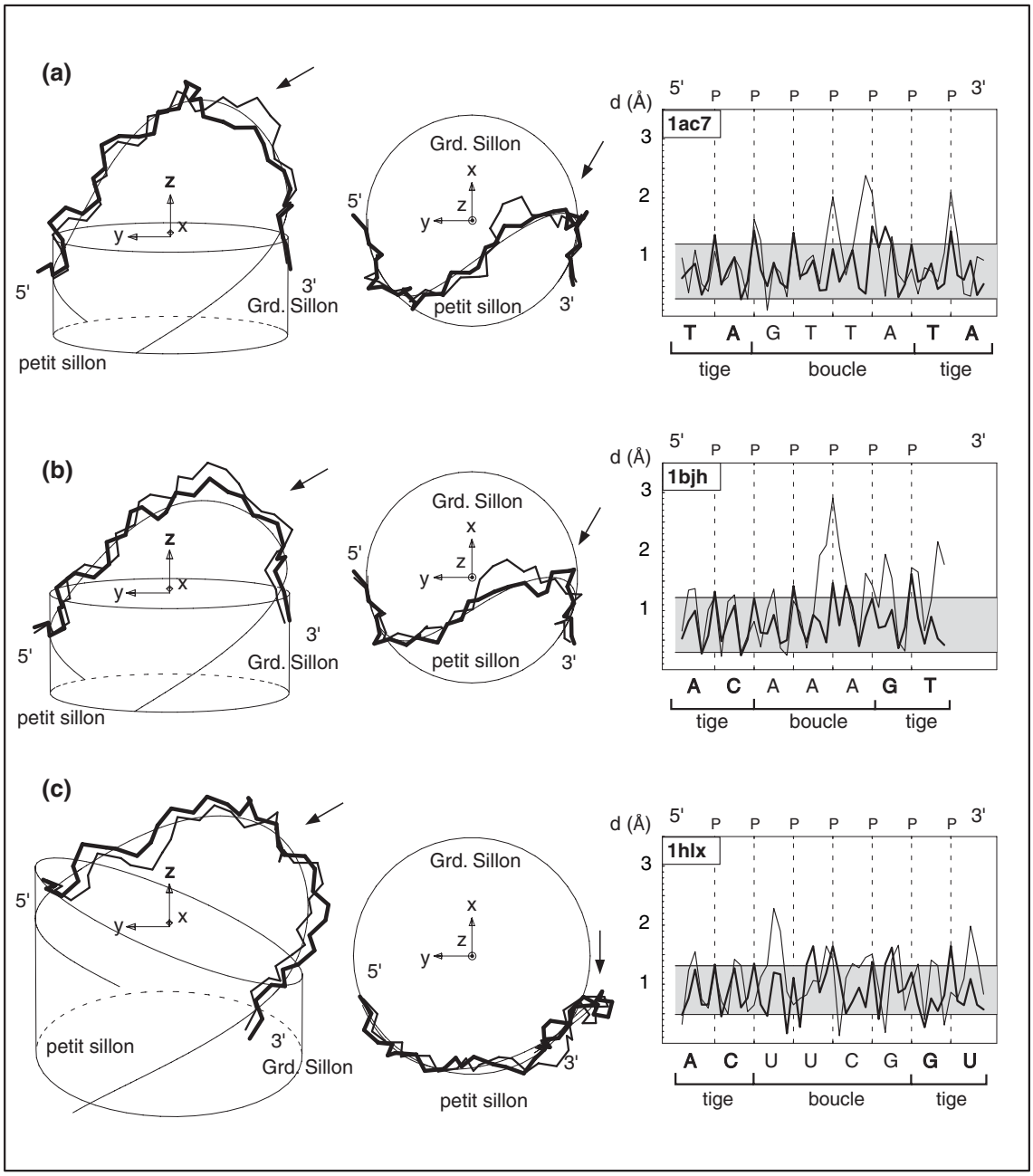


Fig. III.3 : *Comparaison des trajectoires des chaînes sucre-phosphates des modèles BCE et des conformations PDB :* *À gauche et au centre, les vues superposées de la chaîne sucre-phosphate de la structure PDB en gras, et la structure correspondante BCE_{min} et en trait fin le fil calculé par la théorie de l'élasticité pour modéliser la trajectoire de la boucle avec BCE. À gauche, les structures sont vues depuis le petit sillon. Au centre, les structures sont vues de haut le long de l'axe de la double hélice (Oz). À droite, les graphes représentent la distance d en Å des atomes de la chaîne sucre-phosphate au fil élastique pour les deux structures dans le sens 5'→3' de la séquence, en fonction de l'abcisse curviligne de la projection de l'atome sur le fil. Ces vues et ces graphes sont représentatifs des trois classes de molécules étudiées : (a) la molécule 1AC7-GTTA pour les tétra-boucles d'ADN, (b) la molécule 1BJH-AAA, pour les tri-boucles d'ADN, et la molécule 1B36-UUCG pour les tétra-boucles d'ARN. Sur les vues de gauche et du centre, les éléments graphiques sont à l'échelle et exprimés dans le même repère de référence (xyz) indiqué au centre de chaque vue. Les tailles différentes des cylindres des structures (a-b) et de la structure (c) correpondent aux rayons des hélices B inférieurs à ceux des hélices A.*

III.2 reprend les valeurs moyennes des RMSd calculés sur l'ensemble des 126 conformations étudiées. Ils sont compris entre 0,67 Å et 1,56 Å sur la partie en boucle, et entre 1,09 Å et 1,36 Å sur la tige et la boucle. L'accord entre les structures augmente fortement lorsque le nucléotide N_3, qui est le moins bien ajusté, est exclu du calcul. Les RMSd sont alors compris entre 0,20 Å et 1,26 Å sur la partie en boucle et entre 0,97 Å et 1,10 Å sur la tige et la boucle. Dans la molécule 1AC7-GTTA, le nucléotide N_3 de la boucle est le moins bien défini par la RMN. Dans la structure PDB, il pointe dans le solvant, ce qui s'explique par le faible nombre de contraintes de distances dérivées de la RMN. Le mauvais accord peut donc s'expliquer en particulier par un manque de données expérimentales. Pour les structures 1BJH-AAA, 1XUE-GCA et 1ZHU-GCA, ce manque d'accord peut suggèrer que la zone du "sharp-turn", dans laquelle se trouve le troisième nucléotide est moins bien résolue que le reste de la molécule. Pour les molécules 1AUD-UUCG, 1B36-UUCG, 1C0O-UUCG et 1HLX-UUCG, le moins bon accord du nucléotide N_3 s'explique par notre incapacité, à ce jour, de tenir compte du passage en conformation C2'-endo de ce sucre.

III.4 Modélisation de la structure globale des tri- et tétra-boucles d'ADN et des tétra-boucles d'ARN

III.4.1 Paramètres de construction Ω et χ

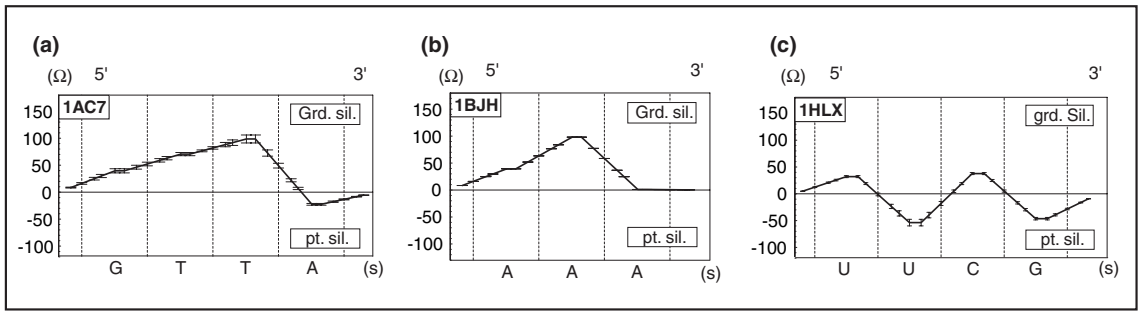
Pour chaque conformation de chaque fichier PDB une conformation BCE_{opt} est calculée de façon automatique. Les paramètres de construction quantitatifs associés aux degrés de liberté Ω et χ sont donc accessibles pour chacune de ces structures. Les valeurs moyennes de ces paramètres sont reportées dans le tableau III.3. Elles montrent une grande homogénéité à l'intérieur d'une même classe de molécules. L'homogénéité de ces valeurs traduit le mode de structuration similaire de toutes les tri-boucles d'ADN et des tétra-boucles d'ARN. Ces valeurs numériques sont associées à la rotation des blocs d'atomes des sucres autour du fil élastique. Elles peuvent être représentées graphiquement sous forme de profils donnant la valeur dont est tourné chaque bloc de la boucle en fonction de l'abscisse curviligne du bloc considéré (*cf.* FIG. : III.4). Ces profils montrent la similarité des modes de structuration à l'intérieur des familles des tri- et tétra-boucles étudiées et révèlent

Identificateur PDB	Ω_1	Ω_3	Ω_3	Ω_4	$\Delta\chi_1$	$\Delta\chi_2$	$\Delta\chi_3$	$\Delta\chi_4$
ADN								
1AC7-GTTA	**39,6**	**70,6**	**98,6**	**-22,6**	**16,7**	**19,9**	**26,7**	**45,4**
	±3,9	±2,9	±7,4	±1,6	±3,2	±3,4	±2,6	±2,8
1BJH-AAA	**39,3**	**98,2**	**1,1**	/	**3,3**	**31,3**	**73,3**	/
	±0,2	±0,1	±0,1	/	±0,3	±0,6	±0,1	/
1XUE-GCA	**30,9**	**98,6**	**-0,5**	/	**-5,0**	**-11,5**	**77,3**	/
	±0,0	±0,0	±0,0	/	±0,0	±0,0	±0,1	/
1ZHU-GCA	**32,6**	**89,7**	**0,5**	/	**-0,7**	**-5,9**	**76,9**	/
	±0,3	±2,2	±0,6	/	±1,1	±0,7	±0,2	/
ARN								
1AUD-UUCG	**29,9**	**-92,7**	**41,1**	**-40,3**	**1,9**	**-14,7**	**-24,6**	**-117,1**
	±33,1	±36,6	±14,7	±7,5	±6,2	±22,3	±7,9	±3,4
1B36-UUCG	**33,5**	**-57,4**	**43,2**	**-49,2**	**31,8**	**15,2**	**-52,9**	**-109,8**
	±3,8	±6,1	±3,4	±4,3	±5,6	±8,1	±7,9	±4,0
1C0O-UUCG	**32,4**	**-63,6**	**39,7**	**-44,1**	**21,1**	**5,5**	**-16,8**	**-111,9**
	±1,7	±3,2	±1,2	±1,8	±2,6	±3,8	±1,6	±1,0
1HLX-UUCG	**32,0**	**-53,7**	**37,7**	**-46,4**	**12,3**	**12,3**	**-17,6**	**-115,7**
	±1,9	±6,1	±1,9	±2,3	±6,0	±5,9	±3,9	±4,4

Tab. III.3 : ***Valeurs moyennes des angles*** Ω ***et*** χ ***calculées lors de la modélisation des structures*** BCE_{opt} ***:*** *Les valeurs de* Ω_i *et* χ_i *sont calculées en prenant comme structure de référence* BCE_{ori} *et exprimées en degrés. Les angles* $\Delta\chi_i$ *sont donc égaux à la différence de la valeur entre l'angle* χ *dans une hélice régulière et la valeur du même angle dans la conformation* BCE_{opt}*.*

des régularités de profils différents. Du côté 5' de la boucle, à la jonction avec le brin I de la tige, les valeurs de Ω sont proches de zéro. Elles atteignent linéairement un maximum au niveau de l'avant dernier nucléotide de la boucle, avec une valeur de Ω similaire compris entre 89,7° et 98,6°. Cette similarité dans les profils de reconstruction est à mettre en parallèle avec les similarités de positionnement des bases des boucles de ces épingles à cheveux. Nous avons déjà vu dans la section I.4.1 que les tri- et tétra-boucles d'ADN se structurent de la même façon. Les bases du côté 5' de la boucle s'empilent les unes sur les autres du côté du grand sillon. Les profils montrent la même chose de façon quantitative, car des valeurs positives de Ω orientent les bases de la boucle vers le grand sillon.

Les profils des tétra-boucles d'ARN UUCG, sont très différents. Ces différences tiennent à deux facteurs. D'une part, la géométrie de l'hélice en forme A des structures d'ARN, et celle de l'hélice B dans les structures d'ADN, définissent des espaces conformationnels différents accessibles par des rotations Ω. Il faut donc



Tab. III.4 : ***Profils de construction*** Ω***=f(s) :*** *Rotation des blocs de la chaîne sucre-phosphate de la boucle en fonction de l'abscisse curviligne de la projection de l'atome pivot du bloc. Les profils correspondent à la modélisation de la première conformation des fichiers PDB de trois structures représentatives des différentes classes de molécules étudiées. (a) La tétra-boucle d'ADN 1AC7-GTTA, (b) La tri-boucle d'ADN 1BJH-AAA et (c) la tétra-boucle d'ARN 1HLX-UUCG. Les valeurs positives de* Ω *tendent à tourner les blocs du côté du grand sillon de la tige, et les valeurs négatives, vers le petit sillon.*

s'attendre à ce que les valeurs de Ω qui permettent d'empiler les bases soient différentes. D'autre part, la géométrie des empilements des bases des boucles UUCG est très différente des géométries des boucles d'ADN. Dans ces structures, la deuxième base de la boucle se place dans le petit sillon de la tige, et la troisième tend à s'empiler sur l'appariement formé entre les bases extrémales de la boucle U1$\cdots$G4. Cette géométrie différente se traduit par des profils de construction alternés, où la première base est légèrement tournée vers le grand sillon pour s'empiler sur le dernier plateau de paire de bases de la tige, la deuxième est tournée vers le petit sillon pour s'y placer, la troisième est tournée vers le grand sillon pour s'empiler sur le plateau de paire de bases formé dans le boucle, et le dernier nucléotide est tourné vers le petit sillon pour former l'appariement. L'analyse des valeurs de Ω est donc en accord avec les données qualitatives du tableau I.2. Ce degré de liberté est donc un paramètre qui permet d'évaluer quantitativement de façon conceptuellement simple le mode de structuration des bases dans les boucles d'acides nucléiques.

Les angles de rotation, Ω, autour de la tangente à la trajectoire utilisés dans BCE sont un outil de description compact du mode de structuration des molécules. Ils permettent également de mettre en relief certaines propriétés. Les nucléotides de la boucle semblent littéralement "tomber en place de façon naturelle" par les rotations d'angle Ω, dans les positions correctes données par les conformations PDB. En particulier, de simples et faibles rotations du premier et du dernier nucléotide autour du fil élastique suffisent pratiquement à mettre en place l'appariement dans la boucle. Cela suggère que l'appariement G$\cdots$A des boucles -GTTA- et -GCA-, l'appariement

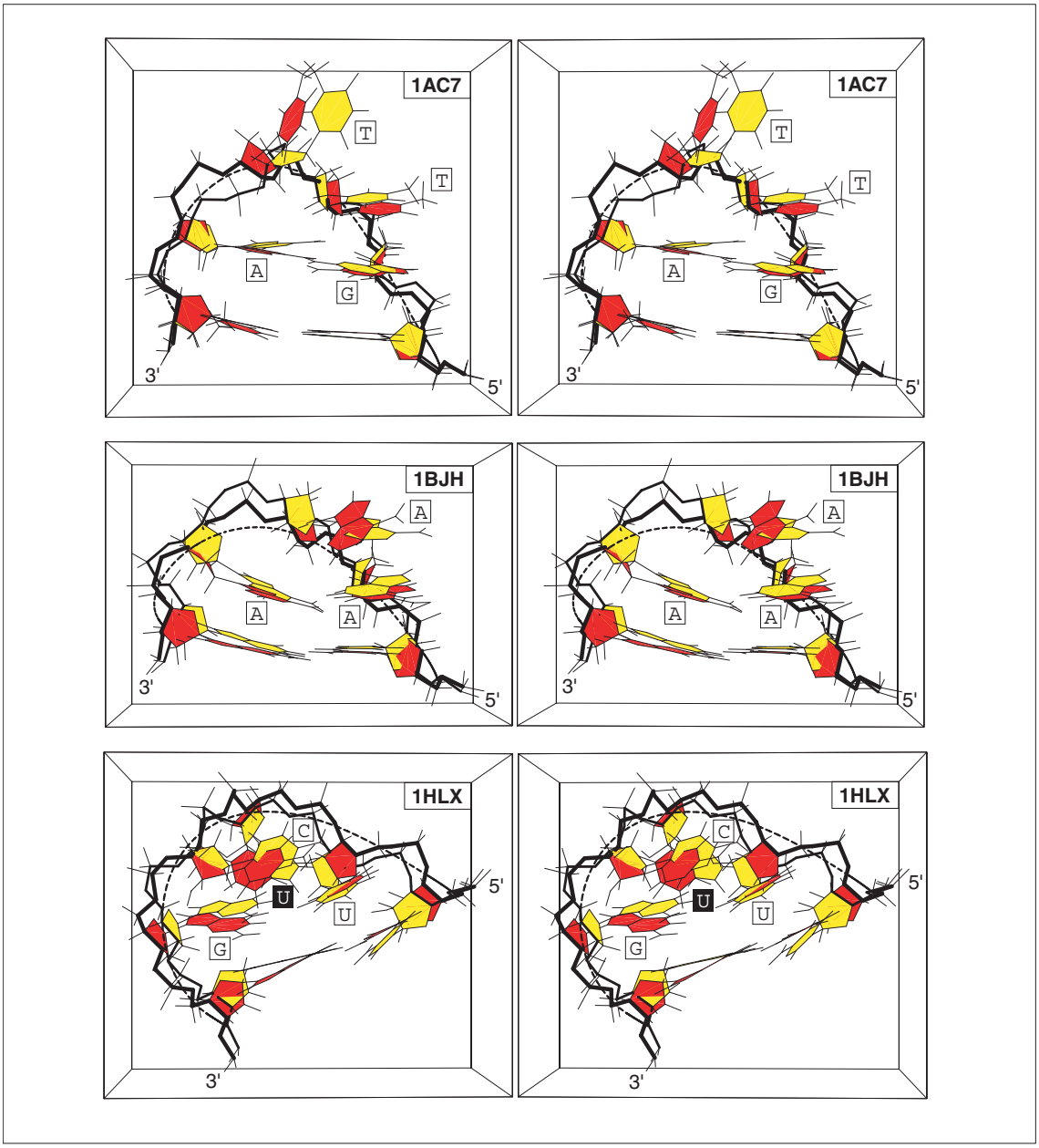


Fig. III.4 : ***Vues stéreoscopiques de la superposition des conformations BCE_{min} et PDB*** *de trois conformations représentatives des différentes classes de molécules étudiées : les bases sont en rouge et la chaîne sucre-phosphate en gras pour la structure de référence PDB; en jaune, la structure modélisée BCE_{min}; en noir et en trait léger, les courbes associées aux hélices de la tige et le fil élastique de la boucle; en haut la conformation 1AC7-GTTA, au milieu, la conformation 1BJH-AAA, et en bas la conformation 1HLX-UUCG.*

A···A dans la boucle -AAA- et l'appariement U···G dans la boucle -UUCG- ne devraient plus être considérés comme des mésappariements, mais plutôt comme le meilleur appariement possible étant donnée la trajectoire adoptée par la chaîne sucre-

phophate à cet endroit. Jusqu'à maintenant, ces mésappariements étaient considérés comme des facteurs majeurs contribuant à la stabilisation de la boucle par la mise en place d'interactions stabilisantes de type liaison hydrogène ou d'empilement. Avec l'étude de ces structures et des structures -TTT- où il n'y pas d'appariement dans la boucle [4, 80] , nous montrons que la chaîne sucre-phosphate adopte "d'abord" une trajectoire de type élastique, et qu'ensuite, selon la nature des bases de la boucle, et selon la géométrie données par la théorie de l'élasticité, des appariements peuvent, ou non, se mettre en place. Ces appariements doivent donc être vus comme un facteur additionnel de stabilité qui peut se mettre en place si la trajectoire BCE du squelette le permet.

III.4.2 Comparaison des structures BCE_{min} et PDB

Plusieurs méthodes sont employées pour évaluer l'accord entre les structures modélisées BCE_{min} et les structures de référence PDB. La première consiste à comparer visuellement les structures superposées (*cf.* FIG. : III.4). La seconde consiste à calculer des RMSd entre les structures atomiques (*cf.* TAB. : III.5).

Identificateur PDB	**Tous les atomes**		**Tous les atomes sans N_3**	
	Boucle	Tige + Boucle	Boucle	Tige + Boucle
ADN				
1AC7-GTTA	1,62 ± 0,10	1,35 ± 0,10	0,97 ± 0,09	0,92 ± 0,11
1BJH-AAA	1,57 ± 0,00	1,38 ± 0,01	1,27 ± 0,01	1,24 ± 0,01
1XUE-GAA	1,27 ± 0,00	1,34 ± 0,01	0,97 ± 0,00	1,22 ± 0,01
1ZHU-GAA	1,31 ± 0,05	1,32 ± 0,02	1,08 ± 0,03	1,21 ± 0,03
ARN				
1AUD-UUCG	1,94 ± 0,24	1,70 ± 0,19	1,74 ± 0,29	1,55 ± 0,18
1B36-UUCG	2,11 ± 0,10	1,73 ± 0,09	1,54 ± 0,12	1,28 ± 0,10
1C0O-UUCG	1,92 ± 0,04	1,64 ± 0,04	1,43 ± 0,03	1,26 ± 0,04
1HLX-UUCG	1,89 ± 0,08	1,55 ± 0,08	1,62 ± 0,11	1,37 ± 0,11

TAB. III.5 : *Accord entre tous les atomes des structures BCE_{min} et les structures PDB :* *Les RMSd sont calculés pour chaque conformation du fichier PDB. Les valeurs reportées correspondent à la moyenne et à l'écart-type sur l'ensemble des structures du fichier PDB. Les atomes pris en compte sont tous les atomes de la chaîne sucre-phosphate, des sucres et des bases. La "Tige" inclut les deux derniers plateaux de paires de bases de la tige. Le nucléotide N_3 est le troisième nucléotide dans la séquence de la boucle.*

Les RMSd montrent le très bon accord des conformations modélisées avec l'approche BCE et des structures de référence PDB. Ils sont compris entre 1,27 Å et 2,11 Å sur la boucle, et entre 1,32 Å et 1,73 Å sur la structure globale. Comme pour la chaîne sucre-phosphate, l'accord augmente lorsque le troisième nucléotide de la boucle est exclu du calcul. Les RMSd sont alors compris entre 0,97 Å et 1,74 Å sur la boucle et entre 0,92 Å et 1,55 Å sur la tige et la boucle. Ces accords sont très bons au regard de la résolution des structures résolues à partir de données dérivées de la RMN qui est comprise entre 1 Å et 1,5 Å [95]. Ces accords sont remarquables étant donnée la démarche *a priori* que nous avons utilisée et le peu de degrés de liberté (Ω et χ) à notre disposition.

III.5 Conclusion

La déformation d'une chaîne macromoléculaire de quelques nucléotides comme une barre mince soumise à la théorie de l'élasticité est une approche conceptuellement très simple du repliement des chaînes d'ADN et d'ARN. Avec cette idée simple, nous avons montré : D'une part qu'il est possible de déformer un simple brin en hélice d'ADN pour retrouver les boucles des épingles à cheveux d'ADN dont la longueur varie de trois à quatre nucléotides. D'autre part, que cette même approche permet de retrouver les structures en boucle des épingles à cheveux d'ARN de séquence UUCG à partir d'un simple brin en hélice d'ARN. La trajectoire globale des chaînes sucre-phosphates de ces molécules semble donc être prédictible *a priori* si l'on connaît les directions aux extrémités de la chaîne avec la théorie de l'élasticité appliquée à la flexion des barres minces.

Les formes très différentes des épingles à cheveux d'ADN et d'ARN sont très bien reproduites par une approche identique et un formalisme commun. Les formes différentes des boucles sont dues aux conditions aux extrémités différentes utilisées pour calculer la trajectoire du fil élastique de la boucle. Ces dernières sont simplement une conséquence de la différence des géométries des hélices d'ADN et d'ARN qui forment la partie en tige des épingles à cheveux. Ces résultats tendent à démontrer que la chaîne sucre-phosphate, bien que de nature discontinue, car formée d'un enchaînement discret d'atomes, adopte un comportement de fil flexible à une échelle globale de plusieurs nucléotides. Ce comportement élastique semble jouer un rôle prépondérant dans la forme adoptée par les chaînes sucre-phosphates des boucles d'ADN et d'ARN. Jusqu'à aujourd'hui plusieurs types d'interactions ont été

invoqués pour expliquer la stabilité exceptionnelle des épingles à cheveux étudiées : les liaisons hydrogène, les interactions d'empilement, les interactions hydrophobes. Nos résultats montrent que la chaîne sucre-phosphate suit une trajectoire de moindre déformation et de moindre énergie où les angles de torsion conservent des valeurs proches des valeurs intiales rencontrées dans les hélices d'ADN-B ou d'ARN-A. Cela suggère que pour ces molécules, les propriétés élastiques de la chaîne sucre-phosphate jouent un rôle structural et énergétique qui contribue de façon importante à la stabilité extraordinaire de certaines épingles à cheveux au moins au même titre que les autres interactions.

Suivant la description habituelle, les épingles à cheveux d'acides nucléiques sont formées d'une tige en double hélice appariée fermée par une boucle simple brin de nucléotides non- ou més-appariés. En accord avec nos résultats, ces mésappariements (G$\cdots$A dans les tri- et tétra-boucles d'ADN, A$\cdots$A dans les tri-boucles d'ADN et U$\cdots$G dans les tétra-boucles d'ARN), devraient plutôt être considérés comme faisant parte de la boucle et comme les meilleurs appariements possibles satisfaisant les conditions géométriques imposées par la forme élastique BCE de la chaîne sucre-phosphate. Suivant le même raisonnement, les appariements Watson-Crick ne sont pas les meilleurs appariements possibles étant donnée la trajectoire BCE de la boucle, mais doivent être les meilleurs appariements possibles étant donnée la forme hélicoïdale de la chaîne sucre-phosphate dans les hélices.

Enfin, le nouveau degré de liberté et paramètre de modélisation, l'angle Ω, est un paramètre quantitatif cohérent de description des boucles comportant un appariement. Il permet d'unifier la description au moyen d'un formalisme commun et quantitatif les modes de structuration des nucléotides des chaînes d'ADN comme des chaînes d'ARN. Il permet de décrire et de modéliser les structures à l'échelle mésoscopique des nucléotides au moyen d'un nombre limité de degrés de liberté.

III.6 ARTICLE

1086–1096 *Nucleic Acids Research, 2003, Vol. 31, No. 3*
DOI: 10.1093/nar/gkg196

DNA tri- and tetra-loops and RNA tetra-loops hairpins fold as elastic biopolymer chains in agreement with PDB coordinates

Guillaume P. H. Santini, Christophe Pakleza and Jean A. H. Cognet*

Laboratoire de Physico-chimie Biomoléculaire et Cellulaire, UMR 7033 CNRS, T22-12, Université Pierre et Marie Curie, 4 place Jussieu, 75252 Paris cedex 05, France

Received July 24, 2002; Revised November 7, 2002; Accepted November 23, 2002

ABSTRACT

The biopolymer chain elasticity (BCE) approach and the new molecular modelling methodology presented previously are used to predict the tri-dimensional backbones of DNA and RNA hairpin loops. The structures of eight remarkably stable DNA or RNA hairpin molecules closed by a mispair, recently determined in solution by NMR and deposited in the PDB, are shown to verify the predicted trajectories by an analysis automated for large numbers of PDB conformations. They encompass: one DNA tetraloop, -GTTA-; three DNA triloops, -AAA- or -GCA-; and four RNA tetraloops, -UUCG-. Folding generates no distortions and bond lengths and bond angles of main atoms of the sugar–phosphate backbone are well restored upon energy refinement. Three different methods (superpositions, distance of main chain atoms to the elastic line and RMSd) are used to show a very good agreement between the trajectories of sugar–phosphate backbones and between entire molecules of theoretical models and of PDB conformations. The geometry of end conditions imposed by the stem is sufficient to dictate the different characteristic DNA or RNA folding shapes. The reduced angular space, consisting of the new parameter, angle Ω, together with the χ angle offers a simple, coherent and quantitative description of hairpin loops.

INTRODUCTION

In the preceding article in this issue, we have postulated that the backbone of single-stranded DNA hairpin loops behaves as a continuous, inextensible and flexible thin rod. With this simple hypothesis, the tri-dimensional trajectory of this elastic line was derived from the theory of elasticity and we have shown how it can be used to predict the structures of the sugar–phosphate backbone of DNA hairpins. We have shown also how single-stranded trinucleotide B-DNA TTT could be folded into hairpin loops, G-TTT-C or C-TTT-G, where most torsion angles are preserved, to match four different sets of NMR data or five different molecular conformations (1–4). In this approach, called biopolymer chain elasticity (BCE), the trajectories of the two helical backbones of the hairpin stem define the geometry of the extremities of the hairpin loop. In the theory of elasticity of thin rods, the geometry of end conditions dictates the shape of the trajectories. Therefore the different shapes of DNA and RNA hairpin loops should be predicted or should result from the different geometries imposed by the stem structures. Double helical B-DNA and A-RNA differ in two respects. Firstly the planes of base pairs and of helical extremities are perpendicular to the helical axis in B-DNA whereas they are tilted in A-RNA. Secondly B-DNA helix has a smaller radius. As shown in Figure 1, when the backbones trajectories of the loops (in dark red or blue) are in perfect continuity with the backbones trajectories of the double helix (in light colour), we predict that the backbone of the DNA tetraloop must go over the surface of the DNA cylinder, whereas the trajectory of RNA backbone is circumscribed on or slightly outside the RNA cylinder as shown in Figure 1. In this article, we investigate whether the BCE methodology and its tri-dimensional predictions that were presented previously for DNA triloops (5) and build on previous ideas (6,7) can be applied not only to other hairpins of different structures, of different lengths, tri- and tetra-loops of DNA, but also to RNA tetraloops.

To this end, we have selected eight different molecules of which structures have been recently determined in solution by NMR and deposited in the Protein Data Bank (PDB) (8). They encompass: one DNA tetraloop, -GTTA-; three DNA triloops, -AAA-, -GCA-, -GCA-; and four RNA tetraloops, -UUCG-, as presented in Table 1. The DNA triloops and RNA tetraloops have been the subject of many determinations over the last decade and have given rise to well defined solution structures since early NMR studies (9,10). They have been used as test systems for theoretical studies (11–13). For brevity and convenience, the molecules are referred by their PDB identifications. These eight molecules have several features in common. They are all remarkably stable (7,14) and the loop is closed by a side-by-side sheared mispair (15,16) G·A (1ac7), A·A (1bjh), G·A (1xue, 1zhu), or by a head to side U·G mispair (17) (1aud, 1b36, 1c0o, 1hlx). Note that the DNA triloops studied here are structurally different from the TTT hairpins studied previously (5). In the latter, the first and last

*To whom correspondence should be addressed. Tel: +33 1 44 27 27 50; Fax: +33 1 44 27 75 60; Email: cognet@ccr.jussieu.fr

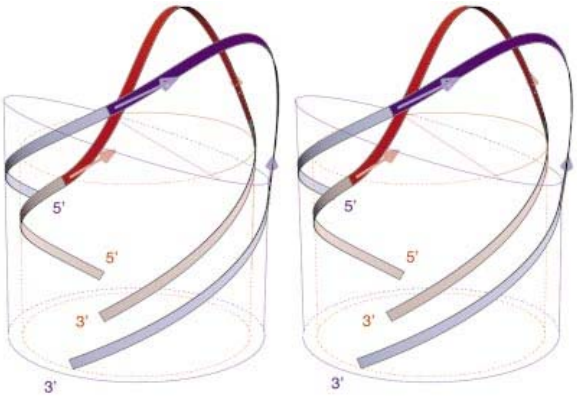


Figure 1. Superimposed stereo views into the minor groove of the computed elastic rod curve for a DNA tetraloop hairpin, shown in red, and for an RNA tetraloop hairpin shown in blue. The radii of cylinders and circles are those of the sugar–phosphate backbones, DNA (red) and RNA (blue). The mean planes of base pairs are indicated by the top sections of the cylinders.

nucleoside of the loop were not stacked on the top base pair of the stem. They never formed a mispair. They were located in the minor groove, in the major groove or in the solvent and did not interact with each other.

Well defined DNA and RNA tetraloops come essentially in three different folds (7,18). Type-I loops are only observed for DNA. They adopt a conformation with the first three bases at the 5′-end of the loop forming a more or less continuous stack on the 3′-end of the stem (1ac7) (15). Type-II loops are found in both DNA (CTTG) (18) and RNA (1aud, 1b36, 1c0o, 1hlx). As indicated in Table 1, Nb is turned into or towards the minor groove and Nc lies over the closing base pair NaNd. A third fold, type-III, is only observed in RNA and is described by a continuous stacking of Nd, Nc and Nb on the 5′-end of the stem.

DNA hairpins can perform many important and diverse biological functions as recently established by numerous experiments and as briefly reviewed in the previous article in this issue. The DNA tetraloop -GTTA- (1ac7) is related to telomeric and centromeric structures (15). The DNA triloops -AAA- (1bjh) and -TTT- are important components of the adenoassociated virus 2 (4,5). The two DNA triloops -GCA- are encountered in human centromere repeats (1xue) and in centromeric GNA triplets (1zhu) and are important to account for the observed expansion of triplet repeats (5). RNA hairpins have been known to play essential structural and biological roles for several decades (14,19). In particular, the hairpin contained in 1aud is part of the polyadenylation inhibition element bound to the RNP domain of the human U1A protein (20). In 1b36, the -UUCG- tetraloop was added to stabilise the structure of one of the two domains essential for catalysis in a ribozyme molecule (21). Similarly in 1c0o, the stable -UUCG- loop was added to close an RNA metal hexammine binding site from the P5 helix of the catalytic core of the *Tetrahymena* group I intron ribozyme (22). In 1hlx, the tetraloop is the capping part of the P1 helix from group I self-splicing introns (17).

Advances in synthetic and spectroscopic techniques have recently extended the size and the accuracy of RNA molecular structures that can now be solved by NMR (20). The solution structures retained here for analysis were determined between 1995 and 1999, from large collections of NMR data. Due to their sizes, to the different complex protocols used, and to the rapid evolution of computer programs, the complete data may be only partly available and it may be difficult to analyse in an identical way to that of the original authors and as we have done in the previous article in this issue. For these reasons, the theoretical molecular structures built with the BCE approach were not compared with NMR-derived distances and to a single molecular structure as previously described (5), but directly to available PDB solution structures. Note that the eight corresponding PDB files contain many molecular conformations, in numbers from 10 up to 31 per coordinates file (Table 1), because solution structures were derived from NMR. Their total number is 126. Due to this large number, the computer program introduced previously, S-mol (5), was

Table 1. Molecular structures selected from the PDB, with PDB identification, number of structures in square brackets and tetraloop folding type (7,18), original authors, DNA or RNA sequences of PDB structures and of theoretical BCE models, and locations of the loop bases in the Major (M) groove, in the minor (m) groove, stacked in the central part of the helix (c) or in the solvent outside the structure (solv.)

PDB identification	Authors	DNA or RNA sequence solved experimentally & sequence used in theoretical models	Na	Nb	Nc	Nd
DNA						
1ac7 [10] Type I	van Dongen *et al.* 1997 (15)	d(...cta-GTTA-tagg...) & d(gcta-GTTA-tagc)	c/M	M	M / solv	c/m
1bjh [16]	Chou *et al.* 1996 (16)	d(gtac-AAA-gtac) & d(gcac-AAA-gtgc)	c/M	M	c	-
1xue [10]	Zhu *et al.* 1996 (30)	d(...gaat-GCA-atgg...) & d(gcat-GCA-atgc)	c/M	M	c	-
1zhu [10]	Zhu *et al.* 1995 (38)	d(caat-GCA-atg) & d(gcat-GCA-atgc)	c/M	M	c	-
RNA						
1aud [31] Type II	Allain *et al.* 1997 (20)	r(...gucc-UUCG-ggac...) & r(gccc-UUCG-gggc)	c/M	m / solv	M	c
1b36 [10] Type II	Butcher *et al.* 1999 (21)	r(...gcgc-UUCG-gcgc...) & r(gcgc-UUCG-gcgc)	c/M	m / solv	M	c
1c0o [19] Type II	Colmenarejo and Tinoco 1999 (22)	r(...gguc-UUCG-gguc...) & r(gcuc-UUCG-gggc)	c/M	m / solv	M	c
1hlx [20] Type II	Allain and Varani 1995 (17)	r(...uaac-UUCG-guug...) & r(gcac-UUCG-gugc)	c/M	m / solv	M	c

Absence of fourth nucleotide in the loop is denoted (-). The loop bases are marked Na, Nb, Nc and Nd in the 5′ to 3′ direction.

enhanced to deal with automatic comparisons. This is an important change that required specific modifications of the BCE methodology as explained in Materials and Methods and below.

In this article, our main focus is to search a general theoretical approach, which is capable of: (i) predicting a priori the tri-dimensional course of the sugar–phosphate chains, not only of DNA hairpin molecules structurally different from TTT hairpins, but also of RNA hairpins, (ii) generating models close to solution structures from these predictions and from large numbers of given PDB conformations, and (iii) characterising the importance of the sugar–phosphate chain and of its elastic properties in the folding process.

MATERIALS AND METHODS

Original molecular structures, PDBid

Original molecular conformations are from the PDB (8) and are referred by their PDBid: 1ac7, 1b36, 1xue, 1zhu, 1aud, 1bjh, 1c0o and 1hlx.

Initial stem and loop model building by molecular mechanics

All initial structures were generated from canonical B-DNA or A-RNA (23).

Theoretical molecular structures, BCE

A registered software Smol© (5) was extended under UNIX and Linux environments using Mathematica (24), Geomview (25) and C languages to build and to compare BCE models with solution conformations of PDB files.

The complete DNA or RNA sequences of the theoretical molecular structures, given in Table 1, were simplified with the two following rules. The sequence of the loop and of the first two base pairs in the stem is identical to original PDB molecular structures. The length of stems is reduced to four base pairs and the remaining sequence of the stem is set to d(GC).d(GC) or r(GC).r(GC). Note that all PDB conformations proposed under a given PDB identification were used for building the theoretical structures. The length, L, of the capping rod was obtained as previously described (5) by fitting a helical line to the atoms of the main sugar–phosphate backbone (O5′, C5′, C4′, C3′, O3′, P) of a single-stranded helical A-RNA or B-DNA and by minimising the root-mean-square of the sum of squared distances to the helical line. For A-RNA radius of helical line was 9.35 Å and its pitch was 30.85 Å/turn. For B-DNA, values were respectively 8.35 Å and 33.74 Å/turn. Molecules were folded into hairpin loops using prescribed geometric boundary conditions.

Setting all PDB conformations in the laboratory reference frame

PDB conformations are moved onto BCE molecular models by a translation-rotation coordinate transformation.

$$\vec{OM}_{\text{in the global reference frame}} = R_{\text{rotation matrix}} \cdot \vec{OM}_{\text{in local reference frame}} + \vec{V}_{\text{vector of translation}}$$

Optimised molecular structures, 'BCE3Ωopt'

Molecular structures provided in PDB files and summarised in Table 1 were used at the third step of theoretical molecular modelling to optimise the rotation angles about the elastic line, Ω, and the glycosidic torsion angles, χ, of each nucleoside in the loop independently from other nucleosides. This was performed by a least square fit on homologous atom positions to give optimised BCE models, BCE3Ωopt.

Final theoretical molecular structures, 'BCE4finalm'

BCE molecular models were energy refined without restraints. Energy refinements were carried out with the program AMBER (5,26,27) without any restraints and with a large stopping root-mean-square energy gradient criterion 0.5 kcal/(mol.Å) to yield final molecular models, BCE4finalm.

RMSd analysis

RMSd are computed after superposing the two sets of matching atoms by a translation-rotation coordinate transformation.

RESULTS

The BCE approach enhanced to treat multiple PDB conformations

Folding a DNA or an RNA hairpin loop with the BCE approach can be described as a three-step procedure, completed by a short energy refinement step to restore backbone bond lengths and bond angles (5). A short and intuitive account of the procedure modified to treat multiple PDB conformations is given below and in Figure 2. (i) Single-stranded A-RNA or B-DNA are basically considered as a continuous and flexible thin rod in the following practical manner. These polymers are generated along helical lines, which are also viewed as elastic lines. The main atoms of the sugar–phosphate backbone (O5′, C5′, C4′, C3′, O3′, P) play a key role because they are attached to this line and because they are used to define the origins of local reference frames for all remaining atoms in the nucleotide. As a result, there are six different groups of atoms per nucleotide. The polymer may thus be viewed as a succession of individual solid blocks of atoms attached to the elastic line. Using this basic framework where all backbone atoms are made part of the elastic line as shown for A-RNA in Figure 2A, the biopolymer chain can be bent and twisted smoothly using elasticity theory of thin rods into a given loop with prescribed end conditions (Fig. 2B and C). This step yields a elastic curve, BCE1curve (Fig. 2D), which can be fitted onto the double helical stem (Fig. 2C). Note that the tri-dimensional trajectory of the elastic line is uniquely determined for end conditions of Figure 2B and C. (ii) Transportation of the biopolymer chain onto the elastic line step yields a molecular model, BCE2basicxyz (Fig. 2A, D and E). Crucial parameters are the length of the loop, tri- or tetra-loop, and the geometry of end conditions imposed by the A-RNA or B-DNA helices. (iii) A useful feature provided by this formalism is that each block of atoms, and consequently an entire nucleoside, can be rotated about the elastic line with an angle, Ω. Each nucleoside block can be rotated independently to match NMR-derived distances as in the previous

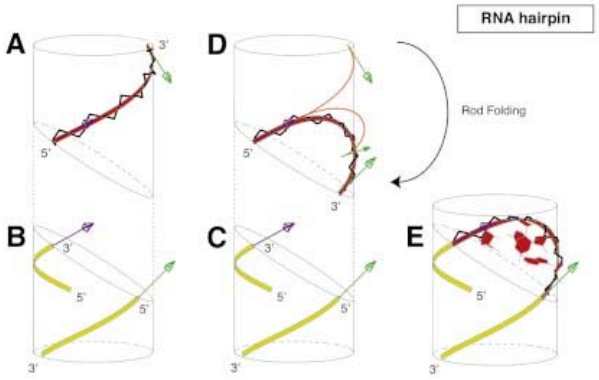


Figure 2. Schematic overview of the construction process of an RNA tetraloop hairpin molecular structure using the BCE approach proposed in the previous article in this issue (5). (**A**) A continuous and flexible thin rod, represented by a ribbon for better visibility, and shown in red is associated to a four nucleotides helical segment; (**B** and **C**) two helices of RNA are generated along helical lines, shown in yellow; (**D**) the flexible rod is bent smoothly into the capping elastic solution curve so that the tangents at its extremities, shown as blue and green arrows, match those of the two helices; helical segments and the capping rod are dissociated for clarity; (**E**) they are shown fully assembled; the complete molecular structure is the basic BCE molecular model and is computed after global deformation by keeping track of the translations and the rotations required to leave unchanged atoms in their nearest local reference frames along the helical segment.

article in this issue, or in order to match each one of the molecular conformations given in a PDB file as in this study. This step is defined here as an automated optimisation of angles, Ω, and of glycosidic torsion angles, χ, and yields optimised BCE molecular models, BCE3Ωopt. (iv) Individual molecular blocks are displaced by the folding procedure without internal deformation. However the chemical bonds and bond angles of the main atoms of the sugar–phosphate backbone (O5′, C5′, C4′, C3′, O3′, P) are modified by the BCE folding procedure. This is why each molecular structure is very shortly energy refined without restraints to restore backbone bond length and bond angles. This step yields the final theoretical molecular model, BCE4finalm.

In this article, we compare the large number of original molecular conformations supplied by a PDB file to their corresponding BCE4finalm molecular models. As summarised in Table 1, these molecules differ from one another in nature, DNA or RNA, in sequence, in length, and in protocol used to determine their solution structures. This may be a source of heterogeneity that is observed in the PDB files. To circumvent this difficulty each original molecular conformation in a PDB file is first translated in an absolute reference frame where all theoretical molecular models are folded. Deformations introduced in the sugar–phosphate backbone are examined at important steps of the folding. We are then in a position to compare each original model conformation of the PDB file to the model structure derived from our theoretical approach.

Multiplicity and heterogeneity of the PDB structures are overcome by setting each conformation in an absolute coordinate frame

As all structures under study were derived from NMR data, their corresponding PDB files contain many proposed solution conformations (Table 1). A direct view of the first ten conformations of different PDB files demonstrates a wide heterogeneity as shown in Figure 3. For 1ac7 (Fig. 3A), the loop appears either very well determined or very rigid, whereas the stem appears either less well determined or more flexible. This view has the advantage of focusing on detailed features of the loop structures (15). The situation is reversed with 1b36 where the main focus is on the central region (Fig. 3B) (21). With the PDB file, 1xue (Fig. 3C), the molecule appears well determined or rigid at every atom positions, whereas with 1c0o, it appears homogeneously underdetermined or flexible (Fig. 3D). This heterogeneity in the PDB structures originates from the arbitrary choice of presentation of superposed molecules. It depends on the molecule and its properties (DNA or RNA, size and sequence, free or bound to a protein) and on the local nature of the two types of information derived from NMR data (torsion angle values from J-couplings and short distances <6 Å from NOE data).

These observations introduce a supplementary difficulty to build the theoretical models at the three different stages of: (i) adjustment of helical thin rod onto the stem to set the elastic curve, BCE1curve, (ii) production of the basic model structure, BCE2basicxyz, (iii) optimisation of angles, Ω and χ of all nucleotides in the loop, BCE3Ωopt. For these matching operations and optimisations to make sense, both molecular models, PDB and theoretical conformations, must be set in the same reference frame coordinates. Since the PDB conformations under study are superposed in arbitrary reference frames, we may choose an absolute and unique reference frame to perform all building operations. It is chosen according to Cambridge conventions on nucleic acids (28). *z* is the axis of the double helical stem and the first stem base pair contiguous to the loop is used to set the origin, O, and the directions and orientations of axes, O*x* and O*y*. The loop of the theoretical model is then automatically built on top of this stem structure with the correct sequence, length and geometry (A-RNA or B-DNA) to yield the basic BCE model. At this point, the nucleotides in the loop have not been rotated, i.e. the loop has not been optimised. It is this unique BCE model that serves as a reference to set the coordinates of the *n* conformations of the PDB file. This is accomplished by superposing any given PDB conformation onto this BCE model. The matching subset is restricted to the sugar–phosphate backbone of the loop and to the first two base pairs of the stem to avoid giving too much weight to the loop or to the stem. Note at this step that the nucleosides in the loop cannot be used since they do not possess correct conformations. Starting from this unique BCE model, all theoretical models are then built by optimising Ω and χ angles of loop nucleosides to each of the *n* conformations in the PDB file as explained in Materials and Methods. Optimised BCE models are energy refined without restraints to yield final theoretical molecular models, BCE4finalm.

Analysis and detailed comparison of the 126 pairs of theoretical and PDB structures is a long task due to the number

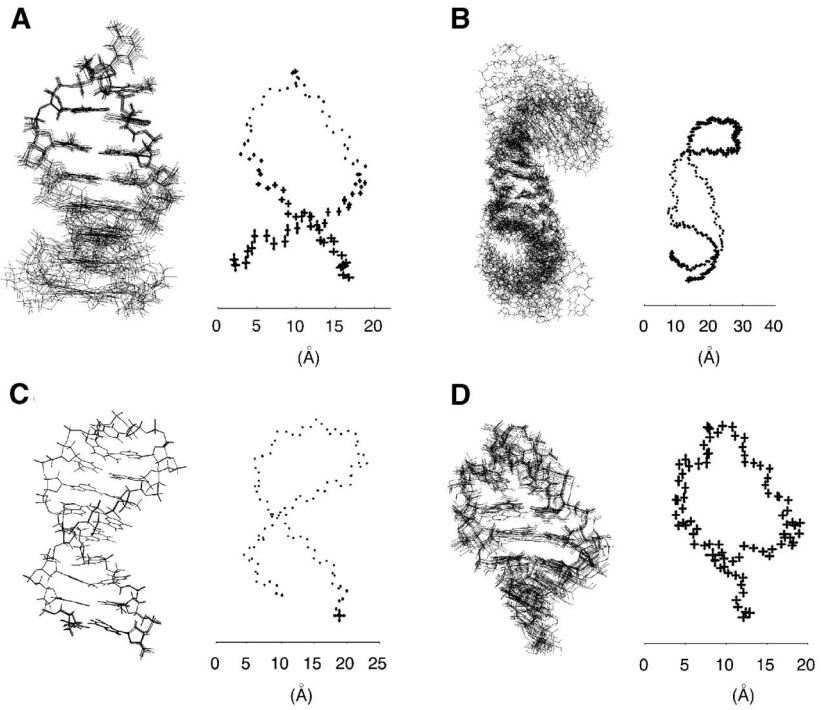


Figure 3. The first ten superimposed views (left) with the standard deviations of the main atoms of the sugar–phosphate backbone plotted at mean atom positions (right) of the *n* molecular conformations available in the PDB for the following PDB identifications and sequences: (**A**) *n* = 10, 1ac7, DNA tetraloop GTTA, (**B**) *n* = 10, 1b36, RNA tetraloop UUCG, (**C**) *n* = 10, 1xue, DNA triloop GCA, (**D**) *n* = 19, 1c0o, RNA tetraloop UUCG. The molecules are of different sizes, and the scale in Å is provided for comparison.

of molecules and to the complexity of each PDB conformation that reflect in part, the intrinsic properties of the sequence, original NMR data as well as modelling protocols used to derive the solution conformations. Presentation of these results is greatly simplified because they can be regrouped into three main homogeneous classes which exactly correspond to the three categories of molecules under study: the DNA tetraloop, DNA triloops and RNA tetraloops. Three representative molecules 1ac7, 1bjh and 1b36 of these categories are sufficient to illustrate all results (Figs 4–7).

Quantitative deformation of the sugar–phosphate main chain

The basic BCE folding procedure described in Figure 2 generates no physical distortions of the initial molecular structure except for the bond lengths and bond angles of the main atoms of the sugar–phosphate backbone. As shown by the dashed lines of Figure 4 (right and left), deformations introduced are generally small: <0.1 Å for bond lengths and <10° for bond angles, except in the region of the sharp turn of B-DNA molecules and in different locations of UUCG RNA molecules. In these regions, deformations are, respectively, generally <0.25 Å and <25°. Note that both bond lengths and bond angles generally oscillate with positive and negative values and that, as expected, both types of plots are well correlated. As shown by the continuous lines of Figure 4, bond lengths and bond angles tend practically to normal values after a short energy refinement without restraints: small oscillations are on the order of thermal fluctuations of bond lengths and bond angles in double helical B-DNA or A-RNA.

Agreement of main chain atoms between theoretical and PDB structures

Three different methods are used here to compare and to show a very good agreement between the trajectories of the main atoms of the sugar–phosphate chains of theoretical models, BCE4finalm and of PDB conformations. Direct and visual comparisons are given in Figure 5 (left and centre) with the superpositions of theoretical and PDB structures and of the elastic line. A quantitative comparison is provided with the plot of distance (d) of main atoms of the sugar–phosphate chain to the elastic line as shown in Figure 5 (right). In these plots, d is <1.2 Å for the stem and for most of the loop except in the region of the sharp turn in DNA hairpins and in the UU region of RNA hairpins. Both sugar–phosphate chains oscillate practically in phase in the loop as in the stem region about the central elastic line. Another means of comparison is the computation of a global mean distance or RMSd for different subsets of atoms as summarised in the 'backbone' columns of Table 2. RMSd are in the range 0.67–1.56 Å for the main backbone atoms of the loop and 1.09–1.36 Å for the 'stem+loop'. These values improve when the third nucleotide, Nc, is omitted from the matching set, respectively, 0.20–1.26 Å and 0.97–1.10 Å. This is expected for 1ac7 since Nc is the least well defined residue from NMR restraints (15). For 1bjh, 1xue and 1zhu, it suggests that Nc, which is in

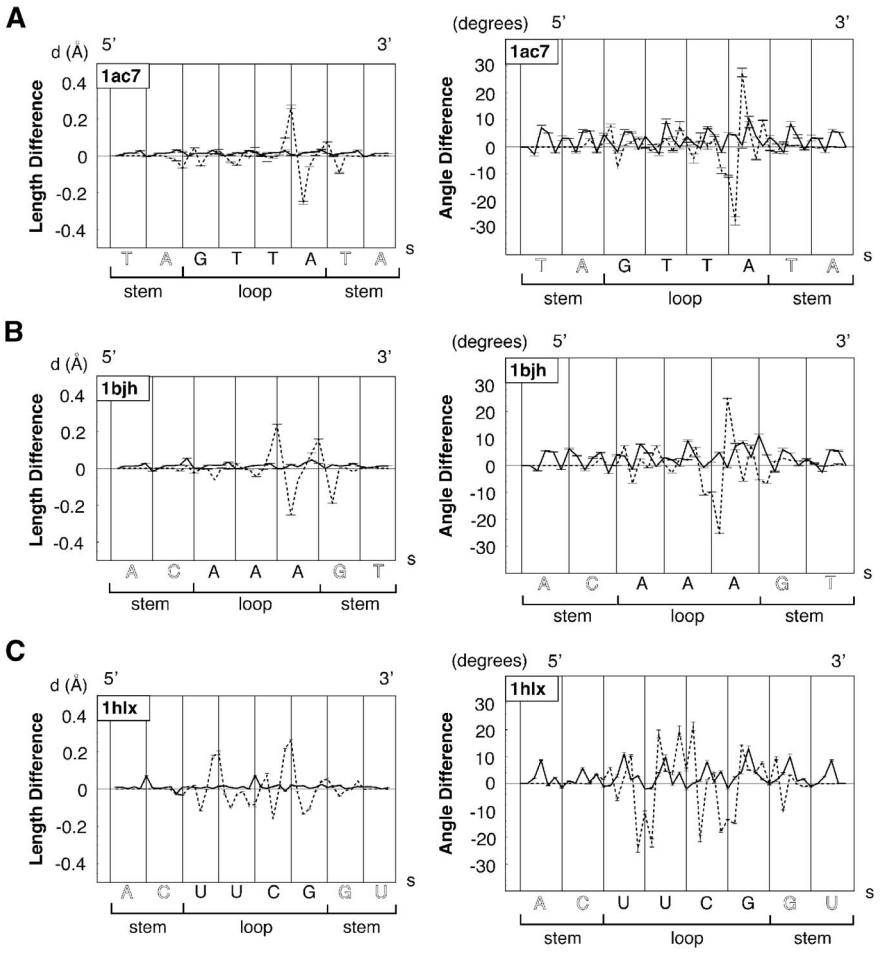


Figure 4. Typical plots of mean bond length differences in Å (left) and of mean bond angle differences in degrees (right) for the main chain atoms of the sugar–phosphate backbone (O5′, C5′, C4′, C3′, O3′, P) at two building steps along the loop sequence in the 5′→3′ direction as a function of arclength, s, in Å. Plots are representative of the three main classes of molecules, DNA tetraloop, DNA triloops and RNA tetraloops, and are computed with the following loop sequences and PDB identifications: (**A**) GTTA (1ac7); (**B**) AAA (1bjh); (**C**) UUCG (1b36). Dashed lines correspond to the third folding BCE step: differences of bond length or bond angle are between optimised, BCE theoretical structural models, BCE3Ωopt, and reference values from standard helical nucleic acids models (B-DNA or A-RNA before folding). Continuous lines correspond to the last building step: values are computed between energy refined theoretical molecular models, BCE4finalm, and reference values of standard helical nucleic acids. Error bars are calculated from the whole set of molecular conformations in the PDB file.

the sharp turn region, is less well resolved than the rest of the molecule. For 1aud, 1b36, 1c0o and 1hlx, it evaluates in part the cost of letting Nc in C3′ endo. Note that agreement is best for the DNA triloops and that representative molecules 1bjh and 1b36 of Figures 4–7 are characterised by the highest RMSd values, indicating that other PDB molecular conformations are better fitted by BCE models.

Agreement between theoretical and PDB structures

Agreement is very good for the DNA tetraloop, DNA triloops and for RNA tetraloops as shown by the direct and visual comparisons in Figure 6 with the superpositions of the theoretical molecular model and of its PDB conformation. Detailed RMSd for all atom subsets are summarised in the 'All atoms' columns of Table 2. They are in the range 1.27–2.11 Å for the loop and 1.32–1.73 Å for the 'stem+loop'. As above these values are improved when the third nucleotide is omitted from the matching set, respectively: 0.97–1.74 Å and 0.92–1.55 Å. Agreements are very good when compared with estimated accuracy of NMR-derived solution structures, 1–1.5 Å (29).

Ω Profiles as a function of sequence

Rotation angles, Ω, of blocks of atoms about the elastic line in the final theoretical models follow one of the three remarkable profiles shown in Figure 7 for the DNA tetraloop, DNA

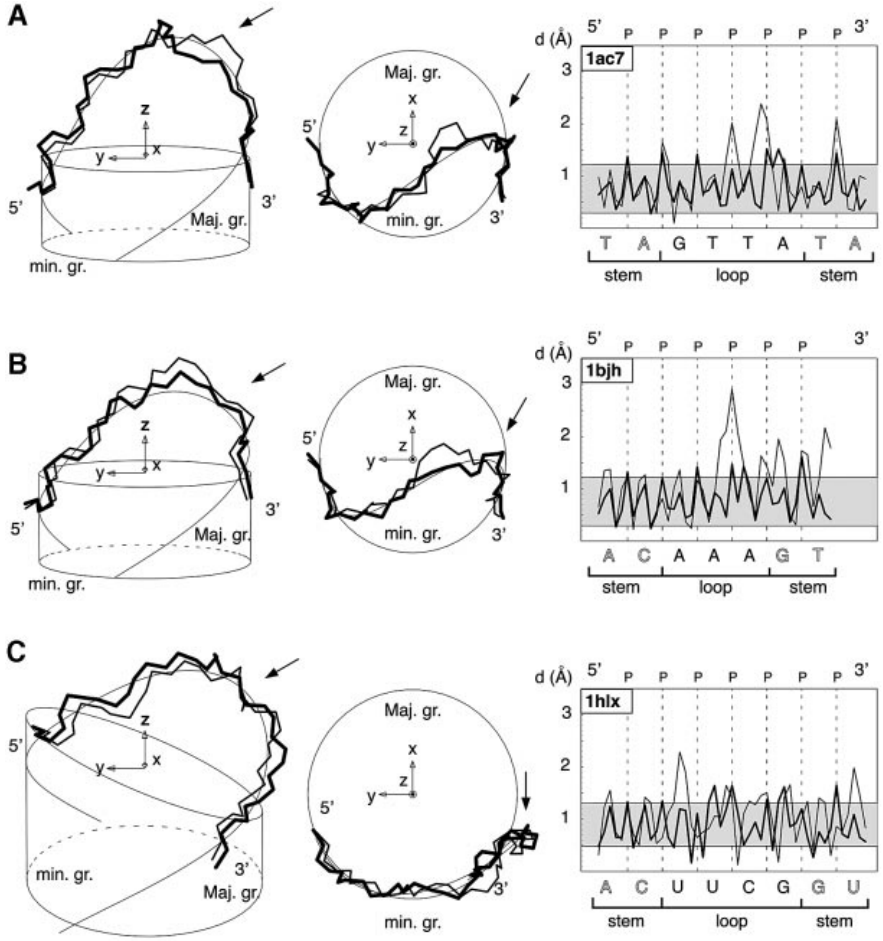


Figure 5. Left and centre: superimposed views of the sugar–phosphate backbones of a molecular conformation available in the PDB, of its corresponding optimised theoretical molecular structure, highlighted in bold, calculated with the BCE approach and of its computed elastic rod curve shown as a thin line (left: view into the minor groove and centre: top view along the helical axis). Right: plot of the distance, d, in Å, of main atoms of the sugar–phosphate chain to the elastic line for the two molecular structures along the loop sequence in the 5′→3′ direction, as a function of arclength, s, in Å. These views and plots are representative of the three main classes of molecules: DNA tetraloop, DNA triloops and RNA tetraloops. Loop sequences and PDB identifications are respectively: (**A**) GTTA (1ac7); (**B**) AAA (1bjh); (**C**) UUCG (1b36). In the left and centre views, the molecular structures are centred in the same global reference frame; radii of the cylinders and circles are those of the sugar–phosphate backbones: (A–B) DNA or (C) RNA. In the right views, d is shown for the best molecular structure.

triloops and RNA tetraloops. Mean and standard deviation of Ω angle values are given for the nucleosides in Table 3. They are in good agreement with qualitative minor or major groove indications of Table 1. As discussed previously (7,18), the differences in Ω values between the DNA tetraloop and the RNA tetraloop, follow from the fact that in the DNA tetraloop the Nb and Nc nucleotides stack upon each other. In the chosen RNA tetraloop Nb folds into the minor groove, while Nc stacks upon the underlying base pair. Surprisingly, we observe that the DNA tetraloop profile and the DNA triloop profiles are very similar. As shown in Figure 7A and B, values of Ω at the 5′ and 3′ ends of the loops are close to zero, and are maximal in the range 89–99° (see also Table 3) for both classes of molecules. Therefore both types of Ω profiles are very close: the monotonous rise of Ω occurs over the first 3 nt for the DNA tetraloop, and over the first two for DNA triloops. Note RNA tetraloop profiles are different for reasons that may also result from the geometric predictions given in Figure 1.

DISCUSSION

Quantitative deformations of the sugar–phosphate main chain

The BCE methodology permits global deformations of the macromolecule with small deformations of the sugar–phosphate chain. The helical line of B-DNA or A-RNA is chosen to pass in the middle of the main chain atoms. Therefore, curving of this elastic line upon folding of the macromolecular chain introduces alternatively compression of chemical bonds for atoms inside the regions of curvature and expansion for atoms outside. This observation explains in part

Table 2. Average and standard deviations in parentheses of RMSd in Å between the final theoretical molecular models, computed from a continuous and flexible thin rod model or 'BCE' model, versus published molecular conformations deposited in the PDB

PDB identification	Main backbone atoms		Main backbone atoms without Nc		All atoms		All atoms without Nc	
	Loop	Stem + loop	Loop	Stem + loop	Loop	Stem + loop	Loop	Stem + loop
DNA								
1ac7	1.29 (0.07)	1.22 (0.13)	0.98 (0.06)	0.98 (0.15)	1.62 (0.10)	1.35 (0.10)	0.97 (0.09)	0.92 (0.11)
1bjh	0.91 (0.01)	1.19 (0.01)	0.37 (0.00)	0.99 (0.01)	1.57 (0.00)	1.38 (0.01)	1.27 (0.01)	1.24 (0.01)
1xue	0.67 (0.00)	1.22 (0.01)	0.20 (0.00)	1.07 (0.01)	1.27 (0.00)	1.34 (0.01)	0.97 (0.00)	1.22 (0.01)
1zhu	0.76 (0.13)	1.15 (0.03)	0.32 (0.05)	1.04 (0.05)	1.31 (0.05)	1.32 (0.02)	1.08 (0.03)	1.21 (0.03)
RNA								
1aud	1.01 (0.11)	1.09 (0.23)	0.92 (0.12)	1.03 (0.21)	1.94 (0.24)	1.70 (0.19)	1.74 (0.29)	1.55 (0.18)
1b36	1.56 (0.12)	1.36 (0.10)	1.26 (0.13)	1.05 (0.11)	2.11 (0.10)	1.73 (0.09)	1.54 (0.12)	1.28 (0.10)
1c0o	1.36 (0.04)	1.24 (0.04)	1.10 (0.03)	0.97 (0.03)	1.92 (0.04)	1.64 (0.04)	1.43 (0.03)	1.26 (0.04)
1hlx	1.29 (0.07)	1.20 (0.07)	1.15 (0.05)	1.10 (0.08)	1.89 (0.08)	1.55 (0.08)	1.62 (0.11)	1.37 (0.11)

Different sets of atoms are taken into account in the RMSd computations with the following notations: 'All atoms' are all nucleotides atoms; 'Main backbone atoms' are: P, O5′, C5′, C4′, C3′, O3′; the 'stem' includes the first two base pairs below the loop. In columns 'without Nc', the third nucleotide, Nc, is not included in the computations.

the oscillatory character of the plots of Figure 4 and also why the bond lengths and bond angles are well restored upon a short energy refinement step. Finally, as shown here and in the different context of the preceding article in this issue, the short energy refinement step gives rise to practically no global deformations of the hairpin structure.

Complexity, multiplicity and heterogeneity of PDB structures: agreement between theoretical and PDB structures

Macromolecules such as DNA and RNA are intrinsically complex and deformable objects, and are therefore difficult to study and to compare with theoretical hairpin molecules. Setting all PDB hairpin coordinates in an absolute reference frame was necessary due to the use of arbitrary reference frames in PDB files. Owing to the flexibilities of the stem, loop and hinge region, we have chosen what seemed to be the best compromise where the weights are proportional to the sizes of the matching sets in stem and loop regions. This method has the advantage of unifying the building procedure of all theoretical structures.

In addition to all these sources of heterogeneity, some of the molecules under study possess outstanding features which may perturb the stem structures. In DNA molecule, 1xue, two unpaired guanines from opposite strands intercalate between sheared G·A base pairs below the first two stem base pairs (30). The 30 nt RNA molecule, 1aud, is part of an RNA–protein complex with 102 amino acids (20). The sequence of 1c0o contains G·U base pairs at the second and third base pairs in the stem, which binds a cobalt hexammine ion (22). Moreover, the numbers of NMR-derived constraints per nucleotide differ depending on the regions of the molecule: 40 for the tetraloop structure, 28 for the stem and an average of 35 for the entire molecule in 1hlx; this results in a higher precision for the loop (17).

All studies on UUCG loops report that the sugars of the two central nucleotides UC in the loop are in C2′ endo conformations whereas all other sugars in the loop or in the stem remain in C3′ endo (17,20–22,31,32). This feature was not taken into account in this preliminary study, and future extension of the folding computer program, S-mol, to DNA or RNA chains with variable puckers and with pucker-dependent chain length (7) should improve the regions of agreements between theoretical and PDB conformations.

These observations and the very good agreements between theoretical and PDB hairpin molecular structures show that the BCE approach and the building method yield robust molecular models. At the present stage of development, they should constitute good starting structures for extensive computational studies based on Metropolis Monte Carlo simulations (33–35) or on molecular dynamics studies (11–13,36,37), where detailed contributions to folding can be examined.

Trajectory of main chain atoms in DNA and RNA hairpins, Ω profiles as a function of sequence and number of nucleotides in the loops

DNA and RNA chains possess an intrinsic BCE that can account for the overall folding shape of these two chemically and geometrically different molecules. This property provides the theoretical grounds for a practical description of nucleotide locations in terms of Ω angles about the central elastic line. It is remarkable that this description appears to be simple transpositions for DNA triloops and the DNA tetraloop. This may be accounted for by the closure of these hairpins by mismatches as explained below.

Loops are closed by a 'mispair' that matches the geometry imposed by the BCE backbone

As remarked before (7,18), the stress induced in the CCCG tetraloop (18) may explain the conversion from Watson–Crick to Hoogsteen base pairing that is observed when pH is lowered. The formation of unusual base pairing for the closing base pair such as Hoogsteen C^+G, or such as GA and UG is a stabilising factor, because the C1′-C1′ distance is shorter than in Watson–Crick base pair, which reduces the stress induced in the loop. The BCE approach should offer a quantitative description to model the loop stress.

The differences in Ω variation throughout the loop (Fig. 7) between DNA and RNA tetraloops are a direct consequence of the choice of loops. The positive Ω values seen in the Type-I, DNA tetraloop is a direct consequence of the continuous stacking. UUCG has a type-II fold: Nb lies then in the minor

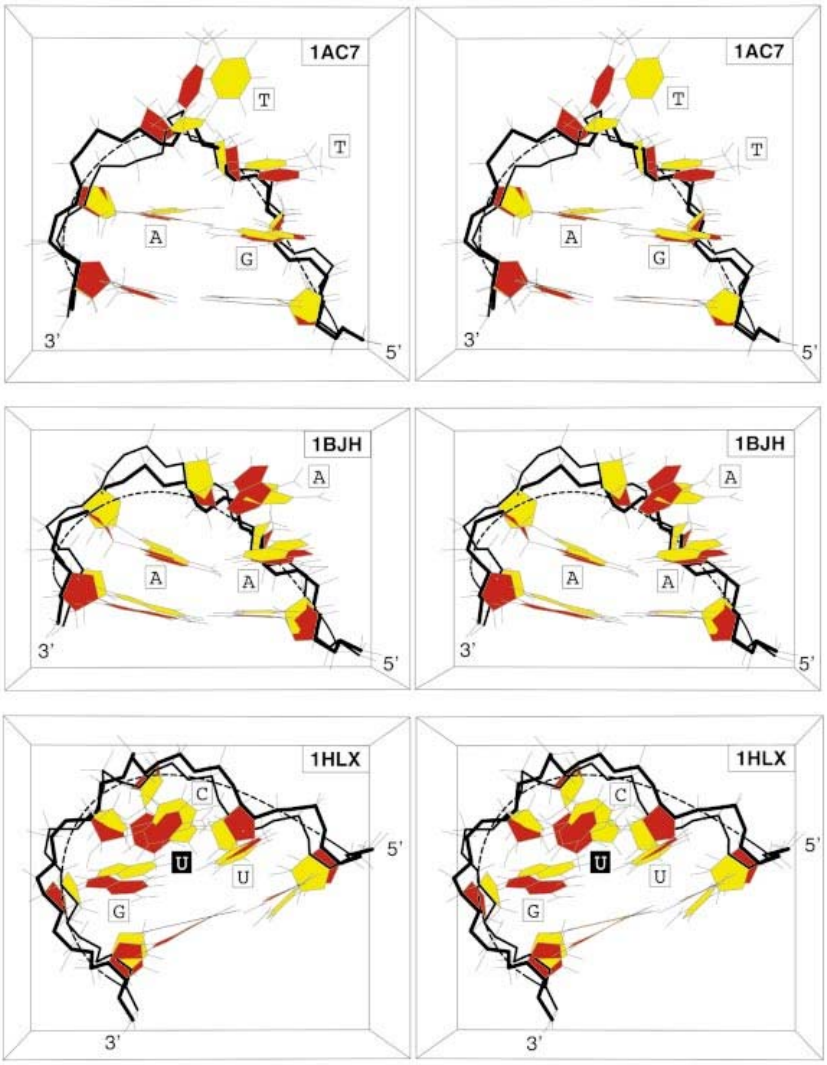


Figure 6. Superimposed stereo views into the major groove of hairpin molecular structures derived by the original authors, in yellow, and of the final molecular structures derived from the elastic curve approach, in red: (**A**) GTTA (1ac7); (**B**) AAA (1bjh); (**C**) UUCG (1b36). These views are representative of the three main classes of molecules under study: DNA tetraloop, DNA triloops and RNA tetraloops. The sugar–phosphate chain is highlighted in bold.

groove so that Ω is negative, Nc stacked on top closing base pair, so that Ω is positive. A type-II DNA tetraloop would show this same pattern in Ω variation. The BCE methodology gives a compact description of these folds. Another remarkable feature consistent with this analysis is that the loop nucleotides appear to literally fall into place upon Ω rotations, i.e. into the correct positions given in the PDB solution conformations. In particular, simple rotations of the first and of the last nucleotides about the elastic line are sufficient to form the required 'mispairs'. This suggests that the G·A base pairing encountered in the DNA GCA or GTTA hairpins, the A·A pairing in the AAA hairpin, and the U·G base pairing in RNA UUCG hairpins should no longer be considered as 'mismatches' but rather as the best possible base pairings capable of fulfilling the geometric conditions imposed by the BCE hairpin fold. Up to now, these mispairs were regarded as major contributors to the stability because they augment stacking and the number of hydrogen bonding interactions. In contrast, these observations and those obtained previously (5) indicate that the sugar–phosphate backbones adopt a BCE conformation, whether a mismatch is formed or not (5), and that mispairs are an additional stabilising factor, if permitted by the BCE backbone. From this perspective, the most conceptually economical way to fold the DNA triloops is to regard them as hairpins with 3 nt in the loop and not as 1-nt loop. In the same way, loop -GTTA- should be regarded as a tetraloop and not as a hairpin with 2 nt in the loop, since the structure of the sugar–phosphate chain can be deduced from the geometry of the B-DNA stem and since the mispair G·A can be easily formed by two simple Ω rotations (Fig. 7) to add

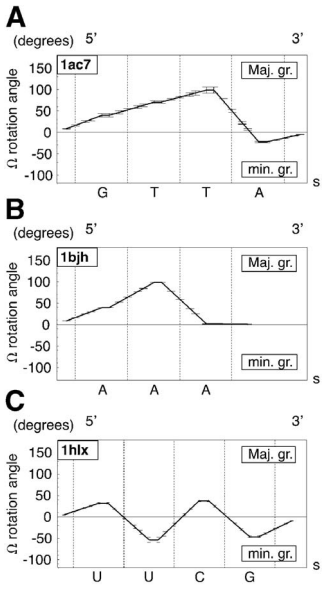


Figure 7. Representative plots of the nucleotide sub-blocks rotation angle, Ω, in degrees, about the elastic rod curve along the loop sequence in the 5′→3′ direction as a function of arclength, s, in Å, in the final theoretical models for the three main classes of molecules under study: DNA tetraloop, DNA triloops and RNA tetraloops: (**A**) GTTA (1ac7); (**B**) AAA (1bjh); (**C**) UUCG (1b36).

Table 3. Average rotation angles, Ω_x, and standard deviation of the loop bases, N_x, of the final theoretical molecular models computed from the BCE curve with PDB conformation structures. Ω_a is the rotation angle of N_a, Ω_b of N_b, Ω_c of N_c and Ω_d of N_d

PDB identification	Ω_a	Ω_b	Ω_c	Ω_d
DNA				
1ac7	39.6 (3.9)	70.6 (2.9)	98.6 (7.4)	–22.6 (1.6)
1bjh	39.3 (0.2)	98.2 (0.1)	1.1 (0.1)	-
1xue	30.9 (0.0)	88.6 (0.0)	–0.5 (0.0)	-
1zhu	32.6 (0.3)	89.7 (2.2)	0.5 (0.6)	-
RNA				
1aud	29.9 (33.1)	–92.7 (36.6)	41.1 (14.7)	–40.3 (7.5)
1b36	33.5 (3.8)	–57.4 (6.1)	43.2 (3.4)	–49.2 (4.3)
1c0o	32.4 (1.7)	–63.6 (3.2)	39.7 (1.2)	–44.1 (1.8)
1hlx	32.0 (1.9)	–53.7 (6.1)	37.7 (1.9)	–46.4 (2.3)

stabilising H bonding and stacking interactions. The same reasoning holds also for RNA UUCG loops. Although more molecules need to be studied in terms of this perspective, all these results appear remarkably coherent.

CONCLUSION

Bending a few nucleotides segment of a macromolecular chain as a thin rigid rod of elasticity theory is one of the simplest conceptual models to fold DNA or RNA macromolecules into hairpins. With this simple idea, we have shown that single-stranded B-DNA can be deformed into hairpin loops that match not only all published NMR data available for trinucleotide TTT loops (5), but also the PDB structures of tri- and tetra-loops of DNA. We have shown in addition that single-stranded A-RNA can be deformed with the same folding methodology into UUCG tetraloops. Note the shapes of DNA and RNA hairpins are different, but are well reproduced by the same methodology applied with the different end conditions imposed by B-DNA or A-RNA helical geometries. These results tend to demonstrate that elastic properties of the sugar–phosphate chains play a key role to understand the folding shapes of both DNA and RNA into hairpins. Up to now, several main types of interactions have been invoked to explain the remarkable stability of all hairpins under study: specific hydrogen bonding, stacking and hydrophobic interactions. The sugar–phosphate chains appear to fold along the smoothest lines of least deformation energy (given by elasticity theory) and most torsion angles remain close to their initial values (B-DNA or A-RNA). It suggests that, for these molecules, the elastic properties of sugar–phosphate chains are an important structural and energetic contribution to hairpin folding that may account for their extraordinary stability.

According to usual descriptions, hairpins are double helical base-paired stems capped by a loop sequence of unpaired or of mismatched nucleotides. In the proposed view, these strange mismatches (G·A in tetra- and tri-loops, A·A in triloop AAA, or U·G in UUCG) should rather be considered as very good base pairings that satisfy the geometric requirements imposed by the BCE fold. Note in contrast that Watson–Crick base pairs would not meet these requirements well.

The new parameter angles, Ω, offer a very coherent simplification of the descriptions of hairpin loops containing G·A, A·A or U·G base pairings. More studies are needed to check whether other hairpins can be reproduced with the BCE approach and described in terms of parameter angles, Ω. If so, they would provide the first quantitative measurements to classify and to understand the structures of DNA and RNA hairpin loops and possibly of many other important biological macromolecules.

ACKNOWLEDGEMENTS

It is a pleasure to thank Ms C. Cordier for revision of the English text and our colleagues of the L.P.B.C. for constant support: Mr J. Bolard, M. Ghomi and P.-Y. Turpin. C.P. acknowledges the support of the MENESR and of the Fondation pour la Recherche Médicale. J.A.H.C. was supported by the Université P. et M. Curie and the Département des Sciences Chimiques du CNRS.

REFERENCES

1. Boulard,Y., Gabarro-Arpa,J., Cognet,J.A.H., Le Bret,M., Guy,A., Téoule,R., Guschlbauer,W. and Fazakerley,G.V. (1991) The solution structure of a DNA hairpin containing a loop of three thymidines determined by nuclear magnetic resonance and molecular mechanics. *Nucleic Acids Res.*, **19**, 5159–5167.
2. Mooren,M.M.W., Pulleyblank,D.E., Wijmenga,S.S., van de Ven,F.J. and Hilbers,C.W. (1994) The solution structure of the hairpin formed by d(TCTCTC-TTT-GAGAGA). *Biochemistry*, **33**, 7315–7325.
3. Kuklenyik,Z., Yao,S. and Marzilli,L.G. (1996) Similar conformations of hairpins with TTT and TTTT sequences: NMR and molecular modeling evidence for T.T base pairs in the TTTT hairpin. *Eur. J. Biochem.*, **236**, 960–969.

4. Chou,S.H., Tseng,Y.Y. and Chu,B.Y. (2000) Natural abundance heteronuclear NMR studies of the T3 mini-loop hairpin in the terminal repeat of the adenoassociated virus 2. *J. Biomol. NMR*, **17**, 1–16.
5. Pakleza,C. and Cognet,J.A.H. (2003) Biopolymer chain elasticity: a novel concept and a least deformation energy principle predicts backbone and overall folding of DNA TTT hairpins in agreement with NMR distances. *Nucleic Acids Res.*, **31**, 1075–1085.
6. Haasnoot,C.A.G., Hilbers,C.W., van der Marel,G.A., van Boom,J.H., Singh,U.C., Pattabiraman,N. and Kollman,P.A. (1986) On loopfolding in nucleic acid hairpin-type structures. *J. Biomol. Struct. Dyn.*, **3**, 843–857.
7. Hilbers,C.W., Heus,H.A., van Dongen,M.J. and Wijmenga,S.S. (1994) The hairpin elements of nucleic acid structure: DNA and RNA folding. In Eckstein,F. and Lilley,D.M.J. (eds), *Nucleic Acids and Molecular Biology*. Springer Verlag, Berlin Heidelberg, pp. 56–104.
8. Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F.,Jr, Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.
9. Cheong,C., Varani,G. and Tinoco,I.,Jr (1990) Solution structure of an unusually stable RNA hairpin, 5′GGAC(UUCG)GUCC. *Nature*, **346**, 680–682.
10. Hirao,I.Y., Kawai,S., Yoshizawa,Y., Nishimura,Y., Ishido,K., Watanabe,K. and Miura,K. (1994) Most compact hairpin-turn structure exerted by a short DNA fragment, d(GCGAAGC) in solution: an extraordinarily stable structure resistant to nuclease and heat. *Nucleic Acids Res.*, **22**, 576–582.
11. Miller,J.L. and Kollman,P.A. (1997) Theoretical studies of an exceptionally stable RNA tetraloop: observation of convergence from an incorrect NMR structure to the correct one using unrestrained molecular dynamics. *J. Mol. Biol.*, **270**, 436–450.
12. Miller,J.L. and Kollman,P.A. (1997) Observation of an A-DNA to B-DNA transition in a nonhelical nucleic acid hairpin molecule using molecular dynamics. *Biophys. J.*, **73**, 2702–2710.
13. Zacharias,M. (2001) Conformational analysis of DNA-trinucleotide-hairpin-loop structures using a continuum solvent model. *Biophys. J.*, **80**, 2350–2363.
14. Varani,G. (1995) Exceptionally stable nucleic acid hairpins. *Annu. Rev. Biophys Biomol. Struct.*, **24**, 379–404.
15. van Dongen,M.J.P., Mooren,M.M.W., Willems,E.F.A., van der Marel,G.A., van Boom,J.H., Wijmenga,S.S. and Hilbers,C.W. (1997) Structural features of the DNA hairpin d(ATCCTA-GTTA-TAGGAT): formation of a G-A pair in the loop. *Nucleic Acids Res.*, **25**, 1537–1547.
16. Chou,S.-H., Zhu,L., Gao,Z., Cheng,J.-W. and Reid,B.R. (1996) Hairpin loops consisting of single adenine residues closed by sheared A.A and G.G pairs formed by the DNA triplets AAA and GAG: solution structure of the d(GTACAAAGTAC) hairpin. *J. Mol. Biol.*, **264**, 981–1001.
17. Allain,F.H.-T. and Varani,G. (1995) Structure of the P1 helix from group I self-splicing introns. *J. Mol. Biol.*, **250**, 333–353.
18. van Dongen,M.J.P., Wijmenga,S.S., van der Marel,G.A., van Boom,J.H., and Hilbers,C.W. (1996) The transition from a neutral-pH double helix to a low-pH triple helix induces a conformational switch in the CCCG tetraloop closing the Watson–Crick Stem. *J. Mol. Biol.*, **263**, 715–729.
19. Gesteland,R.F., Cech,T.R. and Atkins,J.F. (eds) (1999) *The RNA World*. (2nd Edn) Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, p. 709.
20. Allain,F.H.-T., Howe,P.W.A., Neuhaus,D. and Varani,G. (1997) Structural basis of the RNA-binding specificity of human U1A protein. *EMBO J.*, **16**, 5764–5774.
21. Butcher,S.E., Allain,F.H.-T. and Feigon,J. (1999) Solution structure of the B domain from the hairpin ribozyme. *Nature Struct. Biol.*, **6**, 212–216.
22. Colmenarejo,G. and Tinoco,I.,Jr (1999) Structure and thermodynamics of metal binding in the P5 helix of a group I intron ribozyme. *J. Mol. Biol.*, **290**, 119–135.
23. Arnott,S., Campbell-Smith,P. and Chandrasekaran,R. (1976) Atomic coordinates and molecular conformations for DNA-DNA, RNA-RNA, and DNA-RNA helices. In Fasman,G.D. (ed.), *CRC Handbook of Biochemistry and Molecular Biology*. Vol. 2, CRC Press Inc., Cleveland, OH, pp. 411–422.
24. Wolfram Research, Inc. (1999) *Mathematica*, Version 4, Champaign, IL, p. 1470.
25. Geomview, The Geometry Center, University of Minnesota, Minneapolis, USA.
26. Weiner,S.J., Kollman,P.A., Case,D.A., Singh,U.C., Ghio,C., Alagona,G. and Weiner,P.K. (1984) A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.*, **106**, 765–784.
27. Weiner,S.J., Kollman,P.A., Nguyen,D.T. and Case,D.A. (1986) An all atom force field for simulations of proteins and nucleic acids. *J. Comput. Chem.*, **7**, 230–252.
28. Dickerson,R.E., Bansal,M., Calladine,C.R., Diekmann,S., Hunter,W.N., Kennard,O., von Kitzing,E., Lavery,R., Nelson,H.C., Olson,W.K., Saenger,W., Shakked,Z., Sklenar,H., Soumpasis,D.M., Tung,C.S., Wang,A.H.-J. and Zhurkin,V.B. (1989) Definitions and nomenclature of nucleic acid structure parameter., *EMBO J.*, **8**, 1–4.
29. Allain,F.H.-T. and Varani,G. (1997) How accurately and precisely can RNA structure be determined by NMR. *J. Mol. Biol.*, **267**, 338–351.
30. Zhu,L., Chou,S.-H. and Reid,B.R. (1996) A single G-to-C change causes human centromere TGGAA repeats to fold back into hairpins. *Proc. Natl Acad. Sci. USA*, **93**, 12159–12164.
31. Ennifar,E., Nikulin,A., Tishchenko,S., Serganov,A., Nevskaya,N., Garber,M., Ehresmann,B., Ehresmann,C., Nikonov,S. and Dumas,P. (2000) The crystal structure of UUCG tetraloop. *J. Mol. Biol.*, **304**, 35–42.
32. Ban,N., Nissen,P., Hansen,J., Moore,P.B.S. and Steitz,T.A. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science*, **289**, 905–920.
33. Ghomi,M., Victor,J.-M. and Henriet,C. (1994) Monte Carlo simulations on short single-stranded oligonucleotides. I. Application to RNA trimers. *J. Comp. Chem.*, **15**, 433–445.
34. Gabb,H.A., Prevost,C. and Lavery,R. (1995) Efficient conformational space sampling for nucleosides using internal coordinate Monte Carlo simulations and a modified furanose description. *J. Comp. Chem.*, **16**, 667–680.
35. Tung,C.S. (1997) A computational approach to modeling nucleic acid hairpin structures. *Biophys. J.*, **72**, 876–885.
36. Zichi,D.A. (1995) Molecular dynamics of RNA with the OPLS force fields. Aqueous simulation of a hairpin containing a tetranucleodide loop. *J. Am. Chem. Soc.*, **117**, 2957–2969.
37. Auffinger,P., Louise-May,S. and Westhof,E. (1999) Molecular dynamics simulations of solvated yeast tRNA(Asp). *Biophys. J.*, **76**, 50–64.
38. Zhu,L., Chou,S.-H., Xu,J. and Reid,B.R. (1995) Structure of a single-cytidine hairpin loop formed by the DNA triplet GCA. *Nature Struct. Biol.*, **2**, 1012–1017.

Chapitre IV

Les appariements dans les boucles

L'étude des tri- et tétra-boucles d'ADN et des tétra-boucles d'ARN présentée au chapitre précédent a montré que l'approche BCE permet de construire les conformations tridimensionnelles des épingles à cheveux comportant des mésappariements entre les bases extrémales de la boucle au moyen d'un nombre très restreint de degrés de liberté (Ω_i,χ_i) où i varie de 1 à 3 pour les tri-boucles et de 1 à 4 pour les tétra-boucles. La trajectoire de la chaîne sucre-phosphate étant calculée avec la théorie de l'élasticité, les deux degrés de liberté par nucléotide, Ω et χ, semblent suffisants pour mettre en place ces mésappariements. De ce point de vue, il semblerait que ces "mésappariements" ne doivent plus être considérés comme de mauvais appariements selon leur étymologie, mais plutôt comme les meilleurs appariements possibles pour une séquence donnée dans la cadre de la trajectoire tridimensionnelle de la chaîne sucre-phosphate dans la partie en boucle. Afin de confirmer ou d'infirmer cette hypothèse dans le cas des tri-boucles d'ADN, nous allons étudier exhaustivement la formation des appariements entre les bases extrémales des boucles et chercher à établir les meilleurs appariements pour les comparer avec ceux rencontrés dans les tri-boucles publiées. Nous allons mettre en place un protocole d'exploration exhaustif des conformations pour évaluer les interactions de liaison hydrogène et pour déterminer ainsi les conformations d'appariement *a priori* les plus stables.

Après la présentation de notre protocole d'exploration et de l'ensemble des étapes détaillées, nous décrirons les résultats obtenus, c'est-à-dire les meilleurs appariements trouvés *a priori* pour chaque séquence. Nous comparerons ces résultats aux conformations déjà publiées en utilisant les tri-boucles d'ADN

comportant un appariement dans la boucle comme contrôle positif [47, 55, 56, 62–65, 67]. Nous conclurons sur le caractère prédictif de notre approche et sur les perspectives qu'ouvre une voie de modélisation telle que BCE.

IV.1 Protocole d'exploration de la formation des appariements dans les tri-boucles d'ADN

Pour explorer exhaustivement la formation des appariements dans les tri-boucles d'ADN nous devons tenir compte des variables suivantes : la séquence de la boucle des bases extrémales B_1 et B_3, la conformation ANTI ou SYN de ces bases et leurs positions spatiales relatives. Pour chaque conformation des bases B_1 et B_3, il faut tester toutes les liaisons hydrogène suceptibles de se former, et calculer un score de liaison hydrogène pour chacune d'entre elles. L'ensemble est réalisé au moyen du protocole exposé ci-après :

IV.1.1 Choix de la séquence de la boucle

Pour explorer exhaustivement toutes les séquences de tri-boucles d'ADN, il faudrait faire varier la nature de chaque base qui la compose. Si l'on considère que la conformation d'une tri-boucle dépend de la séquence de la boucle et du dernier plateau de paire de bases de la tige, il faut donc faire varier la séquence de chacune de 5 bases. Toutes les séquences possibles de boucles sont donc dénombrées en choisissant chaque nucléotide parmis les quatre bases A, T, G et C formant l'alphabet de l'ADN. Le nombre de séquences possibles s'élève alors à 4^5=1024 séquences différentes.

Dans le cadre de l'exploration des appariements dans les tri-boucles avec BCE, nous négligeons en première approximation l'effet de séquence des bases flanquantes sur l'appariement. Nous considérons dans ce chapitre que l'effet de séquence principal provient de la nature même des bases impliquées dans l'appariement potentiel. La mise en place de la ou des liaisons hydrogène pouvant conduire à la fomation de l'appariement dépend d'abord de la nature et de la géométrie des groupements chimiques donneurs et accepteurs de proton portés les cycles des bases. La nature des bases du dernier plateau de la tige et de la base centrale de la boucle B_2 n'est donc

pas ici considérée comme des variables dans notre calcul. Le nombre de séquences à explorer est donc ainsi réduit aux 4^2=16 paires de bases envisageables entre deux nucléotides arbitrairement choisis.

IV.1.2 Choix de la conformation des bases ANTI ou SYN

Dans une conformation empilée sur un plateau, une base peut présenter l'une ou l'autre de ses deux faces au plateau. Dans le cas d'un simple brin en boucle, cette alternative peut être associée aux conformations ANTI ou SYN des bases. Elles sont caractérisées approximativement par une rotation à plus ou moins 180° environ de la base autour de la liaison glycosidique. Afin d'explorer exhaustivement toutes les conformations appariées possibles, nous devons prendre en compte l'effet de cette rotation sur les géométries d'appariements et sur la formation des liaisons hydrogène entre les bases extrémales de la boucle et ainsi distinguer les quatre cas suivants : ANTI-ANTI, ANTI-SYN, SYN-ANTI et SYN-SYN. De seize explorations différentes données par la variation de la séquence, nous passons à 16×4=64 explorations différentes.

IV.1.3 Exploration du positionnement relatif des bases extrémales

IV.1.3.1 Description de l'exploration des conformations

L'approche BCE offre un cadre d'étude très propice à l'exploration de l'espace conformationnel des bases empilées. Le positionnement de chaque base de la boucle dépend en effet d'un petit nombre de degrés de liberté. Il est donc envisageable de les faire varier systématiquement avec une bonne précision pour rechercher les conformations présentant de bonnes liaisons hydrogène essentielles à la formation d'un appariement stable.

Aux deux degrés de liberté (d.d.l.) par nucléotide utilisés dans l'approche classique BCE (Ω et χ) [80], il faut ajouter le d.d.l. de rotation de redressement d'empilement d'angle Θ_{empil} pour explorer l'espace des conformations des bases empilées. Le raisonnement intuitif conduit donc à dénombrer trois d.d.l. différents pour placer la base dans une conformation empilée : Ω, χ et Θ_{empil}. Pourtant, dans la pratique,

seul le d.d.l. Ω est nécessaire, car les paramètres Θ_{empil} et χ sont déduits de Ω lors de l'exploration de l'espace conformationnel . En effet, il faut rappeler deux caractéristiques de la rotation de redressement d'empilement telle que nous l'avons définie au chapitre II :

- À une valeur donnée de l'angle de rotation Ω , correspond un angle Θ_{empil} et un seul permettant de placer la liaison glycosidique dans un plan globalement parallèle au dernier plateau de paire de bases de la tige, moyennant un angle minimal. Θ_{empil} est donc directement déductible de Ω.
- Une fois la liaison glycosidique placée dans une géométrie globalement parallèle au dernier plateau de la tige par la rotation de redressement d'empilemement d'angle Θ_{empil}, le calcul de χ est automatique. Seules deux valeurs de χ permettent de placer l'ensemble des atomes de la base parallèlement au plan moyen du dernier plateau de l'hélice. Ces deux valeurs de χ conduisent alternativement à placer la base en conformation ANTI ou SYN. Le choix de la conformation de la base étant un paramètre fixé de l'exploration, la définition de χ pour une valeur donnée de Ω est univoque.

Les deux d.d.l. Θ_{empil} et χ ne doivent donc pas être ici considérés comme des paramètres d'exploration, mais plutôt comme des paramètres de reconstruction calculés automatiquement pour explorer l'espace conformationnel auquel on souhaite se limiter, celui des bases empilées sur le dernier plateau de l'hélice.

Ainsi cadrée, l'exploration des appariements dans la boucle ne dépend plus que de deux paramètres d'exploration, les angles Ω_1 et Ω_3, associés aux premier et dernier nucléotides de la boucle. L'exploration de cet espace conformationnel est réalisé en faisant varier chaque angle de rotation Ω des nucléotides 1 et 3 entre -120° et +120° par pas de 0,5° autour de leur position initiale donnée par le repliement d'un simple brin en hélice sur la trajectoire donnée par la théorie de l'élasticité. Pour effectuer cette étude, il suffit donc de générer deux banques de 481 structures de nucléotides empilés sur dernier plateau de l'hélice : une pour chaque nucléotide i=1 et i=3.

IV.1.3.2 Aspects pratiques de l'exploration de l'espace conformationnel

Afin de tester tous les appariements possibles, nous construisons les 64 banques de 481 conformations de bases empilées. Il y a pour chaque position de la base (5' ou

3' de la boucle), 32 banques différentes qui font varier la nature de la base et sa conformation (ANTI ou SYN). Pour explorer les mésappariements entre deux bases dans une conformation donnée, on sélectionne les deux banques de conformations adéquates. Pour explorer la formation des appariements entre un nucléotide en 5' et un nucléotide en 3' de la boucle, on crée tous les couples possibles d'un nucléotide pris d'une banque de conformation avec tous les nucléotides de l'autre banque. Pour chaque couple de dinucléotides N_1/N_3 cela conduit à la construction de 481×481=231361 conformations de dinucléotides possibles par exploration. Dans chacun de ces cas, la formation de toutes les liaisons hydrogène possibles sera évaluée avec la fonction de score de liaison hydrogène (*cf. infra*). Ces explorations donnent lieu à des cartes de liaisons hydrogène éventuelles (*cf.* FIG. : IV.1).

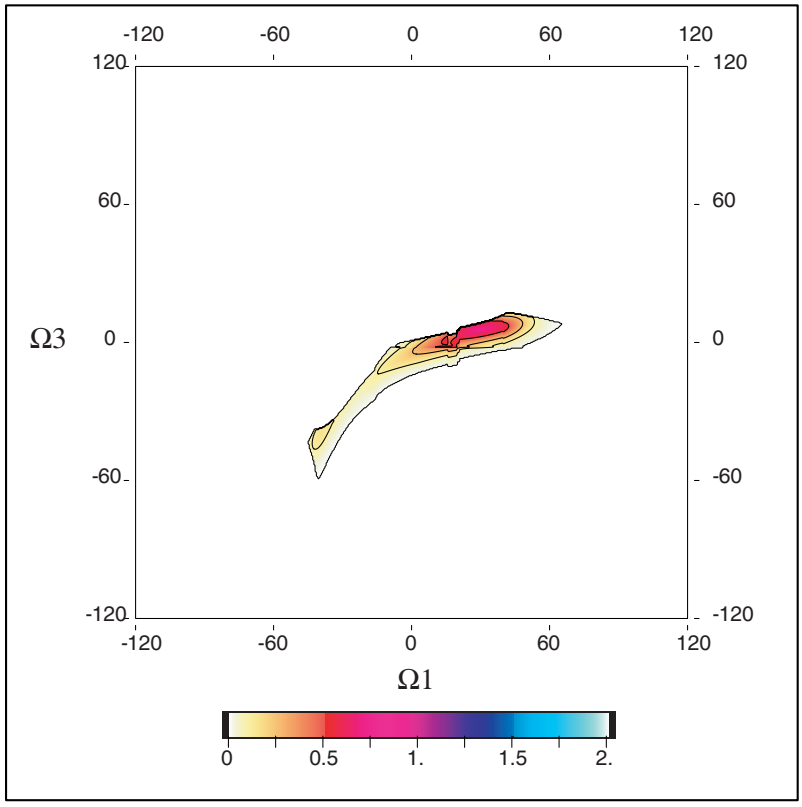


FIG. IV.1 : *Exemple de carte de score de liaison hydrogène :* *Carte de score de la liaison hydrogène A1 N3 - A3 HN6A de l'appariement ADEanti/ADEanti en fonction des angles en degré des rotations autour de la tangente au fil élastique, Ω_1 et Ω_3, des nucléotides extrémaux de la tri-boucle. La coloration de la carte de contour, illustrée par la légende, dépend de la valeur de score de liaison hydrogène pour chaque valeur de Ω_1 et Ω_3. Le score pour une liaison hydrogène unique varie entre 0 et 1 (*cf. PART. : *II.7.3.5), mais peut, lors de la sommation de plusieurs cartes, prendre des valeurs supérieures. La légende, générée automatiquement par notre programme, prend donc en compte des valeurs de score variant entre 0 et 2.*

IV.1.3.3 Analyse des paramètres de construction Θ_{empil} et χ calculés pour explorer l'espace conformationnel des bases empilées

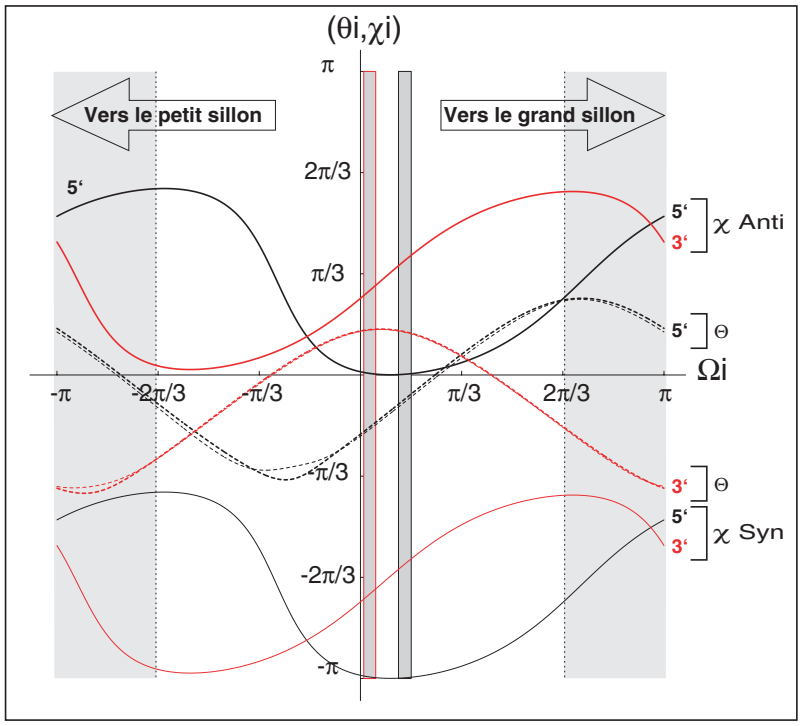


FIG. IV.2 : ***Détermination des angles Θ_{empil} et χ en fonction de Ω pour empiler les bases :*** *Les relations concernant la base en 5' sont en noir et celles de la base en 3' sont en rouge. La relation donnant les angles Θ_{empil} est en traits pointillés et celle donnant χ est en traits continus . Ces relations sont calculées pour placer les bases A, T, G et C dans des géométries empilées sur le dernier plateau de paire de bases de la tige, en fonction de la rotation de valeur Ω variant de -π rad. à +π rad. On distingue deux cas pour χ : en trait gras lorsque la base est en conformation* ANTI *et en trait fin si la base est en conformaiton* SYN. *Les bandeaux gris verticaux sur les côtés du graphe représentent les limites des -120° et +120° entre lesquelles sont explorées la formation des appariements,* i.e. *les liaisons hydrogène. Les bandeaux gris verticaux au centre du graphe représente les zones dans lesquelles sont trouvés les meilleurs appariements, en accord avec les données expérimentales des tri-boucles GAA, GCA, AAA et ATC et GAC (entouré en noir pour la base en 5' et en rouge pour la base en 3' de la boucle).*

Les d.d.l. de construction Θ_{empil} et χ sont calculés automatiquement en fonction de la position de la base considérée dans la boucle (en 5'$\rightarrow N_1$, ou en 3'$\rightarrow N_3$=2 choix), en fonction de sa séquence (A, T, C ou G=4 choix), en fonction de sa conformation (ANTI ou SYN=2 choix) et en fonction de la valeur de rotation, Ω, du nucléotide. Les valeurs de ces angles sont reportées sur les profils de la figure IV.2. Théoriquement les choix multiples doivent produire, au plus, 32 courbes différentes (*cf.* EQ. : IV.1.3.1). Un grand nombre de courbes sont cependant confondues, car

ces tracés sont pratiquement indépendants de la nature de la base A, T, C ou G. Pour une base donnée (en 5' ou en 3'), on observe une seule fonction $\Theta_{empil} = f(\Omega)$, et deux fonctions $\chi = g(\Omega)$, où χ est en effet déterminé à π près. On note que ces fonctions sont différentes selon la position de la base en 5' ou en 3'. En effet, contrairement à la double hélice d'ADN, les positions de ces deux bases ne peuvemt pas se déduire par symétrie dyadique. On met ainsi en évidence que les angles Θ_{empil} et χ calculés, sont quasiment indépendants de la nature de la base et que l'angle de rotation de redressement Θ_{empil} est indépendant de la conformation ANTI ou SYN de la base. Des études exploratoires et de contrôle portant sur les angles Θ_{empil} et χ calculés pour les tétra-boucles d'ADN, d'ARN et dans les hélices de conformation B (courbes non présentées), montrent qu'outre la position 5' ou 3' de la base, le seul facteur modifiant significativement les profils Θ_{empil}=f(Ω) est la position relative de la trajectoire de la boucle par rapport au dernier plateau de paire de bases de la tige.

$$\begin{aligned} & 2 && (\chi\, ou\, \Theta_{empil}) \\ \times\ & 2 && (5'\, ou\, 3') \\ \times\ & 4 && (A\, ou\, T\, ou\, C\, ou\, G) \\ \times\ & 2 && (Anti\, ou\, Syn) \\ =\ & 32 && courbes\, diff\acute{e}rentes \end{aligned} \qquad \text{(IV.1.3.1)}$$

Les profils de χ où les bases sont en conformation ANTI se déduisent de ceux où les bases sont en conformation SYN par une translation verticale de près de 180°, en cohérence avec la définition de ces deux conformations.

Les variations de l'angle Θ_{empil} s'inscrivent dans un intervalle compris en moyenne entre -59° et 44° pour les bases en 5' et entre -49° et 27° pour les bases en 3' des tri-boucles d'ADN, pour Ω variant entre les valeurs extrèmes : -120° et 120°. Comme les angles constatés sont de faible amplitude la déformation de la chaîne sucre-phosphate à l'échelle des angles de torsion est beaucoup plus faible. Nous nous assurons ainsi que notre protocole reste en conformité avec les règles de préservation des angles de torsion initiaux de la chaîne sucre-phosphate que nous nous sommes données.

IV.1.4 Exploration de toutes les liaisons hydrogène possibles

IV.1.4.1 Les groupements proton et accepteur de protons pris en compte pour la formation éventuelle de liaison hydrogène

Pour être exhaustif, il faut tester l'ensemble des liaisons hydrogène suceptibles de se former pour une séquence donnée. Pour cela, nous répertorions tous les accepteurs et tous les donneurs potentiels de protons des cycles des bases des deux nucléotides N_1 et N_3. Le nombre et la nature de ces groupements chimiques ne dépendent que de la nature de la base.

Base	**Protons**		**Accepteurs**
	N-H & O-H	C-H	
ADE	HN6A; HN6B	H2; H8	N1; N3; N7
THY	H3	H7A; H7B; H7C	(O2a; O2b); (O4a; O4b)
GUA	H1; HN2A; HN2B	H8	N3; (O6a; O6b); N7
CYT	HN4A; HN4B	H5	(O2a; O2b); N3

Tab. IV.1 : ***Liste des protons et des atomes accepteurs pouvant être impliqués dans la formation de liaisons hydrogène pour chaque base :*** *Cette liste décrit l'ensemble des groupements donneurs et accepteurs de proton testés dans le cadre de l'exploration de la formation des liaisons hydrogène avec notre protocole. Tous les protons sont pris en compte dans nos calculs à l'exception du proton H6 des pyrimidines qui est exclu des calculs du fait de sa position peu propice à la formation d'une liaison hydrogène. Il faut remarquer que les protons liés à des atomes de carbone distingués dans une colonne propre sont considérés comme susceptibles de former des liaisons hydrogène. Ces liaisons hydrogène de type C-H$\cdots$O seront distinguées après le traitement comme des liaisons hydrogène de plus faible intensité.*

Nous considérons dans cette étude que tout proton des bases peut potentiellement être engagé dans une liaison hydrogène, exceptée le proton H6 des pyrimidines. Sa position sur le cycle, et l'orientation de la liaison Donneur-Proton qui pointe vers le cycle du sucre, se prête mal à la formation de liaisons hydrogène. Les différents groupes donneurs définis pour chaque type de base sont reportés dans le tableau IV.1. Lors du calcul des cartes de liaisons hydrogène, tous les couples possibles sont traités avec la même fonction de score. Cependant, lors de l'analyse, deux classes de liaisons hydrogène sont distinguées :

- Les liaisons hydrogène ordinaires qui impliquent un proton porté par un atome d'azote ou d'oxygène. Ces liaisons hydrogène doivent être les plus stables

du fait de la forte polarité de la liaison entre le proton et l'atome fortement électronégatif auquel il est lié.

- Les liaisons hydrogène de type C-H$\cdots$O qui impliquent un proton lié à un carbone. Ces liaisons hydrogène sont de plus faible énergie à cause de la faible polarisation de la liaison C-H, et peuvent aussi contribuer à la stabilisation d'une conformation.

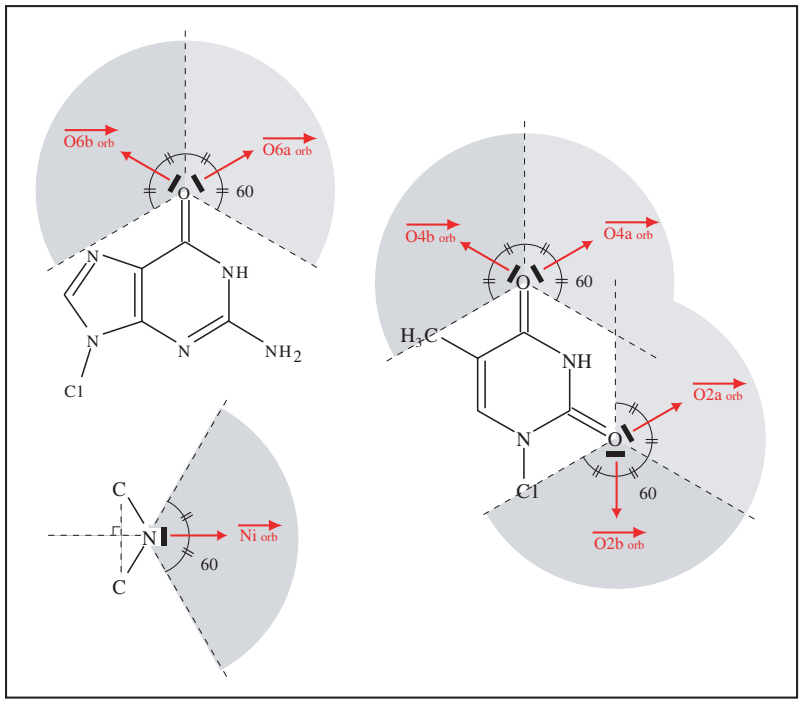


Fig. IV.3 : ***Schéma des différents groupements accepteurs de proton et de la direction des orbitales acceptrices qui leur sont associées.***

Tout atome du cycle de la base présentant un ou plusieurs doublets libres d'électrons est considéré comme accepteur possible de liaison hydrogène. Les différents groupements accepteurs de protons sont répertoriés en fonction de la nature des bases dans le tableau IV.1. Lorsque l'atome accepteur est un oxygène, deux orbitales acceptrices sont considérées, correspondant respectivement aux deux orbitales d'électrons libres portées par l'oxygène sp^2. La figure II.13 schématise les directions et les nomenclatures définies pour différencier les différentes orbitales acceptrices de toutes les bases.

IV.1.4.2 Test de toutes les liaisons hydrogène possibles

Pour un couple de nucléotides de séquence donnée, le nombre de liaisons hydrogène qui peuvent potentiellement se former dépend du nombre de groupements donneurs et accepteurs portés par chaque base. Selon les séquences, le nombre de liaisons hydrogène à tester dans chacune des 481^2=231361 conformations étudiées varie de 18 à 32 (*cf.* TAB. : IV.2).

5′ \ 3′	ADE	GUA	CYT	THY
ADE	**12**+*12*=24	**14**+*14*=28	**12**+*9*=21	**11**+*17*=28
GUA	**14**+*14*=28	**16**+*16*=32	**14**+*10*=24	**12**+*20*=32
CYT	**12**+*9*=21	**14**+*10*=24	**12**+*6*=18	**11**+*13*=24
THY	**11**+*17*=28	**12**+*20*=32	**11**+*13*=24	**8**+*24*=32

TAB. IV.2 : ***Nombre de liaisons hydrogène à tester en fonction du type des bases impliquées dans l'appariement potentiel :*** *En fonction de la nature A, G, C ou T des bases en 5' et 3' de la boucle le tableau donne en gras le nombre de liaisons hydrogène classiques (N-H$\cdots$N, N-H$\cdots$O), en italique le nombre de liaisons hydrogène impliquant un groupement donneur de type C-H, et le total des liaisons hydrogène testées.*

IV.1.4.3 Exploration de tous les appariements possibles

Pour chaque séquence, nous générons la banque de 231361 conformations de dinucléotides pour étudier toutes les conformations des deux bases en faisant varier les valeurs Ω_1 et Ω_3 associées aux deux nucléotides. Nous explorons individuellement pour chaque couple de valeurs Ω_1 et Ω_3 la formation de chacune des 18 à 32 liaisons hydrogène possibles selon la séquence. Pour chaque couple (Ω_1, Ω_3), le score de la liaison hydrogène considérée est calculé. Ceci permet d'établir des cartes de contours de score de liaison hydrogène en fonction de Ω_1 et Ω_3.

IV.1.5 Choix des meilleures liaisons hydrogène par intégration des scores de liaison hydrogène

Les cartes de score de liaison hydrogène sont l'outil principal pour déterminer les meilleures liaisons hydrogène. Le pas d'exploration de 0,5° en Ω_1 et Ω_3 de ces cartes est suffisamment fin pour obtenir une surface continue par interpolation du

premier ordre. Cette surface définit un pic de score pour chaque liaison hydrogène. Le calcul du volume du pic définit le volume de score de liaison hydrogène. Au moyen de ce volume, qui intègre le score de liaison hydrogène en fonction des variations des valeurs de Ω_1 et Ω_3 (*cf.* EQ.: IV.1.5.2), nous pouvons évaluer la stabilité de chaque liaison hydrogène en tenant compte de l'impact de faibles variations conformationnelles de Ω_1 et Ω_3. Nous considérons que les meilleures liaisons hydrogène présentent les pics de plus forts volumes. En effet, le volume tient compte à la fois de :

- la qualité de la géométrie de la liaison hydrogène (score en fonction de Ω_1 et Ω_3)
- la stabilité de la liaison hydrogène pour des variations conformationnelles (intégration du score en fonction de Ω_1 et Ω_3).

En outre, chaque pic admet un maximum définit par un couple ($\Omega_{1,max}$,$\Omega_{3,max}$). On notera pour la suite que ce couple définit la meilleure géométrie possible pour une liaison hydrogène donnée. Cette conformation du dinucléotide (N_1, N_3) sera utilisée comme point de départ pour la construction des conformations complètes (*cf.* PART.: IV.3.4).

Le volume de score est calculé par:

$$Vol = \int_{-120°}^{+120°} \int_{-120°}^{+120°} score(\Omega_1, \Omega_3) \partial\Omega_1 \partial\Omega_3 \qquad \text{(IV.1.5.2)}$$

La fonction de score est un produit d'exponentielles de fonctions quadratiques qui ressemblent à un facteur de normalisation près à des fonctions de densité de probabilité Gaussiennes. Le volume de score de liaison hydrogène est approximativement proportionnel à une probabilité de formation de liaison hydrogène.

Lors de la première analyse, seules les liaisons hydrogène de type N-H$\cdots$O ou N-H$\cdots$N sont prises en compte. Pour chaque conformation retenue, on recherche ensuite la présence éventuelle de liaison hydrogène de type C-H$\cdots$O ou C-H$\cdots$N (*cf.* TAB. : IV.1).

Ce critère n'est cependant pas assez discriminant. Pour obtenir des résultats concordants aux structures publiées nous avons dû tenir compte du caractère défavorable des rotations de fortes valeurs de Ω et $\Theta_{empil.}$.

IV.1.6 Choix de torsion et de flexion minimales

Il s'agit de pondérer le score de liaison hydrogène en chaque point de la courbe par une fonction qui pénalise l'interaction d'autant plus fortement qu'il a fallu déformer la structure initiale par rotation en Ω_i ou Θ_i, avec i=1 ou 3. Ceci revient à considérer que toute rotation en Ω ou en Θ_{empil} est corrélée à un coût énergétique qui est d'autant plus important que les angles de rotations sont forts. L'expression de la fonction de filtre utilisée est la suivante :

$$Filtre\,(\Omega, \Theta_{empil}) = e^{-\frac{\pi}{2}\left(\Omega_1^2 + \Omega_3^2 + \Theta_1^2 + \Theta_3^2\right)}$$

La forme de cette fonction filtre est donnée dans la figure IV.4. Elle est quasiment indépendante de la nature de la base (Θ_{empil} étant non dépendant du type de la base. *cf.* PART. : IV.1.3), et ne dépend que de la position relative de la trajectoire de la boucle par rapport au dernier plateau de paire de bases de la tige. Cette courbe ne varie donc que si l'on change la longueur de la boucle, ou la géométrie de la tige (conformation en hélice A ou B).

Au moyen de cette fonction de filtre, il est possible de réévaluer le score de liaison hydrogène en chaque point, en multipliant le score natif au résultat de l'évaluation de la fonction de filtre au point de coordonnées (Ω_1, Ω_3). La fonction de score de liaison hydrogène devient :

$$S_H\,(\Omega_1, \Omega_3) = S_H \times Filtre(\Omega_1, \Omega_3)$$

Cette pondération des scores de liaisons hydrogène modifie la forme des surfaces interpolées et des pics associés. La position du maximum ($\Omega_{1,max}$,$\Omega_{3,max}$) est déplacée globalement vers le maximum de la fonction de filtre (*cf.* FIG.: IV.4), et le volume est réduit. De façon pratique les volumes loin du maximum de la fonction de filtre sont fortement réduits ce qui peut changer l'ordre des meilleures liaisons hydrogènes.

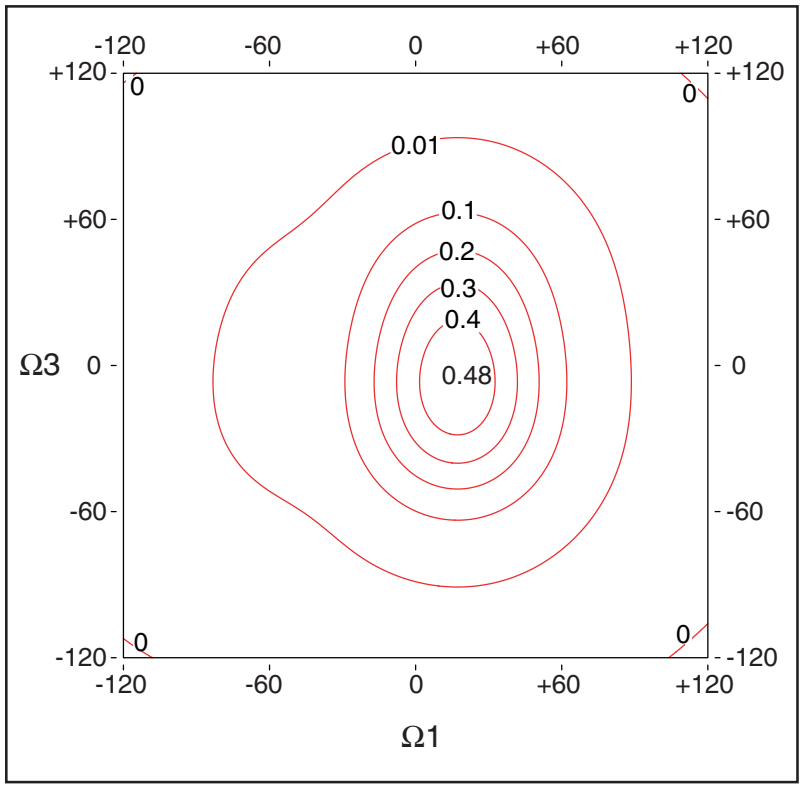


Fig. IV.4 : ***Fonction de filtre de liaison hydrogène en fonction de*** Ω_1 ***et*** Ω_3 ***et*** Θ_1 ***et*** Θ_3 ***:*** *Courbes de niveaux de la fonction de filtre en fonction des paramètres* Ω_1 *et* Ω_3 *et des paramètres* Θ_1 *et* Θ_3 *qui sont respectivement déduits des valeurs de* Ω*. Les valeurs des courbes de niveaux sont mentionnées sur la courbe, et la valeur centrale représente le maximum de la fonction,* i.e. *la pénalisation minimale.*

IV.2 Analyse des cartes de liaisons hydrogène et des appariements identifiés

IV.2.1 Multiplicité, complexité et contrôle des données des cartes de liaisons hydrogène

Pour chacune des 64 explorations différentes notre protocole produit donc entre 18 et 32 cartes différentes correspondant à chaque liaison hydrogène potentielle. Afin d'en faciliter le traitement nous construisons des "cartes sommes" uniques pour chaque exploration sur lesquelles sont représentées toutes les cartes individuelles. Ce type de carte somme permet de visualiser l'ensemble des pics sur une même carte. Les cartes présentées (*cf.* Fig. : IV.8 à IV.17) donnent la forme des pics avant filtrage. La position du maximum de chaque pic avant filtrage est pointée par une barre suivie de

la mention des atomes formant la liaison hydrogène, du volume du pic avant filtrage et du volume du pic après filtrage. Il est ainsi possible de visualiser rapidement les liaisons hydrogène potentielles d'intérêt. Sur ces cartes, les informations concernant les liaisons hydrogène de forte intensité sont marquées en bleu-gras, alors que les informations concernant les liaisons hydrogène de type C-H$\cdots$O ou C-H$\cdots$N, qui sont évaluées de la même façon mais qui sont de moindre intensité sont marquées en noir-italique. Cet outil graphique, permet d'identifier rapidement les liaisons hydrogène potentielles de chaque type, de les "localiser" sur la carte.

Afin de compléter l'analyse de ces explorations, nous produisons des tableaux de liaisons hydrogène qui reprennent les informations principales des cartes. Ces tableaux donnent les trois meilleurs pics de liaison hydrogène pour chaque exploration, ainsi que les paramètres de reconstruction (Ω_1, Ω_3) aux maxima de ces pics, le volume des pics, avant et après filtrage.

Finalement, le dernier outil introduit pour contrôler ces explorations est le calcul automatique des structures appariées N_1-N_3 aux maxima des pics de liaisons hydrogène. La représentation de ces structures permet d'effectuer un contrôle visuel sur l'appariement, et notamment de vérifier la qualité de la géométrie des conformations. En effet, la fonction de score élémentaire que nous utilisons ne tient compte que des encombrements stériques impliquant les atomes de la liaison hydrogène. Il est possible que certains pics de forts volumes correspondent à des géométries particulières de positionnement des bases où les critères de directionnalité et d'éloignement d'une liaison hydrogène sont satisfaits, mais dans lesquelles l'orientation des bases est telle que certaines parties des cycles des deux bases se recouvrent. La vérification visuelle, à ce stade de développement, permet de discriminer de telles conformations.

Dans ce chapitre nous nous limiterons à l'étude détaillée des explorations des appariements A$\cdots$A, G$\cdots$A, A$\cdots$C et G$\cdots$C pour lesquelles des structures expérimentales sont disponibles.

IV.2.2 Les appariements rencontrés dans les structures publiées.

IV.2.2.1 Appariement A···A

Les explorations portant sur l'appariement A···A mettent en évidence deux liaisons hydrogène de fort volume (*cf.* Fig. : IV.8 et Tab. : IV.4) respectivement dans les explorations Anti/Anti et Anti/Syn, et aucun pic dans les explorations Syn/Anti et Syn/Syn (*cf.* Fig. : IV.9 et Tab. : IV.4).

Avant Filtrage				**Après Filtrage**			
Liaison Hydrogène	Ω_1	Ω_3	**Vol.**	**Liaison Hydrogène**	Ω_1	Ω_3	**Vol.**
Anti-Anti							
A1 N3 - A3 HN6A	30°	5,5°	244,2	**A1 N3 - A3 HN6A**	**20,5°**	**2,5°**	**86,9**
A1 N7 - A3 HN6A	-77,5°	-58,5°	137,3	A1 N3 - A3 HN6A	20,5°	38°	7,6
A3 N3 - A1 HN6A	-41,°	46,5°	136,5	A1 N1 - A3 HN6B	-41,5°	43°	2,3
Anti-Syn							
A1 N1 - A3 HN6A	7°	32,5°	194,9	**A1 N1 - A3 HN6A**	**7,5°**	**32,5°**	**50,6**
A1 N3 - A3 HN6B	49,5°	2°	151,8	A1 N3 - A3 HN6B	45,5°	1°	32,7
A3 N7 - A1 HN6B	-18,5°	54,5°	64,3	A3 N7 - A1 HN6B	-17,5°	52,5°	3,6
Syn-Anti							
A3 N1 - A1 HN6A	-13,5°	45°	123,5	A1 N7 - A3 HN6B	13°	27°	21,2
A1 N1 - A3 HN6A	-86,5°	-70,5°	117,6	A3 N1 - A1 HN6A	-13°	44°	12,7
A3 N7 - A1 HN6B	-58,°	-56,5°	74,1	*A3 N7 - A1 HN6A*	*-22,5°*	*18,°*	*4,8*
Syn-Syn							
A3 N7 - A1 HN6A	-22°	41,5°	114,9	*A3 N1 - A1 HN6A*	*-14,5°*	*10,°*	*11,6*
A1 N3 - A3 HN6A	-98,5°	-23°	62,9	A3 N7 - A1 HN6A	-21,5°	40,5°	9,4
A3 N1 - A1 HN6A	-15°	10°	53,9	A1 N7 - A3 HN6A	2°	24,5°	7,5

Tab. IV.4 : ***Appariements ADE···ADE - Volumes de score de liaison hydrogène et paramètres*** Ω_1 ***et*** Ω_3 ***des trois pics de plus fort volume :*** *Pour chaque exploration en fonction de la conformation des bases formant l'appariement, ce tableau donne les trois pics de plus fort volume, la description des atomes impliqués dans la liaison hydrogène et les paramètres de construction* Ω_1 *et* Ω_3 *correspondant au maximum du pic. À gauche avant le filtrage et à droite après filtrage. En gras, les pics de plus fort volume après filtrage et en italique les conformations présentant de forts conflits stériques portant sur plusieurs atomes.*

Dans l'exploration Anti/Anti, la liaison A1 N3···HN6A A3, dont les valeurs de Ω_1 et Ω_3 au maximum du pic sont respectivement de 20,5° et 2,5°, présente un volume de 86,9 après filtrage, ce qui est le pic de volume de score le plus important

de toutes les explorations de cet appariement. Dans l'exploration ANTI/SYN, la liaison A1 N1···HN6A A3, dont les valeurs de Ω_1 et Ω_3 au maximum du pic sont respectivement de 7,5° et 32,5°, avec un volume de 50,6 semble être une alternative possible. L'analyse des conformations associées à ces liaisons hydrogène montre cependant que la liaison en conformation ANTI-SYN doit être moins favorable pour plusieurs raisons :

- En conformation SYN, l'adénine en 5' est beaucoup plus tourné vers le grand sillon (Ω_3=32° en SYN contre 2,5° en ANTI) ce qui correspond à un moins bon empilement sur le dernier plateau de l'hélice figuré en clair sur la figure IV.8.
- La liaison en conformation ANTI est renforcée par une liaison hydrogène de type C-H···N, la liaison A1 H2···N7 A3, dont les valeurs de Ω_1 et Ω_3 au maximum du pic sont respectivement de 20° et 13,5°.

Il semble donc que cette dernière doive être préférée. La dernière confirmation est que cette liaison hydrogène (A1 N3···HN6A A3) est la liaison décrite dans la publication [47] de la structure 1BJH-AAA (*cf.* TAB. : I.6), pour expliquer la formation d'un appariement stable dans la boucle.

IV.2.2.2 Appariement G···G

Les explorations de l'appariement G···G pointent plusieurs liaisons hydrogène dans les explorations ANTI/ANTI et ANTI/SYN (*cf.* FIG. : IV.10 et TAB. : IV.6). Aucun pic de fort volume n'est relevé dans les explorations SYN/ANTI et SYN/SYN (*cf.* FIG. : IV.11 et TAB. : IV.6).

De toutes les explorations, la liaison hydrogène G1 HN2B···N7 G3 de l'exploration ANTI-ANTI est celle dont le pic est de plus fort volume. Les valeurs de Ω_1 et Ω_3 au maximum du pic sont respectivement de 28,5° et 10,5°, et le pic présente un volume de score de 99,3. La conformation au maximum du pic montre que l'empilement est très bon (*cf.* FIG. : IV.10).

Dans l'exploration ANTI/SYN deux liaisons hydrogène apparaissent : les liaisons G1 N3···G3 H1 et G1 N3···HN2A G3, dont les valeurs de Ω_1 et Ω_3 aux maxima des pics sont respectivement de 33° et 29,5° pour la première guanine et -10,5° et -26° pour la dernière. Les volumes de score sont respectivement de 81,3 et 65,4. L'analyse

Avant Filtrage				Après Filtrage			
Liaison Hydrogène	Ω_1	Ω_3	**Vol.**	**Liaison Hydrogène**	Ω_1	Ω_3	**Vol.**
ANTI-ANTI							
G3 N7 - G1 HN2B	30,5°	11,5°	272	**G3 N7 - G1 HN2B**	**28,5°**	**10,5°**	**99,3**
G1 N3 - G3 HN2A	39°	43,5°	158,6	G1 N3 - G3 HN2A	36,5°	42°	22,1
G1 N7 - G3 HN2B	-39°	60°	149,9	G1 N3 - G3 H1	39,5°	29°	11,8
ANTI-SYN							
G1 N3 - G3 H1	37°	-10°	240,6	**G1 N3 - G3 H1**	**33°**	**-10,5°**	**81,3**
G1 N3 - G3 HN2A	32°	-26°	204,1	**G1 N3 - G3 HN2A**	**29,5°**	**-26°**	**65,4**
G3 O6b - G1 H1	7,5°	32,5°	174,9	G3 O6b - G1 H1	8°	32,5°	45,6
SYN-ANTI							
G3 O6b - G1 HN2A	-106,5°	-70°	129,	G1 N7 - G3 H1	6,5°	37°	18,8
G1 O6b - G3 HN2A	-11°	64,5°	114,7	G1 O6b - G3 H1	-10,5°	43,5°	13,7
G1 O6b - G3 H1	-11,5°	44,5°	112,4	G1 N7 - G3 HN2A	6°	51,5°	10,2
SYN-SYN							
G3 N7 - G1 HN2B	-106,5°	-22,5°	77,5	G1 N7 - G3 HN2A	1°	-17,5°	20,4
G1 O6b - G3 HN2A	-15,5°	-9°	74,7	*G1 O6b - G3 HN2A*	*-15°*	*-9°*	*17,7*
G1 N3 - G3 H1	-108°	-63°	67,2	*G1 O6b - G3 H1*	*-12,5°*	*6°*	*13,4*

TAB. IV.6 : ***Appariements GUA$\cdots$GUA - Volumes de score de liaison hydrogène et paramètres Ω_1 et Ω_3 des trois pics de plus fort volume :***
cf. TAB. : *IV.4.*

des conformations et des valeurs des Ω, montre que les deux liaisons hydrogène de l'exploration ANTI-SYN n'en forment en fait qu'une seule bifurquée (*cf.* FIG. : IV.10). L'accepteur est dans les deux cas le N3 de la première guanine qui est placée dans une conformation semblable avec des valeurs de Ω_1 qui sont respectivement de 33° et 29,5°. L'accepteur du proton est alternativement le H1 ou le HN2A de la seconde guanine qui sont aussi dans des conformations très proches, avec des valeurs de Ω_3 qui sont respectivement de -10,5° et -26°. Du fait de la présence de cette liaison bifurquée, la conformation moyenne entre ces deux liaisons hydrogène pourrait être considérée comme plus stable que la liaison unique de la conformation ANTI-SYN. Cependant, 2 éléments semblent pouvoir déstabiliser cette liaison : D'une part, l'empilement de la guanine en 3' est moins bon que dans le cas de la conformation ANTI/ANTI, d'autre part, la formation de la liaison G1 N3$\cdots$G3 H1 qui semble s'accompagner d'une faible gêne stérique entre les hydrogènes des groupements amines des deux bases. En effet, notre approche simple, ne tient pas compte de l'encombrement des atomes. Ici les conformations sont favorables du point de vue des liaisons hydrogène , mais probablement déstabilisées du point de vue stérique.

Donc, le meilleur empilement de la conformation ANTI-ANTI et la déstabilisation stérique probable de la liaison bifurquée nous conduisent à choisir la première liaison hydrogène.

À ce jour, aucune structure en tri-boucle d'ADN comportant cet appariement n'a été résolue. Cependant, dans une analyse comparative des spectres RMN des séquences g(gtac-AAA-gtac) et g(gtac-GAG-gtac), il a été dit [47] que ces deux séquences se structurent de la même façon. La structure AAA ayant été résolue, une hypothèse de formation de liaison hydrogène entre le proton G1 HN2B et l'accepteur N7 G3 a été proposée pour expliquer la stabilité de l'épingle à cheveux de séquence GAG dans la boucle. Cette liaison hydrogène est celle que nous venons d'identifier ci-dessus.

IV.2.2.3 Appariement G···A

Avant Filtrage				Après Filtrage			
Liaison Hydrogène	Ω_1	Ω_3	**Vol.**	**Liaison Hydrogène**	Ω_1	Ω_3	**Vol.**
ANTI-ANTI							
A3 N7 - G1 HN2B	31°	11,5°	277,4	**A3 N7 - G1 HN2B**	**29°**	**10,5°**	**103,**
G1 N3 - A3 HN6A	31,5°	5,5°	247,4	**G1 N3 - A3 HN6A**	**22°**	**2,5°**	**88,3**
G1 N7 - A3 HN6A	-78,5°	-58,5°	139,4	*G1 O6a - A3 HN6B*	*2,5°*	*39°*	*12,2*
ANTI-SYN							
A3 N1 - G1 HN2B	48°	4°	154,	A3 N1 - G1 HN2B	46°	3,5°	32,1
G1 N3 - A3 HN6B	49,5°	1,5°	148,	G1 N3 - A3 HN6B	46,5°	1°	30,8
G1 O6a - A3 HN6A	-7,5°	38,5°	122,5	A3 N7 - G1 HN2B	34°	30°	23,7
SYN-ANTI							
G1 O6b - A3 HN6B	-5,5°	32,5°	82,1	G1 N7 - A3 HN6B	15,5°	27°	23,1
G1 N7 - A3 HN6A	-57,5°	-57°	76,4	G1 O6b - A3 HN6B	-5°	32°	17,8
G1 N7 - A3 HN6B	15°	27°	71,3	*G1 O6b - A3 HN6A*	*-10,5°*	*19,5°*	*0,6*
SYN-SYN							
G1 N3 - A3 HN6A	-98°	-22,5°	83,6	G1 O6b - A3 HN6A	-13°	29°	9,3
A3 N7 - G1 HN2B	-106,5°	-24°	80,1	G1 N7 - A3 HN6A	4,5°	24,5°	8,1
G1 O6b - A3 HN6A	-13,5°	34,5°	56,3	G1 O6b - A3 HN6B	-4,5°	15,5°	7,1

TAB. IV.8 : ***Appariements GUA···ADE - Volumes de score de liaison hydrogène et paramètres*** Ω_1 ***et*** Ω_3 ***des trois pics de plus fort volume :***
cf. TAB. : *IV.4.*

Les explorations concernant l'appariement G···A identifient seulement deux pics de fort volume (*cf.* FIG. : IV.12 et TAB. : IV.8). Les deux sont dans l'exploration ANTI/ANTI. Aucun pic de fort volume de score n'est trouvé dans les

explorations ANTI/SYN, SYN/ANTI et SYN/SYN (*cf.* FIG. : IV.12&IV.13 et TAB. : IV.8). Ces deux pics correspondent aux liaisons hydrogène G1 HN2B···N7 A3 et G1 N3···HN6A A3. Ils présentent des volumes respectifs de 103 et 88,3. Ils apparaissent dans les mêmes gammes de Ω, avec de valeurs de Ω_1 et Ω_3 aux maxima des pics qui sont respectivement de 29° et 22° pour la guanine et 10,5° et 2,5° pour l'adénine. La similarité des valeurs en Ω pour une base donnée indique que les conformations sont proches. Ainsi, la proximité des maxima des pics, ainsi que le recouvrement des deux pics sur la carte (*cf.* FIG. : IV.12), indiquent que ces deux liaisons hydrogène peuvent se former simultanément. En outre, cette conformation présente un bon empilement sur le dernier plateau de paire de bases de la tige comme le montre les conformations de la figure IV.12. Cette conformation semble donc être très favorable à la formation d'un appariement stable dans la boucle.

La présence de ces deux liaisons hydrogène est décrite comme stabilisant conjointement l'appariement dans toutes les structures publiées [55, 56, 62, 63], 1XUE-GCA, 1ZHU-GCA, 1JVE-GAA et 1PQT-GAA (*cf.* TAB. : I.6).

IV.2.2.4 Appariement A···C

Les explorations de la séquence A···C donnent un pic de fort volume de score de liaison hydrogène pour les conformations ANTI/ANTI et ANTI/SYN (*cf.* FIG. : IV.14 et TAB. : IV.10). Aucun pic de fort volume n'est trouvé dans les explorations SYN/ANTI et SYN/SYN (*cf.* FIG. : IV.15 et TAB. : IV.10).

Le pic de plus fort volume concerne la conformation ANTI/ANTI avec un volume de 117,8 contre 101,6 dans l'exploration ANTI/SYN. Dans la première exploration, la liaison hydrogène A1 N3···HN4A C3 présente des valeurs de Ω_1 et Ω_3 au maximum du pic qui sont respectivement de 28° et 2°. Dans la seconde exploration (ANTI/SYN) la liaison hydrogène A1 N3···HN4B C3 identifiée présente des valeurs de Ω_1 et Ω_3 au maximum du pic qui sont respectivement de 30,5° et 6°. Ces deux liaisons hydrogène correspondent à un placement des bases 5' et 3' identique comme le montre la similarité des valeurs de Ω_1 et Ω_3. Pour passer d'une conformation à l'autre il suffit de tourner la base autour de la liaison glycosidique de 180°, ce qui correspond au passage ANTI↔SYN. Dans notre approche d'exploration, aucune pénalité n'est donnée à la rotation de l'angle χ, aux gênes stériques (absentes des conformations observées) ou à la qualité de l'empilement (bon dans les deux cas : *cf.* les conformations dans FIG. : IV.14). C'est un modèle très simple d'exploration

Avant Filtrage				Après Filtrage			
Liaison Hydrogène	Ω_1	Ω_3	**Vol.**	**Liaison Hydrogène**	Ω_1	Ω_3	**Vol.**
ANTI-ANTI							
A1 N3 - C3 HN4A	32°	3°	294,6	**A1 N3 - C3 HN4A**	**28°**	**2°**	**117,8**
C3 O2b - A1 HN6A	-43°	51,5°	109,1	A1 N1 - C3 HN4B	10°	36,5°	24,8
A1 N1 - C3 HN4B	9,5°	37°	98,5	C3 N3 - A1 HN6B	-17°	47°	4,8
ANTI-SYN							
A1 N3 - C3 HN4B	35,5°	7°	267,2	**A1 N3 - C3 HN4B**	**30,5°**	**6°**	**101,6**
A1 N1 - C3 HN4A	5°	40,5°	146,5	A1 N1 - C3 HN4A	9°	40°	31,3
A1 N7 - C3 HN4B	-78,5°	-58,5°	91,5	A1 N3 - C3 HN4A	39,5°	23,5°	1,4
SYN-ANTI							
C3 O2a - A1 HN6A	-28,5°	62°	108,6	A1 N7 - C3 HN4B	1°	25°	27,5
C3 N3 - A1 HN6A	-22,5°	40°	104	C3 N3 - A1 HN6A	-21,5°	39°	8
A1 N7 - C3 HN4B	-0,5°	25°	94,3	C3 O2a - A1 HN6A	-28°	58°	2,8
SYN-SYN							
A1 N7 - C3 HN4A	-1,5°	28°	123,1	A1 N7 - C3 HN4A	0°	28°	33,6
C3 O2a - A1 HN6A	-30,5°	-5,5°	86,4	*C3 O2a - A1 HN6A*	*-30°*	*-4,5°*	*8,1*
A1 N3 - C3 HN4A	-104°	-17,5°	46,1	*C3 N3 - A1 HN6A*	*-22,5°*	*18°*	*4*

TAB. IV.10 : ***Appariements ADE···CYT - Volumes de score de liaison hydrogène et paramètres*** Ω_1 ***et*** Ω_3 ***des trois pics de plus fort volume :***
cf. TAB. : *IV.4.*

où la taille du pic ne permet pas de choisir entre les deux conformations identifiées. Ainsi, la liaison la plus probable est la liaison de la conformation ANTI/ANTI.

Cette liaison hydrogène correspond à la conformation et à la liaison hydrogène décrite dans les publications [65, 67] des structures ATC et AGC comme facteur de stabilisation de l'appariement A-C dans ces boucles (*cf.* TAB. : I.6).

IV.2.2.5 Appariement G···C

Pour l'appariement G-C, l'analyse des cartes est plus délicate. Quatre pics de fort volume sont identifiés. Un pic dans l'exploration ANTI/ANTI, et 3 pics dans l'exploration ANTI/SYN (*cf.* FIG. : IV.16 et TAB. : IV.12). Aucun pic de fort volume de score n'est trouvé dans les explorations SYN/ANTI et SYN/SYN (*cf.* FIG. : IV.17 et TAB. : IV.12).

Dans l'exploration ANTI/ANTI, la liaison hydrogène identifiée est la liaison G1 HN2B···N7 G3 avec un volume de 117,1. Dans l'exploration ANTI/SYN,

les liaisons hydrogène identifiées sont, G1 HN2B···O2a C3, G1 HN2B···N3 C3 et G1 N3···HN4B C3 avec des volumes respectifs de 131,3, 114,6 et 100,4. L'identification du meilleur appariement ne peut donc être trivialement interprétée à partir du volume des pics principaux. Plusieurs critères permettent cependant de faire un choix :

La conformation au maximum du pic de l'exploration ANTI/ANTI est bien empilée sur le dernier plateau de paire de bases de la tige. La conformation au maximum du pic de plus fort volume de l'exploration ANTI/SYN est moins bien empilée comme le montre la position de la cytosine placée-tournée vers le petit sillon de l'hélice. La conformation ANTI/ANTI semble à ce point être un meilleur candidat que la conformation ANTI/SYN.

Avant Filtrage				**Après Filtrage**			
Liaison Hydrogène	Ω_1	Ω_3	**Vol.**	**Liaison Hydrogène**	Ω_1	Ω_3	**Vol.**
ANTI-ANTI							
G1 N3 - C3 HN4A	32,5°	2,5°	292,8	**G1 N3 - C3 HN4A**	**28°**	**1,5°**	**117,1**
C3 O2a - G1 HN2B	23,5°	50,5°	134,8	C3 O2a - G1 HN2B	22,5°	47,5°	23,8
C3 N3 - G1 H1	-5,5°	47°	113,9	G1 O6a - C3 HN4B	-5,5°	38,5°	17,3
ANTI-SYN							
C3 O2a - G1 HN2B	8,5°	-20,5°	335,9	**C3 O2a - G1 HN2B**	**10,5°**	**-17,5°**	**131,3**
C3 N3 - G1 HN2B	30°	11,5°	283,4	**C3 N3 - G1 HN2B**	**27,5°**	**10°**	**114,6**
G1 N3 - C3 HN4B	36°	6,5°	265	**G1 N3 - C3 HN4B**	**31°**	**5,5°**	**100,4**
SYN-ANTI							
G1 N7 - C3 HN4B	2,5°	25°	96,9	G1 N7 - C3 HN4B	3,5°	24,5°	30,5
C3 N3 - G1 HN2B	-108°	-22°	92,	G1 O6b - C3 HN4B	-11,5°	32°	13,
G1 O6b - C3 HN4B	-12,5°	32,5°	73,8	G1 O6b - C3 HN4A	-12°	17,5°	3,1
SYN-SYN							
G1 N7 - C3 HN4A	1°	28°	127,2	G1 N7 - C3 HN4A	2°	27,5°	37,7
G1 O6b - C3 HN4A	-13,5°	36,5°	80,5	G1 O6b - C3 HN4A	-12,5°	35,5°	12,1
G1 N3 - C3 HN4A	-104,5°	-17,5°	62,5	G1 O6b - C3 HN4B	-11°	21°	2,8

TAB. IV.12 : ***Appariements GUA···CYT - Volumes de score de liaison hydrogène et paramètres*** Ω_1 ***et*** Ω_3 ***des trois pics de plus fort volume :***
cf. TAB. : *IV.4.*

Cependant, comme dans le cas de l'appariement GANTI-AANTI, deux pics de l'exploration ANTI/SYN sont très proches en Ω et se recouvrent partiellement. Il s'agit des pics associés aux liaisons hydrogène G1 HN2B···N3 C3 et G1 N3···HN4B C3, dont les valeurs de Ω_1 et Ω_3 sont respectivement 27,5° et 31° pour la guanine et 10° et 5,5° pour la cytosine. Cette proximité des valeurs

de Ω indique que les conformations associées à ces deux liaisons hydrogène sont très proches, et que celles-ci doivent pouvoir se former simultanément. Deux liaisons hydrogène offrant plus de stabilité à la structure qu'une seule, le meilleur appariement pour G···C fait intervenir deux liaisons hydrogène dont la somme des volumes de score est de loin supérieure aux autres volumes isolés.

Ces deux liaisons hydrogène, avec la cytosine en conformation SYN, sont les liaisons identifiées dans la structure publiée 1P0U-GAC [64], comme stabilisant l'appariement dans la boucle publiée (*cf.* TAB. : I.6).

IV.3 Analyse des structures BCE$_{opt}$

IV.3.1 Isomorphie des appariements

Les appariements A···A, G···G et G···A identifiés par nos explorations sont isomorphes (*cf.* FIG. : IV.5), c'est-à-dire que l'on peut passer de l'un à l'autre en changeant une purine par une autre tout en conservant les positions des cycles. Cette substitution permet, malgré la variation de la nature de la base, d'établir d'autres liaisons hydrogène de qualité équivalente. On observe de la même façon une isomorphie entre les appariements A···C et G···C.

Ces isomorphies sont mises en évidence par le calcul des distances entre les C1' des deux nucléotides qui forment l'appariement, et par le calcul des valeurs des angles $(C1' - \widehat{C1' - N(1/9)})$ qui traduisent l'orientation réciproque des bases (*cf.* FIG. : **IV.5**). Il est remarquable de constater la similarité de construction des appariements Pu-Pu et Pu-Py tant du point de vue des valeurs des distances entre atomes C1' (compris entre 8,07Å et 8,20Å) que des angles $(C1' - \widehat{C1' - N(1/9)})$, qui varient entre 102,3° et 108,5° pour la base en 5' et entre -1,8° et 7,8° pour la base en 3'. Le remplacement en 3' de la boucle d'une purine par une cytosine ne semble pas affecter la géométrie de l'empilement ni l'orientation de la base en 3' de la boucle, puisque les angles $(C1' - \widehat{C1' - N(1/9)})$ restent proches.

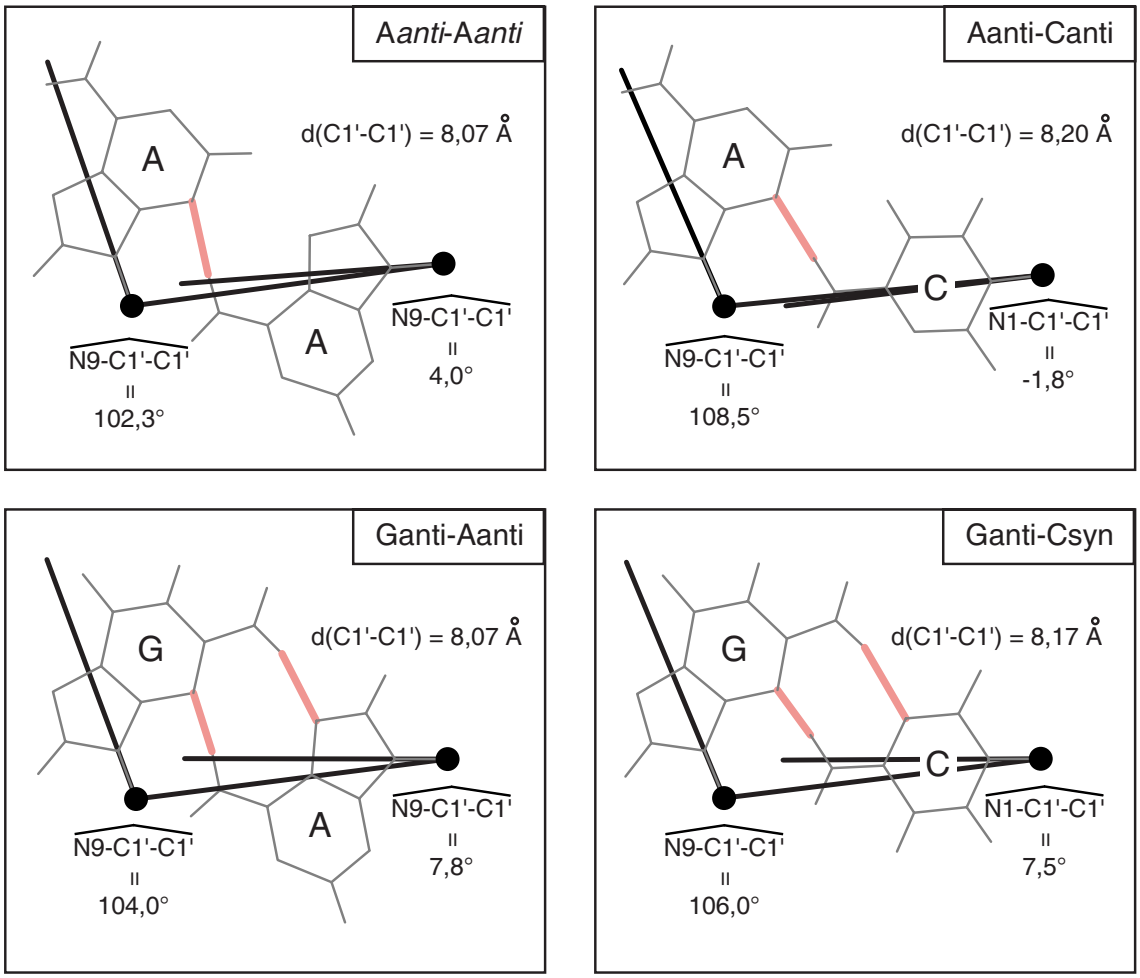


Fig. IV.5 : ***Isomorphies des appariements A···A/G···A et A···C/G···C :*** *Représentation des distances entre atomes C1' en Angström (Å) et des angles C1'-C1'-N(1/9) en degré des appariements théoriques. Superposés, les appariements isomorphes. Côte à côte, les appariements comportant le même nombre de liaisons hydrogène. Les atomes C1' sont figurés par des cercles noirs. En gras noir : les directions des liaisons C1'-N1 et C1'-N9. En gras rose : les liaisons hydrogène. Les conformations théoriques des appariements comportant deux liaisons hydrogène sont obtenues en moyennant les valeurs* Ω_1 *et* Ω_3 *obtenues aux maxima des deux pics dans les explorations correspondantes.*

IV.3.2 Généralisation de l'isomorphie au moyen des couples (Ω_1, Ω_3)

De façon générale, s'il y a isomorphie entre deux appariements identifiés par nos explorations, alors les valeurs des couples (Ω_1, Ω_3) définies pour ces conformations sont nécessairement très proches. Ainsi, pour les appariements isomorphes Pu-Pu, A···A, G···A et G···G (*cf.* Tab. : IV.13), la purine en 5' est tournée d'une valeur de Ω_1 comprise entre 20,5° et 29° et la purine en 3' est tournée d'une valeur de Ω_3

comprise entre 2,5° et 10,5°. De la même façon, pour les appariements isomorphes Pu-Py, A···C et G···C (*cf.* TAB. : IV.13), la purine en 5' est tournée d'une valeur de Ω_1 comprise entre 27,5° et 31° et la purine en 3' est tournée d'une valeur de Ω_3 comprise entre 2° et 10°.

Nature de l'appariement	**Nature de la liaison hydrogène**	**Valeurs aux maxima des pics** Ω_1	Ω_3
Appariement Pu-Pu			
A···A	A1 N3 - HN6A A3	20,5°	2,5°
G···A	G1 HN2B - N7 A3	29°	10,5°
	G1 N3 - HN6A A3	22°	2,5°
G···G	G1 HN2B - N7 G3	28,5°	10,5°
Appariement Pu-Py			
A···C	A1 N3 - HN4A C3	28°	2°
G···C	G1 HN2B - N3 C3	27.5°	10°
	G1 N3 - HN4B C3	31°	5,5°
Moyenne et Écart type		**26,6°±3,6°**	**6,2°±3,7°**

TAB. IV.13 : ***Récapitulation des valeurs de Ω déterminées aux maxima des pics de liaison hydrogène retenus pour chaque appariement identifié :*** *Pour chaque liaison hydrogène le tableau rassemble les valeurs de Ω_1 et Ω_3 aux maxima des pics des tables précédentes. La moyenne et l'écart type sont calculés sur l'ensemble de ces valeurs.*

De plus, on observe que les valeurs des couples (Ω_1, Ω_3) sont très similaires pour les deux familles d'appariements Pu-Pu et Pu-Py. La valeur de Ω_1 de la purine en 5' est comprise entre 20,5° et 31° (26,6°±3,6°) et la valeur de Ω_3 de la base en 3' est comprise entre 2,5° et 10,5° (6,2°±3,7°). Les valeurs des angles Ω dont est tournée chaque base de l'appariement sont donc très proches en 5' et en 3' comme le confirme les faibles écart-types de 3,6° et 3,7° respectivement calculés sur les valeurs de Ω_1 et de Ω_3(*cf.* TAB. : IV.13). L'angle Ω offre donc un bon paramètre quantitatif de construction, de description et de comparaison des appariements. Il met en évidence des modes de structuration similaires que la définition stricte de l'isomorphie n'identifie pas.

IV.3.3 Comparaison des structures BCE$_{opt}$ obtenues

Les appariements Pu-Pu sont donc isomorphes, les appariements Pu-Py le sont aussi, et les appariements Pu-Pu et Pu-Py sont similaires. Par ailleurs, la base centrale de toutes ces tri-boucles s'empile systématiquement sur le plateau formé

par l'appariement de la boucle. Les valeurs Ω_i calculées pour toutes les bases de la boucle sont donc très similaires pour toutes ces molécules : Ω_1 vaut en moyenne 26,6°±3,6°, Ω_2, 78,9°±1,2° et Ω_3, 6,2°±3,7° (*cf.* TAB. : IV.13). Ces valeurs sont cohérentes avec les valeurs trouvées précédemment (*cf.* CHAPT. : III).

Les déformations selon l'angle Θ sont respectivement pour la première et la dernière base de -14,5°±2,8° et de 25,2°±5,1° ce qui reste faible et garantit une faible déformation des angles de torsion de l'hélice initiale lors de la manipulation des blocs rigides.

Les variations de l'angle χ de la liaison glycosidique des appariements ANTI/ANTI sont de 0,8°±0,5°, c'est-à-dire proche de zéro, pour la première base , de -30,3°±0,5° pour la deuxième et de 49,3°±2,2° pour la troisième (valeurs calculées sur les angles différents du tableau IV.14). La différence de la variation de l'angle χ observée pour la base terminale de la boucle GAC, qui est tournée de -126°, provient de la conformation SYN de cette base.

		5'			**Central**		**3'**		
		Ω_1	Θ_1	$\Delta\chi_1$	Ω_2	$\Delta\chi_2$	Ω_3	Θ_3	$\Delta\chi_3$
A···A	AAA-1BJH	20,5°	-19,0°	0,1°	77.9°	-30.9°	2,5°	28,5°	48,6°
G···A	GCA-1XUE-1ZHU	25,5°	-14,6°	1,3°	81,2°	-29,6°	6,5°	29,0°	52,3°
	GAA-1JVE-1PQT	”	”	”	77.8°	-30,8°	”	”	”
A···C	ATC	28,0°	-13,0°	0,6°	79,6°	-30,1°	2,0°	26,4°	47,1°
	AGC	”	”	”	78.6°	-30,3°	”	”	”
G···C	GAC-1P0U	29,0°	-11.6°	1.3°	78,3°	-30,1°	9,0°	26.9°	-126.8°

TAB. IV.14 : ***Paramètres de modélisation des tri-boucles d'ADN théoriques comportant un appariement dans la boucle :*** *Valeurs des angles Ω, χ pour les trois bases de la tri-boucle et Θ pour les bases appariées empilées sur le dernier plateau de paire de bases de la tige. Ces valeurs sont des constantes fonctions de la séquence.*

IV.3.4 Construction des structures BCE$_{opt}$ complètes

Afin de vérifier la qualité des structures appariées identifiées par nos explorations théoriques, nous avons construit les structures complètes pour les comparer aux structures expérimentales PDB ou autres.

Pour construire la structure entière, nous devons compléter la paire de bases appariée retenue avec la séquence de la tige et de la base centrale. Ceci est réalisé en

appliquant le protocole de modélisation BCE : après avoir construit une structure BCE_{ori}, les bases extrémales de la boucle sont tournées des valeurs ($\Omega_1 max$, $\Omega_3 max$) déterminées au maximum du pic de score de liaison hydrogène. Les bases sont ensuite redressées en Θ_{empil} et χ par l'opération de redressement d'empilement des bases. Le placement de la base centrale B_2 de la tri-boucle, est réalisé par la rotation du nucléotide en Ω et en χ, sans utiliser la rotation de redressement d'empilement d'angle Θ_{empil}. Les valeurs Ω_2 et χ_2, sont déterminées pour optimiser l'empilement de la base centrale sur le plateau apparié formé par les deux bases extrémales de la boucle. Dans le cas de ce nucléotide, les deux d.d.l. Ω et χ suffisent pour placer la base centrale dans une géométrie globalement coplanaire au plateau moyen défini par les deux bases de l'appariement.

À l'issue de ce protocole nous obtenons la structure BCE_{opt}. Pour finaliser la construction de la structure complète nous la minimisons pour obtenir la structure BCE_{min}.

IV.4 Analyse des structures BCE_{min}

IV.4.1 Calcul des structures BCE_{min}

La dernière étape consiste à minimiser la structure ainsi obtenue pour relaxer la conformation. Lors de cette minimisation, nous avons cherché le nombre minimal de contraintes d'angles de torsion nécessaire pour retrouver les mêmes angles de torsion que ceux de la chaîne sucre-phosphate des structures publiées (*cf.* TAB. : IV.15). L'introduction de ces contraintes, toutes au niveau de la zone du "sharp-turn" n'affecte pas la structure globale de molécule donnée par BCE. Ces contraintes permettent surtout de corriger l'angle α du troisième nucléotide de la boucle qui est systématiquement en conformation *cis* après repliement de l'hélice simple-brin sur la trajectoire donnée par la théorie de l'élasticité.

Le fichier PDB GCA-1ZHU demande un traitement particulier. Il présente deux classes de conformations qui se distinguent par la valeur de certains angles de torsion de la boucle. Dans quatre des dix conformations du fichier les angles C2(ζ) et A3 (α-β) valent respectivement -109,7°±0,1° (g^-), -94,2°±0,2° (g^-) et 57,3°±0,1° (g^+), alors que dans les six autres, les mêmes angles prennent les valeurs 130,9°±0° (t/g^+), 80°±0,2° (g^+) et -100,0°±0,2° (g^-). Les autres angles sont significativement

Appariement	Structures	Contraintes de minimisation
A···A	AAA-1BJH	(γ) A3
G···A	GCA-1ZHU #1	(γ) A3
G···A	GCA-1ZHU #2 & GCA-1XUE & GAA-1JVE & GAA-1PQT	(α, γ) A3
A···C	ATC et AGC	(γ) C3
G···C	GAC-1P0U	(α, β, γ) C3

Tab. IV.15 : ***Angles de torsion contraints lors de la minimisation d'énergie de la structure BCE_{opt} pour obtenir la structure BCE_{min}.***

semblables. À ce stade, il faut rappeler que les conformations g^+/g^- ou g^-/g^+ de deux angles de torsion consécutifs sont habituellement interdits [96]. Cependant, afin de tenir compte de cette variété conformationnelle expérimentale, nous avons choisi de construire les deux classes de conformations. Pour modéliser la première classe, GCA-1ZHU#1, dont les angles sont en conformation (g^-, g^-, g^+), nous avons introduit une contrainte unique sur l'angle de torsion A3 (α), alors que pour modéliser la deuxième classe, GCA-1ZHU#2, dont la conformation est (t/g^+, g^+, g^-), nous avons introduit deux contraintes sur les angles A3 (α, γ). Les autres fichiers PDB ne comportent qu'une seule classe de molécules à l'échelle des angles de torsion de la chaîne sucre-phosphate de la boucle. Pour ces molécules, une seule modélisation est donc suffisante.

IV.4.2 Comparaison des structures BCE_{min} et expérimentales, localement à l'échelle de la chaîne sucre phosphate et de l'appariement

La comparaison des structures reconstruites avec l'approche BCE aux structures publiées (*cf.* Tab. : IV.17 et Fig. : IV.6) montre l'excellent accord entre les modèles théoriques et les modèles expérimentaux à l'échelle locale. Cette quasi identité est confirmée par le calcul, sur les atomes formant l'appariement et sur ceux de la chaîne sucre-phosphate, de RMSd et des moyennes de déviation entre les angles de torsion.

À l'échelle de l'appariement, le RMSd varie de 0,30 Å à 0,98 Å, ce qui est très faible, comme le confirme leur superposition visuelle dans la figure III.3. De façon identique, les RMSd montrent que la chaîne sucre phosphate de la partie en boucle et de la partie tige et boucle est également très bien résolue par notre approche avec

des valeurs qui varient respectivement entre 0,35Å et 1,22Å et entre 0,20Å et 0,61 Å.

Appariement	Structure	App.	RMSd Boucle + Tige	RMSd Boucle	$\sigma_{dev}tor$ Boucle + Tige	$\sigma_{dev}tor$ Boucle
A···A	AAA-1BJH	0,87Å	0,36 Å	0,35Å	13,86°	10,81°
G···A	GCA-1XUE	0,69Å	0,43Å	0,23Å	7,09°	7,95°
	GCA-1ZHU #1	0,71Å	0,55Å	0,20Å	8,75°	9,60°
	GCA-1ZHU #2	0,69Å	0,54Å	0,22Å	8,09°	9,07°
	GAA-1JVE	0,45Å	0,62Å	0,61Å	39,33°	53,44°
	GAA-1PQT	0,51Å	1,22Å	0,59Å	16,68°	17,70°
A···C	ATC	0,57Å	0,75Å	0,41Å	13,86°	13,39°
	AGC	0,30Å	0,35Å	0,33Å	6,79°	7,03°
G···C	GAC-1P0U	0,98 Å	1,00Å	0,61Å	17,67°(1)	14,45°

Tab. IV.16 : ***Comparaison au moyen de calculs de RMSd des conformations de la chaîne sucre-phosphate et des appariements entre les structures théoriques (BCE après raffinement d'énergie) et les conformations PDB, des tri-boucles d'ADN à des échelles locales sur des sélections d'atomes :*** *RMSd calculés sur l'appariement (App.) et les atomes principaux de la chaîne sucre-phosphate (P, O5', C5', C4', C3' et O3') de la partie en boucle et sur la tige et la boucle entre atomes homologues des structures théoriques et expérimentales. $\sigma_{dev}tor$: Moyenne des écarts des angles de torsion homologues entre la structure théorique minimisée et la structure expérimentale.* (1) *La structure expérimentale est constituée d'une tige de deux plateaux de paires de bases. Les calculs incluant la tige sont donc limités aux deux derniers plateaux de la structure PDB. Le calcul de $\sigma_{dev}tor$ ne peut donc être effectué pour les angles de torsion (α, β) en 5' et (ε, ζ) en 3' de la structure car des atomes sont absents pour calculer ces angles.*

Les valeurs des moyennes des écarts des angles de torsion comfirment à l'échelle des angles, le bon placement des atomes de la chaîne sucre-phosphate. Avec des valeurs comprises à une exception près entre 7° et 17,7°, on montre que les angles des structures théoriques après minimisation sont quasiment identiques à ceux des structures expérimentales. Nous rappelons ici, que pour obtenir ces résultats à l'issue de la minimisaiton nous avons introduit des contraintes en angle de torsion dont le nombre varie entre 1 et 3 angles tous situés au niveau du "sharp-turn". Ce résultat indique donc deux éléments. D'abord, que la modélisation de la zone du "sharp-turn" est toujours une étape délicate. D'autre part que notre approche de modélisation est très efficace pour déterminer la structure des tri-boucles considérées à un, deux ou trois angles près.

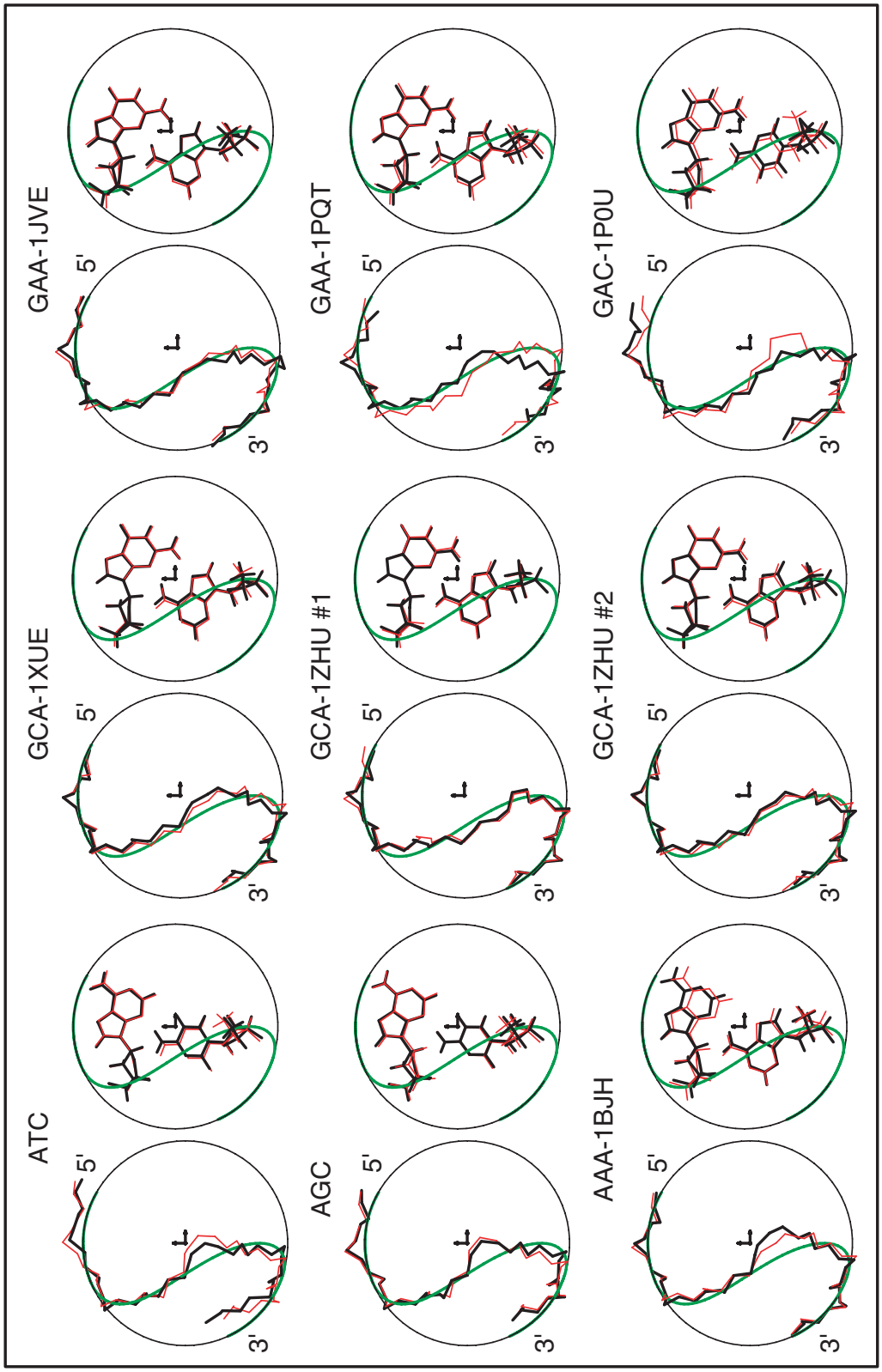


Fig. IV.6 : ***Comparaison des trajectoires des chaînes sucre-phosphates et des appariements des structures théoriques et BCE :*** *Superpositions des trajectoires de la chaîne sucre-phosphate et des bases appariées de la structure théorique minimisée en rouge et de la première structure expérimentale de chaque fichier PDB en noir.*

IV.4.3 Comparaison des structures BCE_{min} et expérimentales, globalement à l'échelle de tous les nucléotides

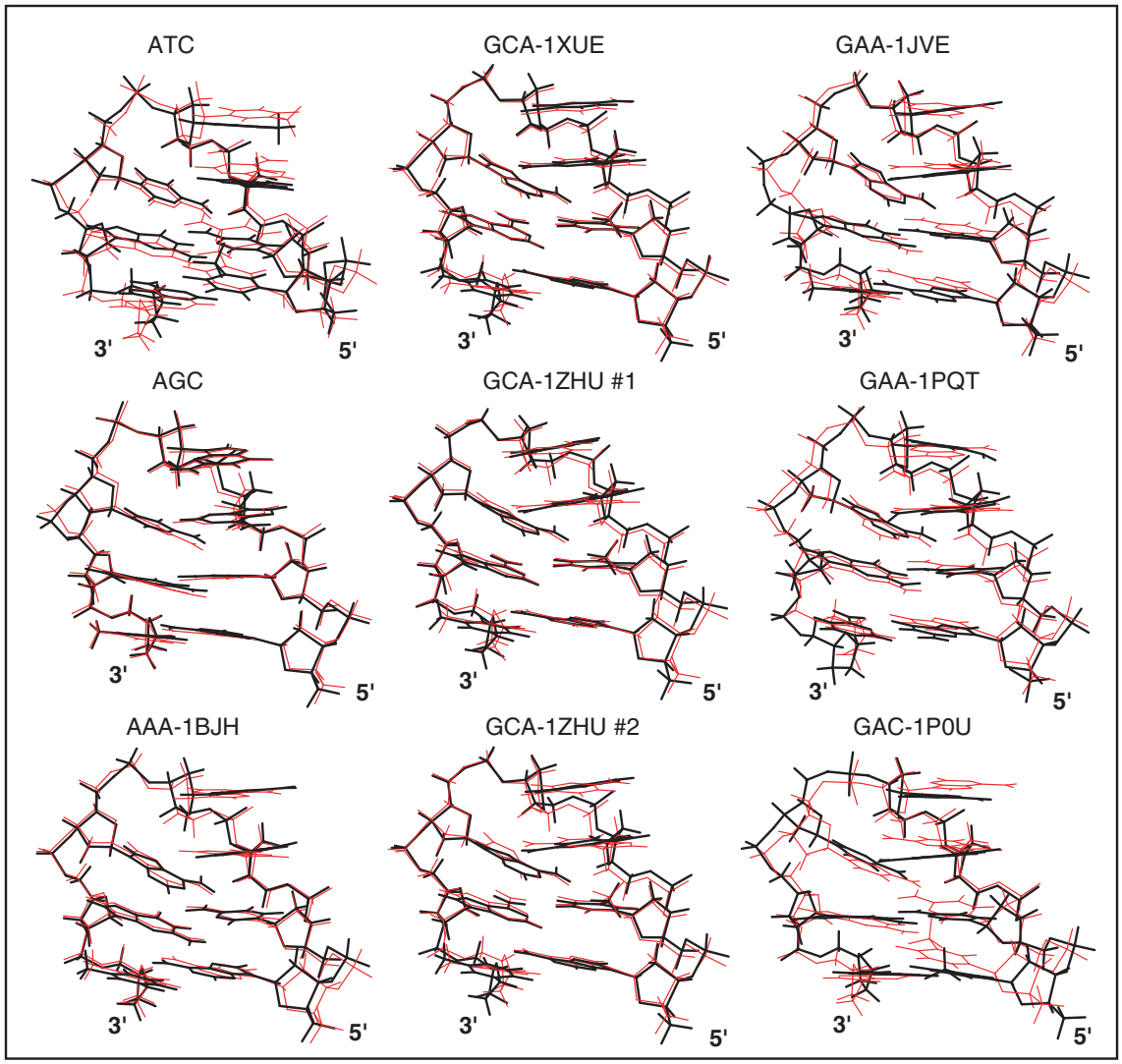


Fig. IV.7 : ***Superposition des structures globales théoriques minimisées et expérimentales des tri-boucles d'ADN comportant un appariement dans la boucle :*** *Représentation des 9 conformations de tri-boucles d'ADN. En noir, la structure théorique minimisée et en rouge la structure expérimentale issue du fichier PDB.*

La comparaison de la molécule à l'échelle globale montre un très bon accord entre les structures théoriques et les structures expérimentales sur tous les atomes(*cf.* Fig. : IV.7). Sur la partie en boucle les RMSd varient entre 0,39Å et 1,28 Å et sur la tige et la boucle entre 0,38Å et 1,33 Å (*cf.* Tab. : IV.17). Ces accords sont très bons, et restent tous dans la fourchette de résolution de la technique expérimentale, la RMN.

		RMSd	
Appariement	**Structure**	**Boucle+ Tige**	**Boucle**
A···A	AAA-1BJH	1,02 Å	1,11 Å
G···A	GCA-1XUE	0,96 Å	0,94 Å
	GCA-1ZHU #1	0,98 Å	0,96 Å
	GCA-1ZHU #2	0,96 Å	0,94 Å
	GAA-1JVE	0,97 Å	1,06 Å
	GAA-1PQT	1,16 Å	0,97 Å
A···C	ATC	1,03 Å	1,00 Å
	AGC	0,38 Å	0,39 Å
G···C	GAC-1P0U	1,33 Å	1,28 Å

Tab. IV.17 : ***Comparaison par RMSd des conformations de la chaîne sucre-phosphate et des appariements des structures théoriques et des appariements des tri-boucles d'ADN à l'échelle globale pour tous les atomes :*** *RMSd calculés sur la tige et sur la tige et la boucle entre atomes homologues de la structure théorique minimisée et la structure expérimentale. La partie en boucle est définie par tous les atomes compris entre les deux sucres du dernier plateau de paire de bases de l'hélice de la tige. La partie tige utilisée dans ce calcul comprend les deux derniers plateaux de paires de bases de la tige.*

IV.5 Discussion

IV.5.1 Un système non pseudo-dyadique

L'exploration des conformations empilées des bases de la boucle montre que, pour satisfaire la formation de liaisons hydrogène, la première base doit toujours être en conformation Anti. En effet, les explorations Syn/Anti et Syn/Syn ne donnent jamais lieu à des pics de fort volume. L'étude des appariements dans une boucle d'épingle à cheveux et dans les hélices d'ADN sont de ce point de vue des problèmes différents.

Les hélices d'ADN possèdent une symétrie pseudo-dyadique et donc toute conformation d'appariement valide pour une paire de bases X···Y donnée, sera possible pour Y···X en retournant le plateau autour de l'axe (Ox). Ce retournement conserve la conformation Syn ou Anti de la base X et Y. Ainsi, un appariement Xanti···Yanti, donne par symétrie pseudo-dyadique un appariement Yanti···Xanti, un appariement Xanti···Ysyn donne un appariement Ysyn···Xanti, et un appariement Xsyn···Ysyn donne un appariement Ysyn···Xsyn. Une autre conséquence de la pseudo-dyadicité porte

sur la conformation des appariements homo-base (A$\cdots$A, G$\cdots$G, C$\cdots$C, T$\cdots$T). Ces appariements peuvent adopter des conformations "sheared" (cisaillées) où l'une des bases est dans le grand sillon et l'autre dans le petit. Par symétrie dyadique les conformations où les positions sont échangées sont automatiquement possibles et de même énergie.

Dans le cas des appariements dans les boucles des épingles à cheveux, cette symétrie n'existe pas. En effet, la trajectoire de la chaîne sucre-phosphate associée aux bases formant les appariements dans la boucle n'est pas pseudo-dyadique. Une conséquence est que la présence d'appariements de type ANTI/SYN n'entraîne pas, dans ce système, la validité automatique de l'appariement ANTI/SYN. C'est pourquoi il n'est pas surprenant que la moitié de ces explorations ne soient pas favorables, ce qui correspond ici à la base 5' de la boucle en conformation SYN. Pour l'appariement homo-A "sheared" on observe que la seule conformation possible correspond au cas où A en 5' est dans le grand sillon et A en 3' dans le petit sillon.

IV.5.2 Constantes de forces et longueur de persistance du simple brin pour la torsion et la flexion

IV.5.2.1 Les hypothèses de travail et définition de la fonction de score

Lors de l'élaboration de notre approche, nous avons postulé dans un premier temps, à partir des résultats obtenus dans le chapitre III, que les conformations des bases appariées dans la boucle peuvent être obtenues directement :

- par une exploration autour de la géométrie de la chaîne sucre-phosphate, fixée et calculée par la théorie de l'élasticité, et
- au moyen d'une fonction de score qui évalue l'écart à la géométrie idéale d'une liaison hydrogène donnée.

La forme donnée à cette fonction de score permet de l'interpréter, de façon approximative comme la probabilité de former cette liaison hydrogène. En effet, la fonction de score est toujours positive, elle prend la forme d'une Gaussienne de maximum égal à 1, et son intégration sur l'ensemble de l'intervalle de définition est

proche de 1. Cette fonction est donc presque normée à 1. Cette fonction de score détecte un nombre important de liaisons hydrogène et se révèle peu discriminante.

Dans un second temps, nous avons donc rajouté un terme supplémentaire à la fonction de score pour tenir compte des conformations de moindre déformation de la chaîne sucre-phosphate en torsion (Ω) et flexion (Θ) pour chacune des bases. Comme précédemment, nous avons donné à ces nouveaux termes une forme de Gaussienne presque normée dont le maximum est égal à 1.

Au total la fonction de score est le produit de deux fonctions évaluant respectivement le score d'une liaison hydrogène donnée et celui des déformations de la chaîne sucre-phosphate par rotation autour du fil ou par flexion du fil élastique de chacune des bases (*cf.* EQ. : IV.5.2.3 & PART. : II.7.3.5 pour la définition des notations).

$$\begin{aligned} Score = & \left(e^{-\tau \left(r_{H-Acc.} - r_0 \right)^2} \cdot e^{-\tau \left(\theta_{don.} - \theta_{0,don.} \right)^2} \cdot e^{-\tau \left(\theta_{acc.} - \theta_{0,acc.} \right)^2} \right) \\ \times & \left(e^{-\tau \left(\Omega_1^2 + \Omega_3^2 + \Theta_1^2 + \Theta_3^2 \right)} \right) \end{aligned} \qquad \text{(IV.5.2.3)}$$

IV.5.2.2 Identification de la fonction de score avec les lois de Boltzmann et de Hooke

Les deux termes de la fonction de score peuvent être redéfinis à partir du formalisme général de la loi de Boltzmann et de loi de Hooke :

D'après la loi de Boltzmann, f(x), la densité de probabilité de l'état x est proportionnelle à :

$$f(x) = \alpha \, e^{-\frac{E(x)}{RT}} \qquad \text{(IV.5.2.4)}$$

où

E(x) est l'énergie du système,

T, est la température absolue et

R, est la constante des gaz parfaits.

La loi de Hooke permet d'exprimer l'énergie d'un système en fonction d'une variable, x, lorsque celui-ci s'écarte d'une norme, x_0 :

$$E(x) = \frac{1}{2}k_x(x - x_0)^2 = \frac{1}{2}k_x\Delta x^2 \qquad \text{(IV.5.2.5)}$$

Il est possible de réécrire la fonction de score en exprimant les termes exponentiels à partir de la loi de Boltzmann (*cf.* Eq. : IV.5.2.4) et de la loi de Hooke (*cf.* Eq. : IV.5.2.5).

$$\begin{aligned} Score = & \left(e^{-\frac{1}{2}\frac{k_r}{RT}\left(r_{H-Acc} - r_0\right)^2} \cdot e^{-\frac{1}{2}\frac{k_{\theta_{don}}}{RT}\left(\theta_{don} - \theta_{0,don}\right)^2} \cdot \right. \\ & \left. e^{-\frac{1}{2}\frac{k_{\theta_{acc}}}{RT}\left(\theta_{acc} - \theta_{0,acc}\right)^2} \right) \\ \times & \left(e^{-\frac{1}{2}\frac{k_\Omega}{RT}\left(\Omega_1^2 + \Omega_3^2\right)} e^{-\frac{1}{2}\frac{k_\Theta}{RT}\left(\Theta_1^2 + \Theta_3^2\right)} \right) \end{aligned} \qquad \text{(IV.5.2.6)}$$

où :

r, en Angström est la distance séparant le proton du groupement accepteur, assimilable à la longueur de la liaison hydrogène ,

θ_{acc} et θ_{don} sont les angles en radians caractérisant respectivement l'orientation de la liaison hydrogène par rapport à la direction de l'orbitale acceptrice et par rapport à la direction de la liaison Donneur-Proton,

Ω et Θ sont les angles en radians caractérisant respectivement la torsion et la flexion pour un nucléotide,

k_r, est une constante phénoménologique de raideur d'élongation et de compression de la liaison hydrogène, en $kcal\,mol^{-1}\,\text{Å}^{-2}$,

$k_{\theta_{don}}$ et $k_{\theta_{acc}}$ sont des constantes de raideur de l'alignement de la liaison hydrogène, en $kcal\,mol^{-1}\,rad^{-2}$,

k_Ω et k_Θ sont respectivement des constantes de torsion et de flexion de la chaîne sucre-phosphate, $kcal\,mol^{-1}\,rad^{-2}$,

R, est la constante des gaz parfaits et,

T, la température du système en Kelvin.

Cette fonction peut être écrite de façon compacte de la façon suivante :

$$Score = \prod_x e^{-\frac{1}{2}\frac{k_x \Delta x^2}{RT}}$$

où

x prend chacune des valeurs r, θ_{acc}, θ_{don}, Ω et Θ .

Chacun des termes de potentiel de cette fonction évalue d'une part les contributions détaillées à la stabilité énergétique d'une liaison hydrogène dans un appariement donné, et d'autre part les contributions à l'énergie de déformation de la chaîne sucre-phosphate en rotation autour du fil et en flexion du fil élastique de chacunes des bases B_1 et B_3 de l'appariement. Dans ce modèle le moteur de la déformation est lié à la formation de liaisons hydrogène. De façon très approximative, comme le système est stable, l'énergie de déformation en torsion et en flexion est de l'ordre de, ou inférieure à, l'énergie de stabilisation apportée par la formation de liaisons hydrogène.

IV.5.2.3 Interprétation en terme de probabilités

La fonction de densité de probabilité d'une Gaussienne fonction de la variable, x, est donnée par :

$$f(x) = \frac{1}{\sqrt{2\,\pi}\sigma_x} e^{-\frac{1}{2}\left(\frac{\Delta x}{\sigma_x}\right)^2} \qquad \text{(IV.5.2.7)}$$

En identifiant chacun des exposants de l'équation IV.5.2.3 et de l'équation IV.5.2.7 il vient :

$$\tau = \frac{1}{2\sigma_x^2} \Leftrightarrow \sigma_x = \frac{1}{\sqrt{2\,\tau}}$$

Nous avons fixé :

$$\tau = \frac{\pi}{2}$$

donc :

$$\sigma_x = \frac{1}{\sqrt{\pi}}$$

Pour x=r Å, numériquement, on a $\sigma_r = 0,564\,\text{Å}$.

Pour x=Ω° (torsion), ou x=Θ° (flexion), numériquement, on a $\sigma_\Omega = \sigma_\Theta = 32,33^\circ$.

Les fonctions $e^{-\tau\,\Delta x^2}$ de la fonction de score IV.5.2.3 ne sont pas des Gaussiennes pures mais tronquées (*cf.* PART. : II.7.3.5). Elles peuvent être approximées par des Gaussiennes pures avec un facteur de pré-exponentielle de normalisation $\frac{1}{\sqrt{2\,\pi}\sigma_x}$ proche de 1, car, pour $\sigma_x = \frac{1}{\sqrt{\pi}}$, on a :

$$\frac{1}{\sqrt{2\,\pi}\sigma_x} = \frac{1}{\sqrt{2}} \simeq 0,707$$

L'ensemble de ce raisonnement est donc assez cohérent.

IV.5.2.4 Équivalence distribution Gaussienne - lois de Boltzmann et de Hooke

En identifiant chacun des exposants de l'équation **??** et de l'équation de probabilité IV.5.2.7 il vient :

$$\frac{1}{2\,\sigma_x^2} = \frac{1}{2}\frac{k_x}{R\,T} \Leftrightarrow k_x = \frac{R\,T}{\sigma_x^2}$$

Pour $\sigma_x = \frac{1}{\sqrt{\pi}}$, il vient :

$$k_x = \pi\, R\, T \tag{IV.5.2.8}$$

Pour les différentes valeurs de x on obtient :

$$\begin{aligned} k_r = \quad \pi\, R\, T = \quad & 1,872\, kcal\, mol^{-1}\, \text{Å}^{-2} \\ k_{\theta_{don}} = k_{\theta_{acc}} = \quad \pi\, R\, T = \quad & 1,872\, kcal\, mol^{-1}\, rad^{-2} \\ k_\Omega = k_\Theta = \quad \pi\, R\, T = \quad & 1,872\, kcal\, mol^{-1}\, rad^{-2} \end{aligned}$$

Le bon accord entre les structures tridimensionnelles calculées par notre approche théorique et les structures expérimentales donne par conséquent une estimation de la grandeur des constantes de torsion, k_Ω, et de flexion, k_Θ, des nucléotides, respectivement autour et le long de la chaîne sucre-phosphate.

IV.5.2.5 Équivalence énergie thermodynamique de torsion et de flexion - loi de Hooke

Les relations [70,86] entre les énergies de déformation en torsion ou en flexion et les longueurs de persistance en flexion, $L_{p_{torsion}}$ et en torsion, $L_{p_{flexion}}$sont données par :

$$\begin{aligned} \Delta G_{torsion} =& R\, T\, L_{p_{torsion}} \times \frac{(\Delta\Omega)^2}{L} \\ \Delta G_{flexion} =& R\, T\, L_{p_{flexion}} \times \frac{(\Delta\Theta)^2}{L} \end{aligned} \tag{IV.5.2.9}$$

Dans ces expressions $\Delta\Omega$ et $\Delta\Theta$ correspondent respectivement aux variations angulaires en torsion et en flexion de la chaîne sucre-phosphate, pour une longueur L. Dans notre traitement, L est la longueur de chaîne d'un nucléotide (L_{1nt}≃5-8Å).

En identifiant l'exposant de l'équation de la fonction de score exprimée au moyen des lois de Boltzmann et de Hooke (*cf.* Eq. : **??**), il est possible d'évaluer la longueur de persistance en fonction des constantes de force de rappel en torsion et en flexion :

$$\frac{1}{2}k_x\Delta x^2 \;=\; R\,T\,L_p\,\frac{\Delta x^2}{2\,L}$$

On obtient ainsi par simplication :

$$k_x \;=\; R\,T\,\frac{L_p}{L_{1\,nt}}$$

En substituant dans cette relation la valeur de k_x par son expression établie dans la relation IV.5.2.8, on obtient :

$$\pi\,R\,T \;=\; R\,T\,\frac{L_p}{L_{1\,nt}} \;\Rightarrow\; L_p \;=\; \pi\,L_{1\,nt}$$

À partir des relations IV.5.2.8 et il vient :

$$L_{P_{torsion}} = L_{p_{flexion}} \simeq 3 \times L_{1\,nt} \qquad \text{(IV.5.2.10)}$$

ce qui donne un ordre de grandeur de la longueur de persistance en torsion et flexion de la chaîne sucre-phosphate. En attribuant une longueur de chaîne pour un nucléotide comprise entre 5 Å et 8 Å, les longueurs de persistance sont de l'ordre de 15 Å à 24 Å.

IV.5.2.6 Équivalence probabilité - loi de Boltzmann et énergie thermodynamique

En identifiant les exposants de l'équation de Boltzmann, où l'énergie est exprimée au moyen de l'équation IV.5.2.9, et de l'équation IV.5.2.7, il vient :

$$\begin{aligned} \frac{1}{2}\left(\frac{\Delta x^2}{\sigma_x}\right) &= \frac{R\,T\,L_p\,\Delta x^2}{2\,L} \\ \Leftrightarrow \qquad \frac{1}{\sigma_x^2} &= \frac{L_p}{L} \end{aligned}$$

$$\begin{aligned} \frac{1}{2}\left(\frac{\Delta x^2}{\sigma_x}\right) &= \frac{R\,T\,L_p\,\Delta x^2}{2\,L} \\ \Leftrightarrow \qquad \frac{1}{\sigma_x^2} &= \frac{L_p}{L} \end{aligned}$$

soit :

$$\begin{aligned} L_p &= \frac{L}{\sigma_x^2} \\ \Leftrightarrow \quad \sigma_x^2 &= \frac{L}{L_p} \end{aligned}$$

IV.5.2.7 Les mesures de la longueur de persistance du simple brin

Pour utiliser la théorie de l'élasticité nous avons postulé que la chaîne sucre-phophate se comporte comme une barre mince rigide et inextensible. En première approximation cette théorie est applicable lorsque la longueur de la barre est inférieure à $L_p/2$ [97]. Nos résultats donnent une longueur de persistance de l'ordre de trois nucléotides (*cf.* Eq. : IV.5.2.10). Cet ordre de grandeur est en accord avec nos hypothèses de départ puisque le rapport $\frac{L_{barre}}{Lp}$ en flexion et en torsion est de l'ordre de 1 et que les deux extrémités de la barre sont fixées par les extrémités de la double hélice.

La mesure de la longueur de persistance de l'ADN simple-brin est un sujet qui donne lieu à de nombreuses études. Différentes techniques expérimentales sont utilisées pour mesurer ce paramètre fondamental qui conditionne les conformations accessibles. Ainsi, une étude de la stabilité des épingles à cheveux en fonction de la longueur de la boucle utilisant un modèle de "fermeture éclair" [70] mécanique et statistique appliqué aux polymères semiflexibles a montré que la longueur de persistence L_p des ADN simple-brins poly-d(T) est proche de 14 Å [98]. Des expériences fondées sur des mesures de dimensions de pelotes statistiques de polymères, par diffusion de la lumière ou par sédimentation dans des conditions de solvant théta, donnent une valeur de L_p=14 Å pour des chaînes de poly-r(U) simple-brin [99, 100]. Des mesures de réponses élastiques d'ADN simple-brin utilisant des pinces optiques donnent une évaluation de L_p à 7,5 Å [101]. Des mesures de distances

entre extrémités de segments de poly-d(T) insérés entre deux duplex d'ADN évaluent L_p dans un intervalle compris entre 13 Å et 28 Å [102]. L'interprétation de coefficients d'auto-diffusion de chaînes simple-brin par des expériences de restitution de la fluorescence après photo-blanchiment donne une longueur de persistance L_p=13 Å [103]. La relaxation de la biréfringence électrique de duplex comprenant des parties simple-brins donne L_p=25-35 Å pour des poly-d(T) et L_p=75 Å pour des poly-d(A) [104].

Au total, la mesure de la longueur de persistance des simple-brins d'ADN est loin d'être établie. Les différences observées reflètent une dépendance en fonction de la force ionique et de la séquence. La longueur de persistance dérivée de notre modéle s'inscrit dans les ordres de grandeurs des valeurs données ci-dessus. Il est intéressant de constater que cette grandeur peut-être estimée uniquement à partir de la recherche des meilleurs appariements dans les tri-boucles et de quelques hypothèses raisonnables. Une prise en compte plus systématique et plus rigoureuse devrait permettre d'atteindre une meilleure évaluation de ce paramètre fondamental, et de mieux comprendre les forces et énergies impliquées dans le maintien des structures en épingle à cheveux.

IV.6 Conclusion

Dans ce chapitre nous avons exploré la formation d'appariements dans les tri-boucles d'ADN à partir d'un concept très simple qui dérive de l'approche BCE. La trajectoire calculée par la théorie de l'élasticité procure une armature autour de laquelle les conformations des bases sont modifiées au moyen de trois d.d.l.. Ce sont dans l'ordre : la rotation autour de la tangente au fil élastique d'angle Ω, la rotation de redressement d'empilement d'angle Θ_{empil}, et la rotation de la base autour de la liaison glycosidique d'angle χ. Le petit nombre de ces d.d.l. permet d'explorer systématiquement l'ensemble de l'espace conformationnel où les bases extrémales de la boucle sont empilées sur le dernier plateau de l'hélice de la tige.

La fonction de score de liaison hydrogène est très simple. Elle discrimine efficacement les appariements les plus stables. Les conformations tridimensionnelles complètes sont construites à partir des paramètres optimaux Ω, Θ et χ obtenus par l'exploration exhaustive des appariements possibles entre les bases extrémales de la boucle et par le placement en conformation empilée de la base centrale de la tri-boucle. Les structures finales sont obtenues par minimisation, avec une incertitude

encore non résolue sur un, deux ou trois angles de torsion situés dans la zone du "sharp-turn". Elles sont très proches des structures résolues à partir de données RMN. Notre approche offre donc un cadre très efficace et remarquable pour calculer la conformation des boucles comportant des appariements.

Ces résultats montrent que les appariements identifiés à ce jour dans les tri-boucles d'ADN ne doivent plus être considérés comme des mésappariements. À l'image des appariements Watson-Crick dans les hélices, les appariements Pu$\cdots$Pu et Pu$\cdots$Py des boucles étudiées ici, sont les meilleurs appariements possibles étant donné les contraintes géométriques imposées par la trajectoire de la chaîne sucre-phosphate de la boucle. Celle-ci est elle-même une trajectoire de moindre énergie, calculée à partir de la théorie de l'élasticité des barres minces.

FIG. IV.8-IV.17 : Cartes de score de liaisons hydrogène en fonction des paramètres Ω_1 et Ω_3 variant entre -120° et +120° et meilleurs appariements de chaque exploration.

Chaque double page présente les résultats d'un appariement (A···A, G···G, G···A, A···C ou G···C) pour chacune des conformations ANTI/ANTI, ANTI/SYN, SYN/ANTI et SYN/SYN, dans l'ordre de lecture. Les cartes présentées donnent la forme des pics avant filtrage. La position du maximum de chaque pic avant filtrage est pointée par une barre suivie de la mention des atomes formant la liaison hydrogène, du volume du pic avant et après filtrage. Les liaisons hydrogène de type C-H···O ou C-H···N de moindre intensité sont marquées en noir-italique, et les autres liaisons hydrogène sont marquées en bleu-gras. Parmis ces dernières, les 3 liaisons hydrogène de plus fort volume après filtrage sont pointées par une flèche bleue. En rouge sont indiquées les courbes de niveau de la fonction de filtre et les valeurs associées à chaque niveau.

Sous chaque carte, les conformations associées aux maxima des pics des 3 liaisons hydrogènes de plus fort volume après filtrage sont représentées dans l'ordre décroisaant de volume: en noir les atomes des nucléotides formant l'appariement tournés des valeurs ($\Omega_{1,max}$,$\Omega_{3,max}$) après filtrage ; en rouge la liaison hydrogène du pic considéré ; en jaune, le fil associé à la trajectoire de la boucle calculée par la théorie de l'élasticité ; en gris le dernier plateau de paire de bases de la tige jouxtant l'appariement de la boucle ; en rose, les liaisons hydrogène du dernier plateau de paire de bases de la tige.

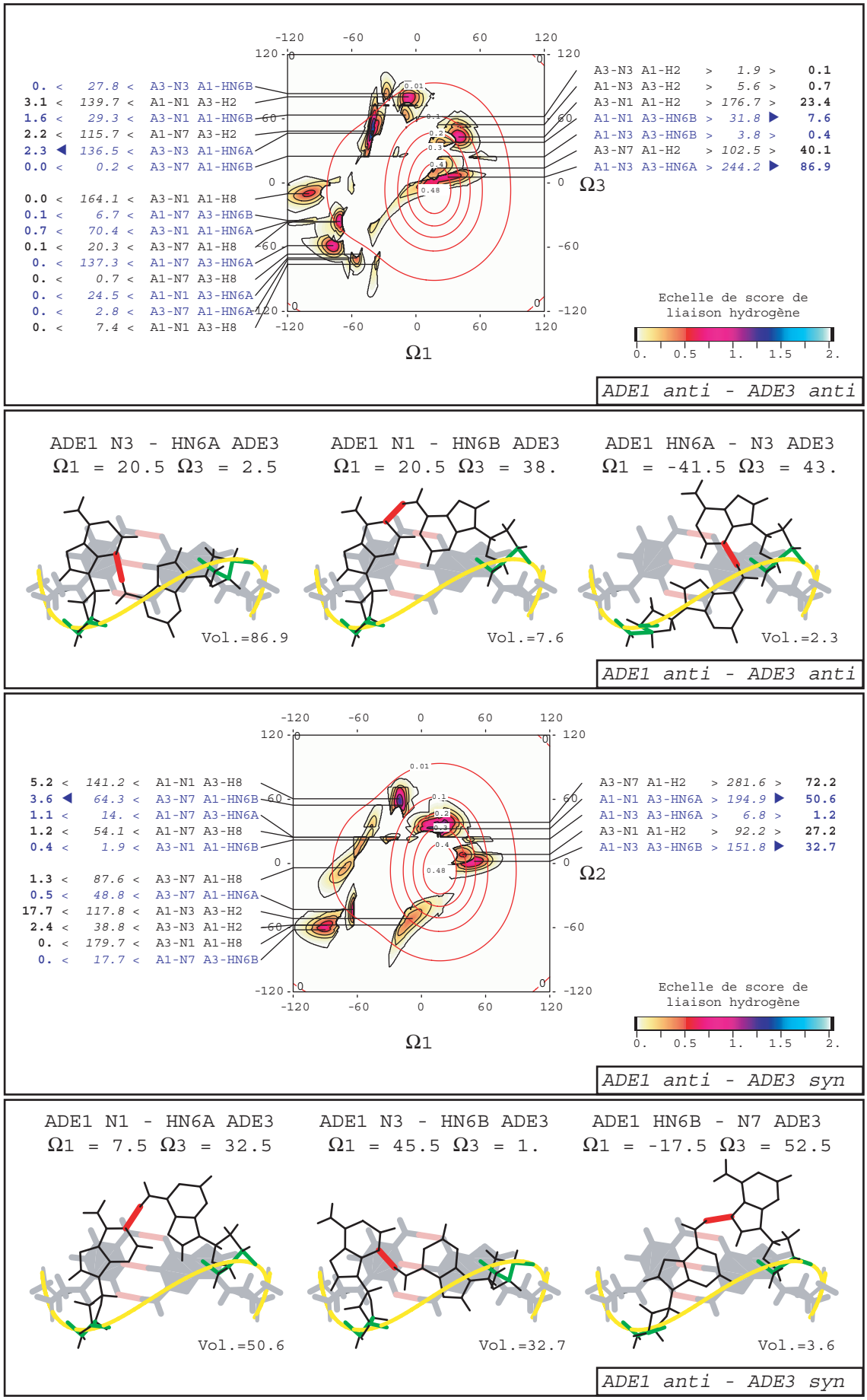


FIG. IV.8 : *Appariements ADE···ADE en conformation* ANTI-ANTI *et* ANTI-SYN - *Cartes de score de liaison hydrogène et meilleurs appariements.*

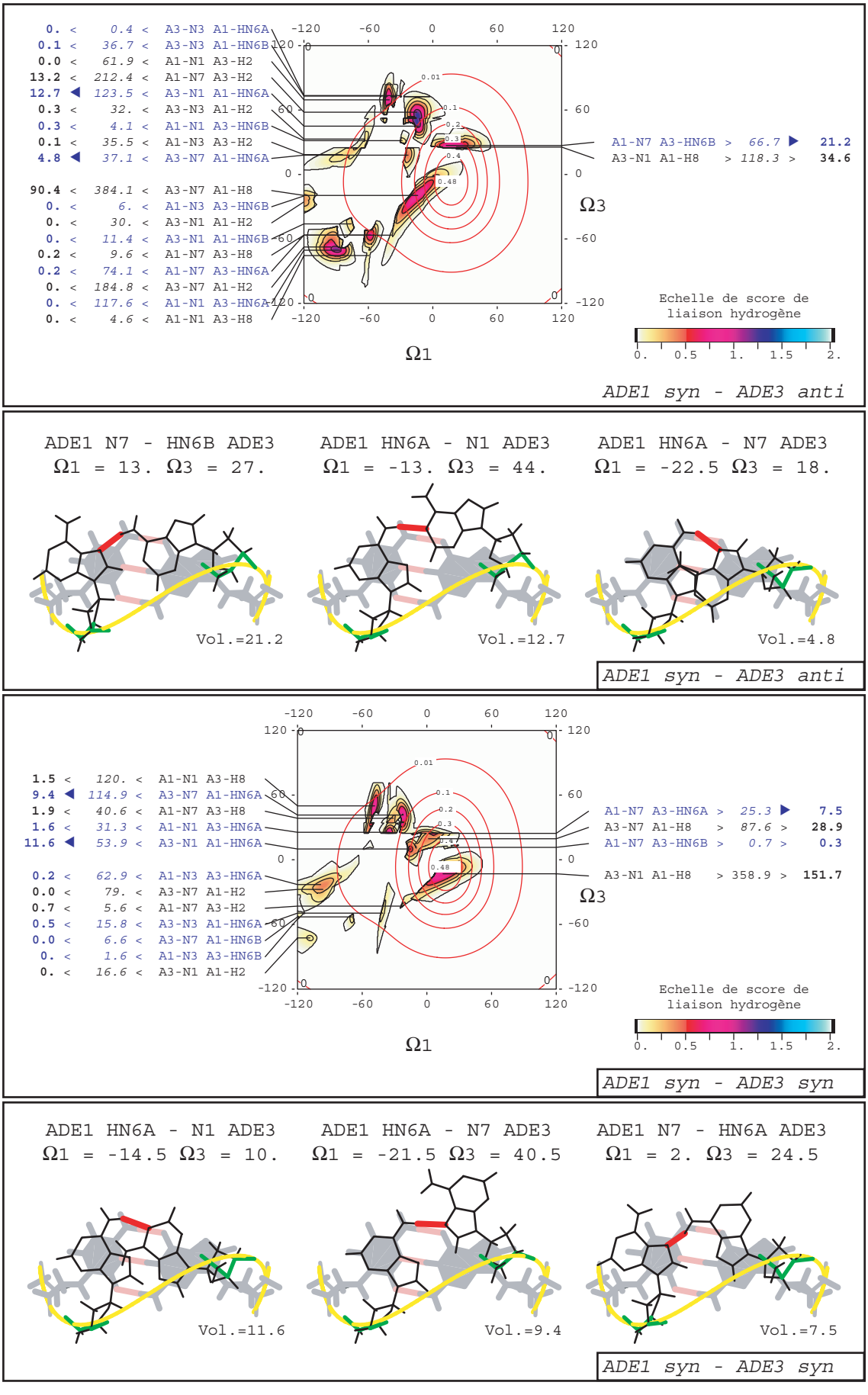


FIG. IV.9 : *Appariements ADE···ADE en conformation* SYN-ANTI *et* SYN-SYN *- Cartes de score de liaison hydrogène et meilleurs appariements.*

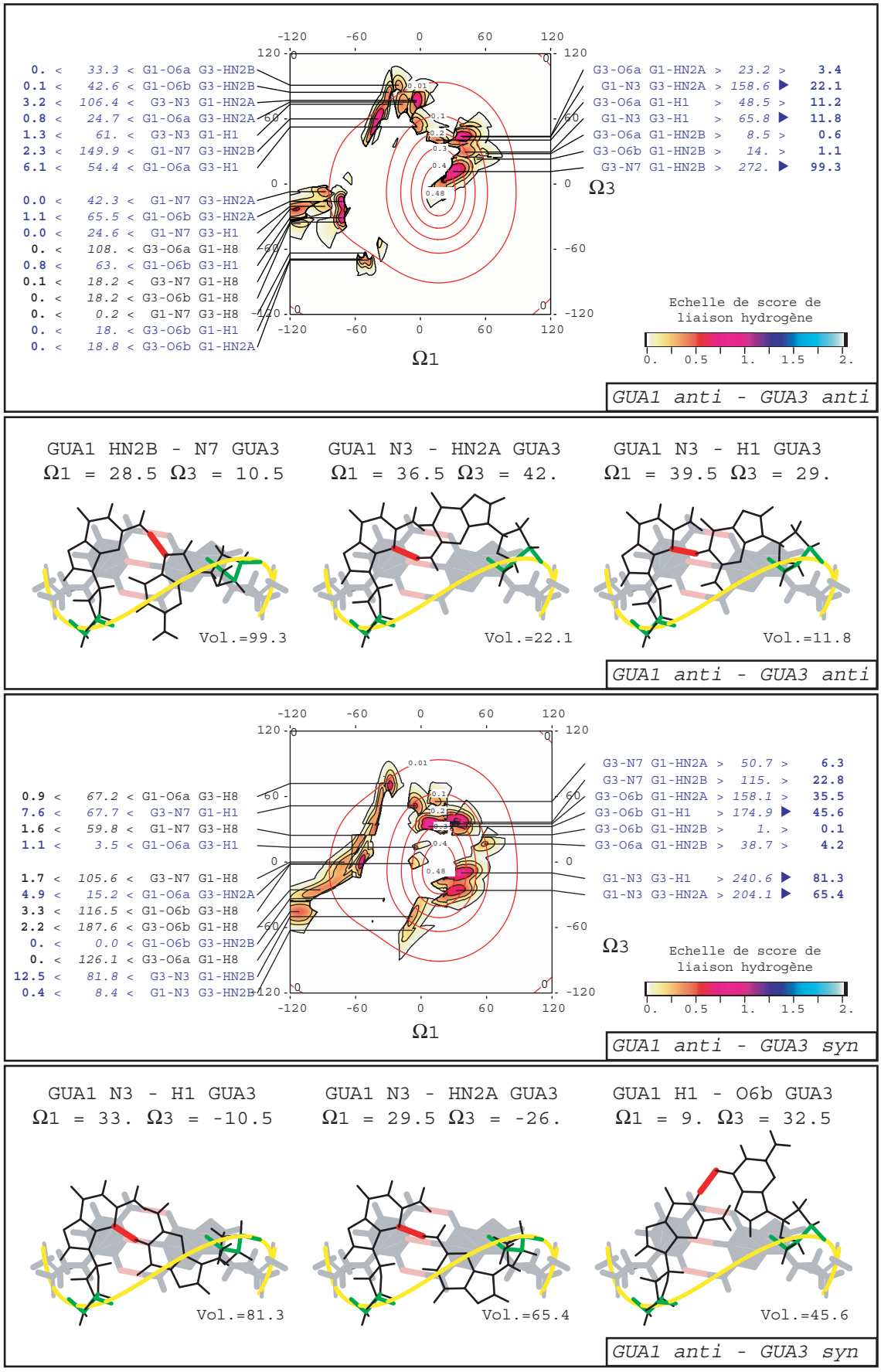


FIG. IV.10 : *Appariements GUA··· GUA en conformation* ANTI-ANTI *et* ANTI-SYN *- Cartes de score de liaison hydrogène et meilleurs appariements.*

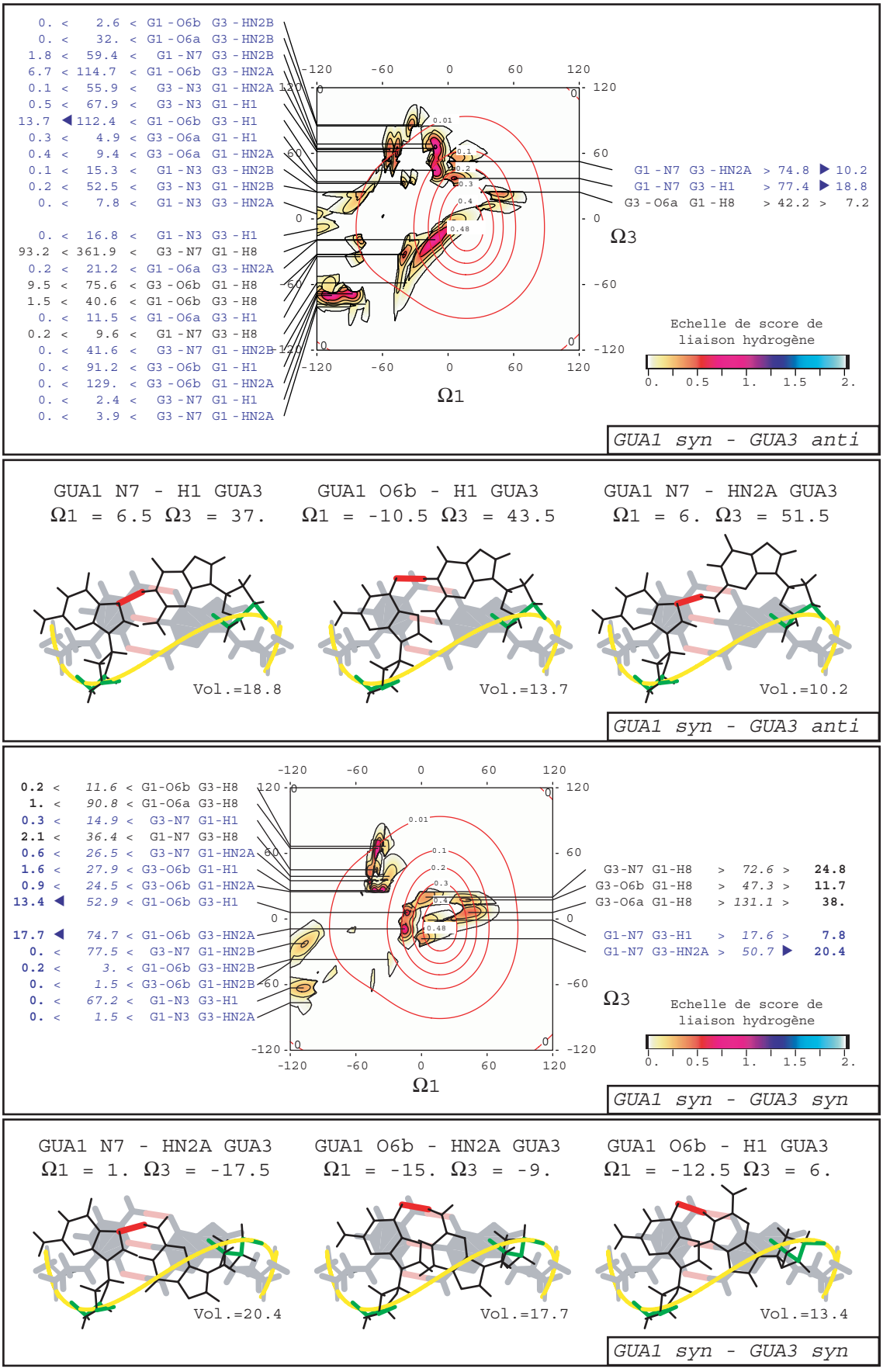


FIG. IV.11 : *Appariements GUA···GUA en conformation* SYN-ANTI *et* SYN-SYN *- Cartes de score de liaison hydrogène et meilleurs appariements.*

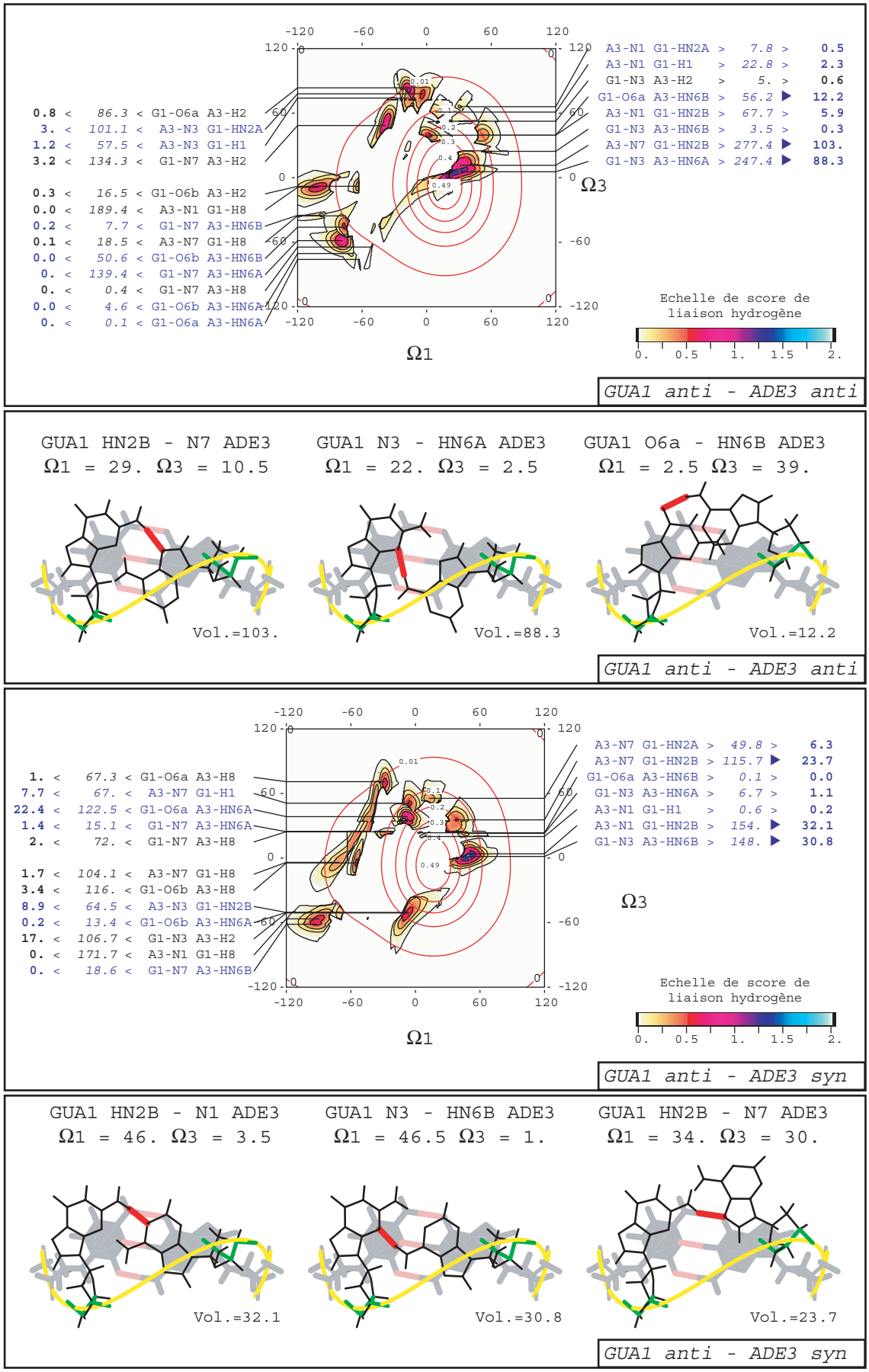


FIG. IV.12 : *Appariements GUA···ADE en conformation* ANTI-ANTI *et* ANTI-SYN *- Cartes de score de liaison hydrogène et meilleurs appariements.*

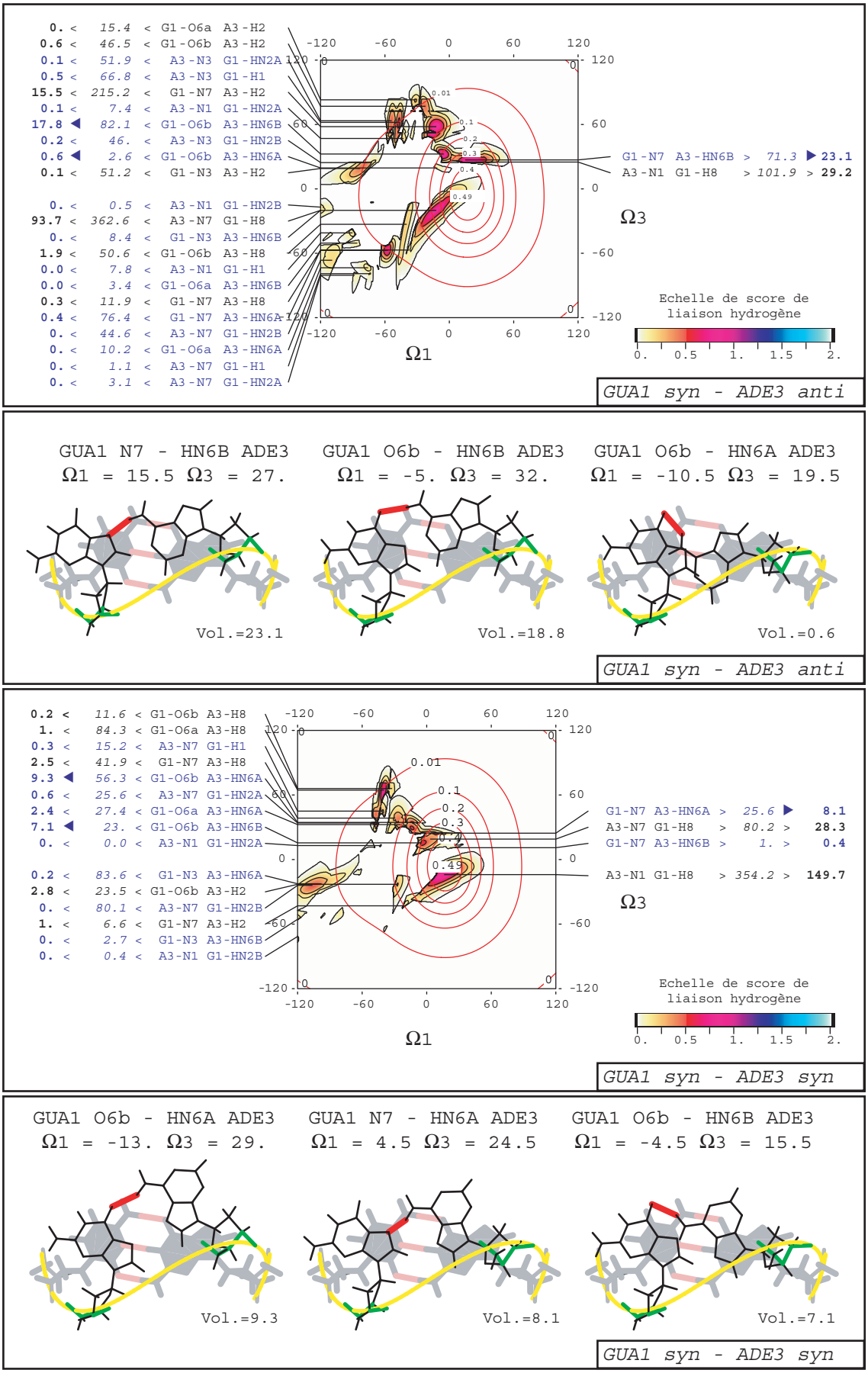


FIG. IV.13 : *Appariements GUA···ADE en conformation* SYN-ANTI *et* SYN-SYN *- Cartes de score de liaison hydrogène et meilleurs appariements.*

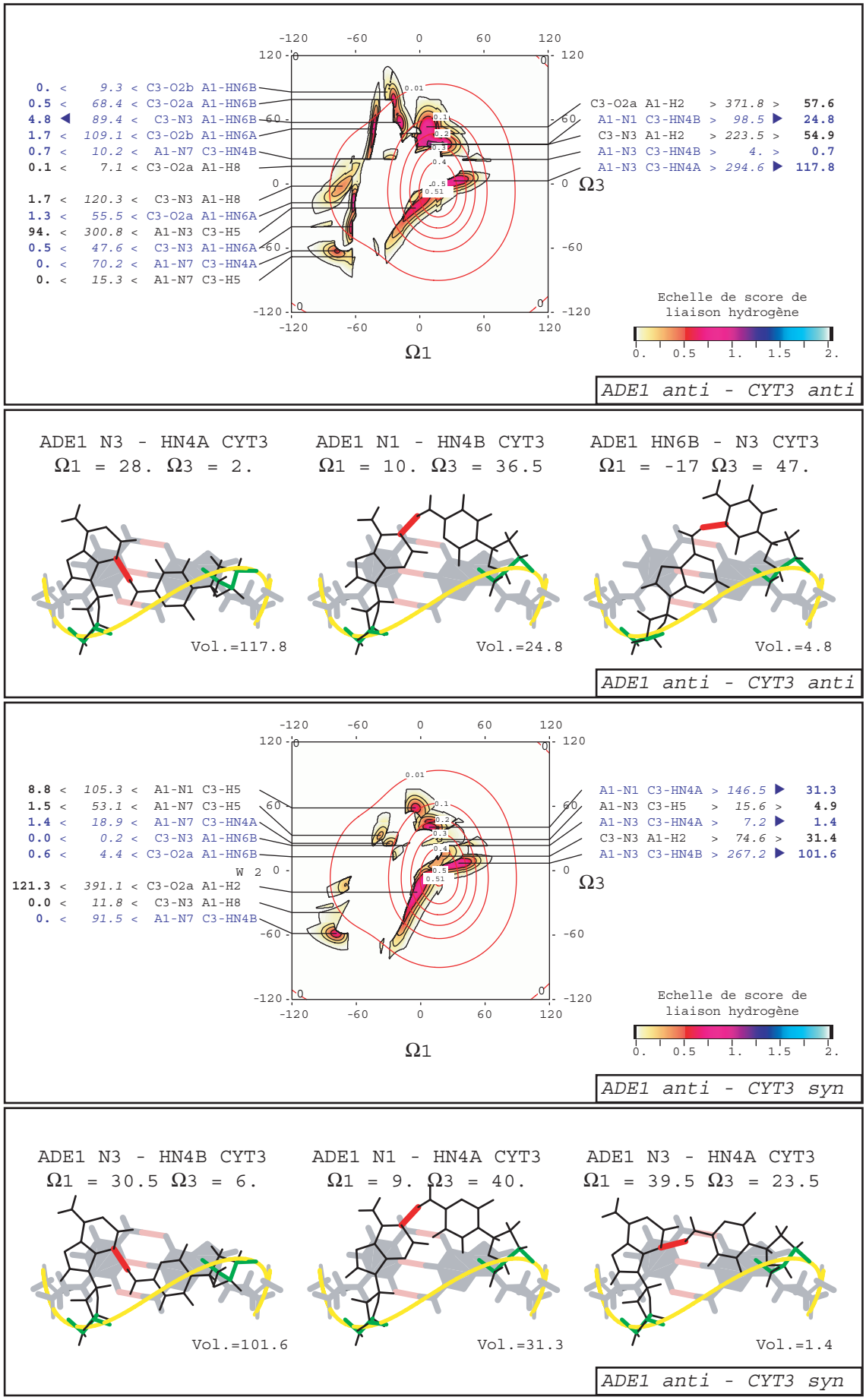


FIG. IV.14 : *Appariements ADE···CYT en conformation* ANTI-ANTI *et* ANTI-SYN - *Cartes de score de liaison hydrogène et meilleurs appariements.*

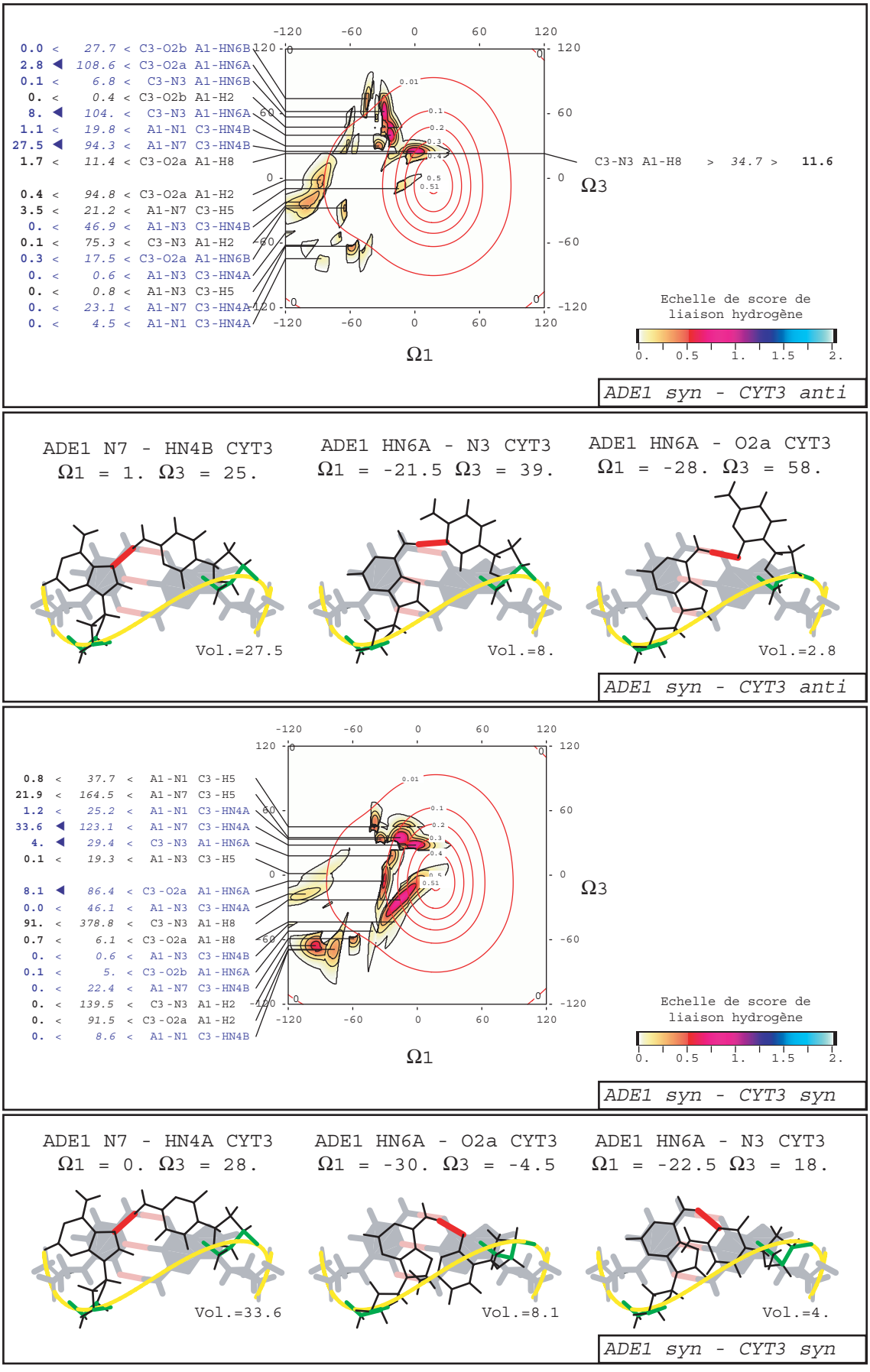


FIG. IV.15 : *Appariements ADE···CYT en conformation* SYN-ANTI *et* SYN-SYN *- Cartes de score de liaison hydrogène et meilleurs appariements.*

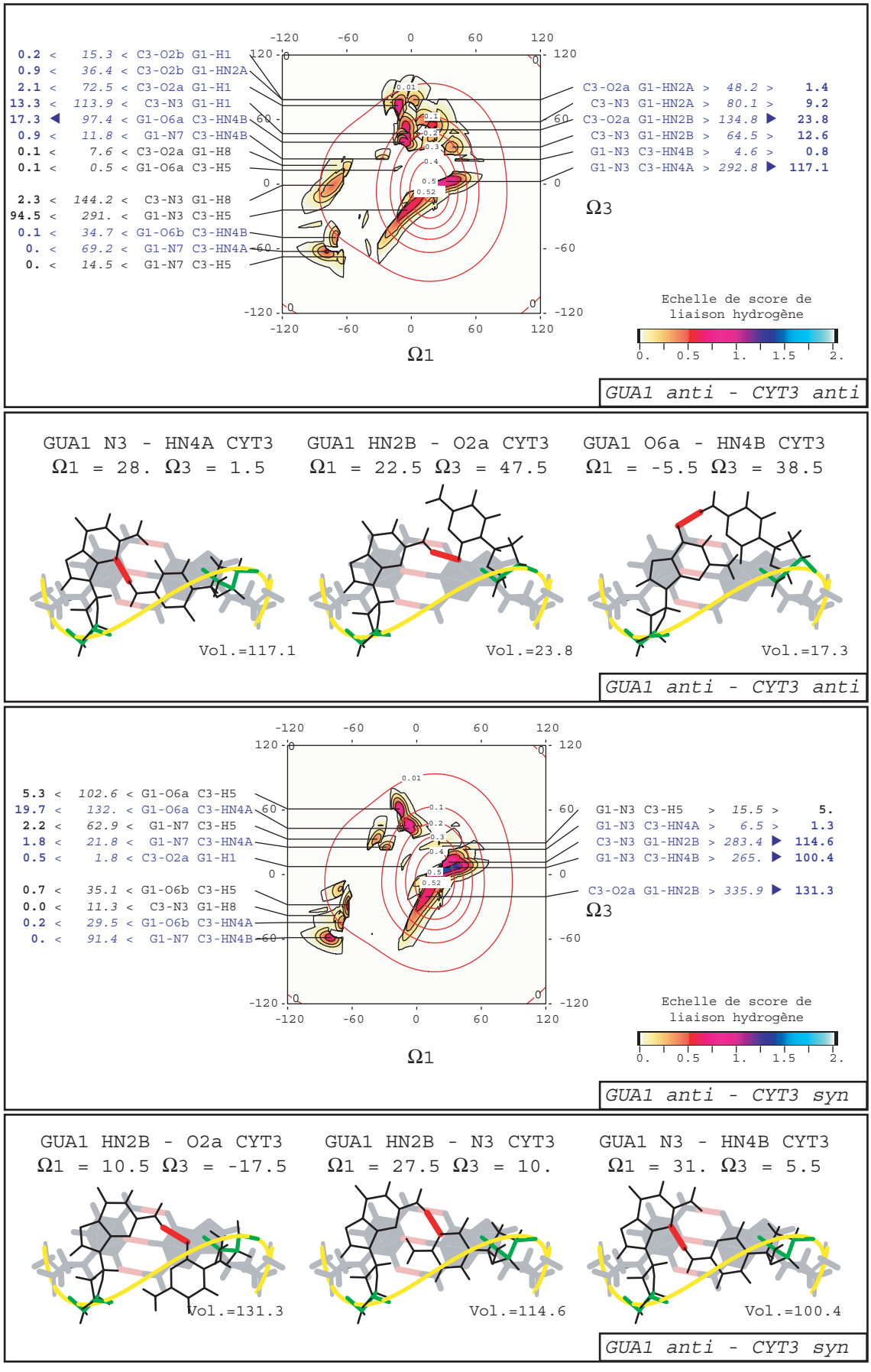


FIG. IV.16 : *Appariements GUA···CYT en conformation* ANTI-ANTI *et* ANTI-SYN *- Cartes de score de liaison hydrogène et meilleurs appariements.*

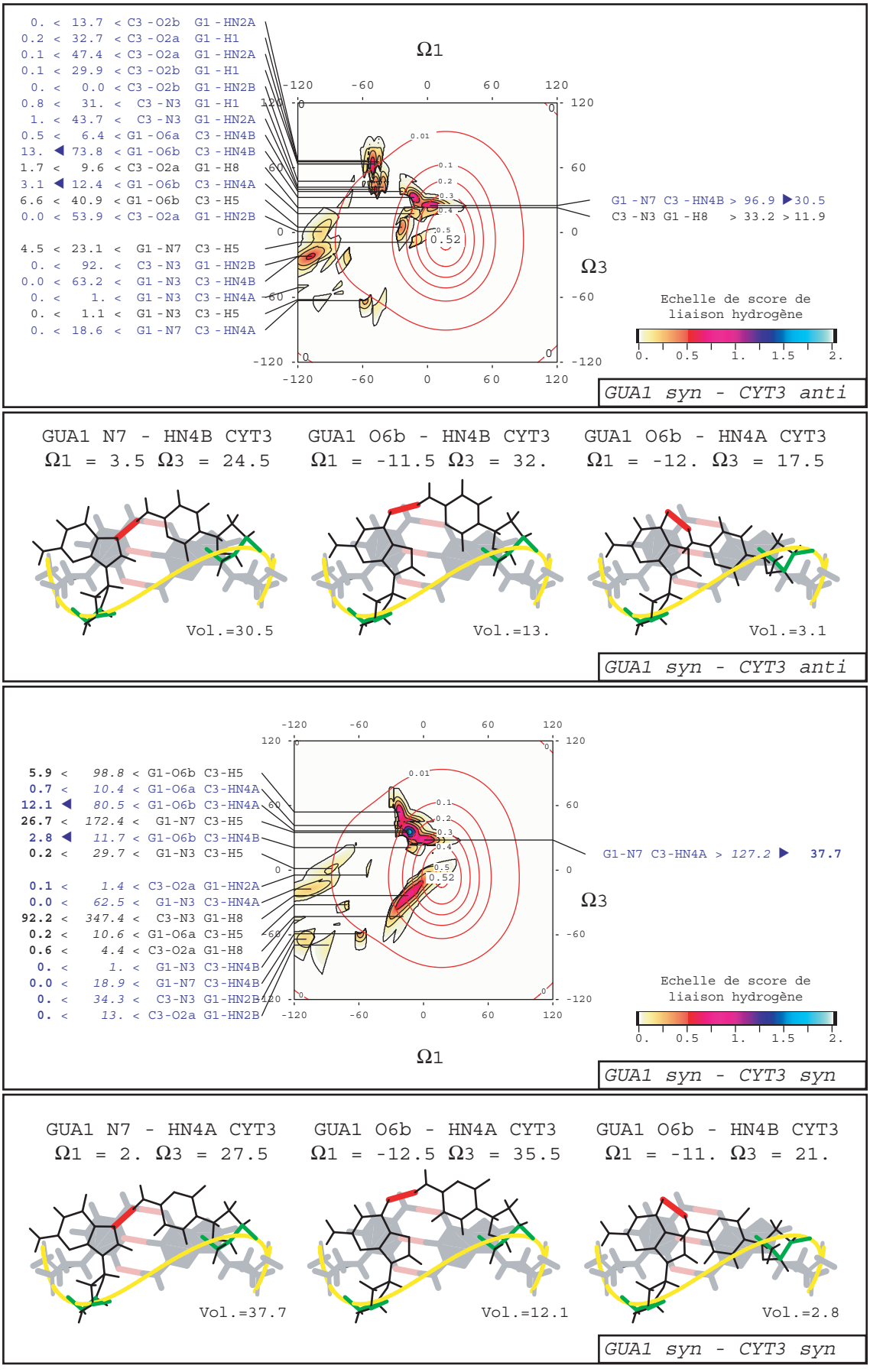


FIG. IV.17 : *Appariements GUA···CYT en conformation* SYN-ANTI *et* SYN-SYN *- Cartes de score de liaison hydrogène et meilleurs appariements.*

Chapitre V

Conclusions et perspectives générales

Conclusions

Dans ce travail de thèse, j'ai étudié la conformation de plusieurs séries de tri- et tétra-boucles d'ADN et d'ARN. Le but était double. D'une part il s'agissait de délimiter la validité de l'approche BCE (Biopolymer Chain Elasticity) pour des boucles d'acides nucléiques de longueur et de nature différentes. D'autre part, il s'agissait de développer de nouveaux outils théoriques et informatiques, et d'explorer les facteurs qui président à la structuration des boucles comportant des appariements.

L'approche BCE est fondée sur un principe de minimum d'énergie qui se décline à différentes échelles. Au cours de ma thèse j'ai procédé de la façon suivante :

(1) **À l'échelle de plusieurs nucléotides :** j'ai vérifié que la trajectoire de la chaîne sucre-phosphate est une trajectoire de moindre énergie donnée par la théorie de l'élasticité des barres minces (*cf.* CHAPT. : III) pour une série d'épingles à cheveux comportant des appariements dans la boucle, d'ADN et d'ARN, comportant entre 3 et 4 nucléotides,

(2) **À l'échelle du nucléotide :** Pour ces boucles, j'ai montré que :

(2.1) le placement en rotation des bases et des blocs d'atomes autour du fil à l'échelle atomique est obtenu en répartissant de façon uniforme la torsion physique imposée par des contraintes externes selon la théorie de l'élasticité (*cf.* CHAPT. : II),

(2.2) le placement en rotation et en flexion des nucléotides N_1 et N_3 formant l'appariement dans la boucle satisfait une règle de déformation minimale (*cf.* CHAPT. : IV).

(3) **À l'échelle de la paire de bases :** j'ai montré que pour les tri-boucles étudiées (*cf.* CHAPT. : IV), l'appariement observé expérimentalement est celui qui forme les meilleures liaisons hydrogène pour une géométrie donnée de la chaîne sucre-phosphate.

(4) **À l'échelle des angles de torsion :** j'ai vérifié, pour les boucles étudiées aux chapitres III et IV, que les points 1. et 2. entraînent la conservation des angles de torsion observés dans les structures expérimentales à l'exception de la zone du "sharp-turn".

(5) **À l'échelle atomique :** j'ai vérifié que les structures théoriques BCE trouvées sont très proches d'un minimum énergétique atteint par minimisation d'énergie.

<u>À l'échelle de plusieurs nucléotides :</u> La modélisation de la chaîne sucre-phosphate comme une barre mince rigide est pertinente.

L'approche Biopolymer Chain Elasticity est une approche simple, efficace et explicative de modélisation des chaînes polymériques d'acides nucléiques. Nous avons montré au chapitre III que l'approche BCE permet de prédire la trajectoire tridimensionnelle de toute une série de tri-boucles d'ADN (1BJH-AAA-, 1XUE-GCA- et 1ZHU-GCA), d'une tétra-boucle d'ADN (1AC7-GTTA) et de plusieurs tétra-boucles d'ARN de séquence -UUCG- (1AUD-UUCG, 1B36-UUCG, 1C0O-UUCG et 1HLX-UUCG) résolues à ce jour. La réussite de notre approche montre donc que pour toutes ces structures, la chaîne simple-brin se comporte en première approximation comme une barre mince flexible et inextensible dont la trajectoire est calculable par la théorie de l'élasticité. Deux conclusions sont importantes :

- Le squelette des acides nucléiques présente des propriétés élastiques pour des simple-brins en boucle dont la longueur varie de trois à quatre nucléotides.

- La conformation de moindre énergie s'applique de la même manière aux chaînes d'ARN et d'ADN. Le principe d'élasticité unifie le cadre de modélisation et explique simplement les différentes formes prise par les boucles d'ARN et d'ADN par les géométries différentes des tiges en hélices en conformation A ou B.

À l'échelle du nucléotide : Le squelette se comporte en torsion et en flexion comme une barre élastique

J'ai développé un formalisme permettant de répartir l'effet de la torsion physique sur le positionnement des blocs d'atomes. Il montre que la rotation d'un bloc en Ω autour de la tangente au fil doit être compris, du point de vue de l'élasticité, comme une déformation en torsion du fil élastique. Pour les boucles d'ADN étudiées comportant un appariement, l'ensemble des nuléotides de la boucle suit une déformation quasi-uniforme en torsion le long du fil jusqu'à la région du "sharp-turn". À cet endroit, une modification forte des angles de torsion compris entre les deux derniers nucléotides de la boucle pourrait relacher la torsion physique.

À l'échelle de la paire de bases : Les appariements dans les boucles sont les meilleurs appariements.

La mise en place de la fonction de score de liaison hydrogène permet d'évaluer, dans cet espace conformationnel, la formation d'appariements. Cette exploration tient compte des contraintes géométriques imposées par la trajectoire particulière de la chaîne sucre-phosphate dans la partie en boucle, et prend en compte la nature et la conformation des bases. Elle permet de retrouver tous les appariements déjà publiés.

La fonction de score de liaison hydrogène discrimine quantitativement les appariements les plus stables et montre ainsi que les appariements décrits ne doivent plus être considérés comme des mésappariements, mais comme les meilleurs appariements possibles étant donné les contraintes géométriques imposées par la trajectoire de la chaîne sucre-phosphate.

Les trajectoires prédites par la théorie de l'élasticité des barres minces dérivent d'un principe de minimum d'énergie, mais sont indépendantes de la nature du

matériau, *i.e.* du module de Young de la barre. Dans cette étude nous avons donc pu calculer la structure tridimensionnelle complète des épingles à cheveux en fonction des tangentes aux extrémités de la double hélice. Ces calculs de trajectoire ont donc été réalisés uniquement à partir de considérations géométriques. Au moyen de l'exploration de l'espace des conformations des bases B_1 et B_3 de la boucle et grâce à la qualité des appariements formés, nous avons été amené à postuler des rigidités en torsion et en flexion de chaîne sucre-phosphate, caractérisées par une longueur de persistance commune de l'ordre de trois nucléotides.

À l'échelle atomique : Calcul et prédiction des structures atomiques complètes des tri-boucles d'ADN comportant un appariement dans la boucle

À partir des paramètres calculés lors de l'exploration des appariements, il est possible de construire, *a priori,* l'ensemble de la structure de l'épingle à cheveux comportant un appariement dans la boucle. Nous avons montré que cette approche théorique permet de retrouver, après minimisation sous contrainte d'un, deux ou trois angles de torsion, la structure complète des tri-boucles d'ADN résolues à ce jour. Tous ces angles de torsion sont localisés dans la zone de plus forte courbure, le "sharp-turn".

L'approche BCE : une approche hiérarchique et multi-échelle ouverte.

L'approche BCE prend en compte les caractéristiques structurales prédominantes de la molécule propres aux différentes échelles, globale, intermédiaire et atomique. Ainsi il est possible de mettre en place une modélisation efficace en faisant appel à un petit nombre de degrés de liberté adaptés aux opérations de déformation à chaque échelle.

Les d.d.l. Ω et χ sont suffisants pour mettre en place les conformations des bases dans les boucles. Ils permettent de former les appariements qui apparaissent dans les parties en boucle. Le développement et l'utilisation du d.d.l. Θ_{empil} permet de tenir compte de l'encombrement stérique dans les conformations où les bases de la boucle sont empilées sur le plateau de l'hélice de la tige. Ce d.d.l. permet ainsi d'explorer simplement l'espace des conformations propices à la formation d'appariements. La

simplicité de cette approche de modélisation permet d'explorer systématiquement cet espace conformationnel au moyen de paramètres de modélisation qui peuvent être utilisés comme paramètres de description.

Perspectives

La prédiction complète de la conformation des tri-boucles

L'exploration doit prendre en compte toutes les possibilités de conformations alternatives. Des conformations stables sont par exemple rencontrées dans les boucles de séquence -TTT-. En effet, dans ces boucles, la première thymine de la boucle se place en conformation non-appariée et non-empilée dans le petit sillon où elle engage des liaisons hydrogène avec les dernier et avant-dernier plateaux de la tige. L'approche devra également exclure les conformations qui présentent des mauvais contacts. Elle devra également éventuellement prendre en compte la qualité des empilements (aire de recouvrement des surfaces des cycles), voire l'exposition au solvant. Pour généraliser l'aspect prédictif, il faut donc enrichir le modèle.

Explorations des appariements dans les hélices et les autres boucles

Ce type d'étude pourrait être reconduit aux cas des tétra-boucles d'ADN, aux tri- et tétra-boucles d'ARN, aux boucles formées par des simple-brins repliés en triple ou quadruple hélices. L'approche peut également être mise à profit pour explorer les mésappariements stables dans les hélices et plus généralement dans les boucles internes.

À plus long terme, ce type d'approche devrait permettre d'étudier des boucles plus complexes telles que la boucle de l'anti-codon, la boucle de reconnaissance de HIV-tar, les aptamères...

Une approche conceptuellement simple qui se prête à un enrichissement progressif

Il est remarquable que nous ayons pu construire complètement des structures aussi complexes que les tri-boucles d'ADN avec aussi peu de contraintes (liaisons hydrogène, torsion et flexion). Nos résultats montrent qu'elles suffisent pour ces conformations. Pour être capable de traîter d'autres cas, il faudra certainement tenir compte d'autres facteurs tels quel l'hydrophobicité, les interactions électrostatiques, l'empilement ou les mauvais contacts atomiques les plus grossiers à l'échelle mésoscopique.

Développement d'un champ de force mésoscopique

Dans cette étude nous avons utilisé les propriétés géométriques de la théorie de l'élasticité pour calculer la trajectoire globale du fil. À plus long terme, il devrait être possible de substituer les conditions géométriques d'encastrement par des contraintes de forces sur le fil ou sur les bases en tenant compte des termes d'énergie de torsion ou de flexion du squelette.

Annexe

A. Paramètres quantitatifs de description des hélices d'acides nucléiques

A.1 Calcul de RMSd

Plusieurs formules sont utilisées dans la littérature sous le terme de RMSd. Le RMSdisp est adapté à la comparaison de deux molécules, alors que le RMS_{dev} est plus général et peut être employé pour plusieurs molécules. Ces acronymes anglo-saxons se développent de la manière suivante :

- "Root Mean Square displacement". Cette formule permet de calculer la racine carrée de la moyenne des distances au carré entre atome homologues de deux molécules de composition et de topologie identique :

$$RMS_{disp}[mol_1, mol_2] = \sqrt{\frac{1}{n_{ato}} \sum_{1=1}^{i=n_{ato}} d^2_{1,2}(i)}$$

où :

mol_1 et mol_2, les deux molécules sur lesquelles le calcul est réalisé,

n_{ato}, le nombre d'atomes dans chaque molécule,

i, le numéro de l'atome, et

$d_{1,2}(i)$, la distance entre les deux atomes i des molécules mol_1et mol_2.

- "Root Mean Square deviation". Cette formule permet de calculer la racine carrée de la moyenne des écarts des distances homologues entre les structure et une structure moyenne :

$$RMS_{dev}[mol_1, mol_2, \ ..., mol_n] = \sqrt{\frac{1}{n_{ato}} \sum_{1=1}^{i=n_{ato}} \left(\frac{1}{n} \sum_{j=1}^{j=n} d_{j,av}^2(i) \right)}$$

où :

mol_i, les molécules sur lesquelles le calcul est réalisé,

n, le nombre de molécules,

n_{ato}, le nombre d'atome des molécules,

i, le numéro de l'atome,

j, le numéro de la molécule

$d_{j,av}(i)$, la distance entre l'atome i de la molécule j et l'atome i de la molécule moyenne. La position de l'atome i correspond au barycentre des atomes homologues des molécules sur lesquelles le calcul est réalisé.

A.2 Description des appariements et des empilements dans les hélices d'acides nucléiques

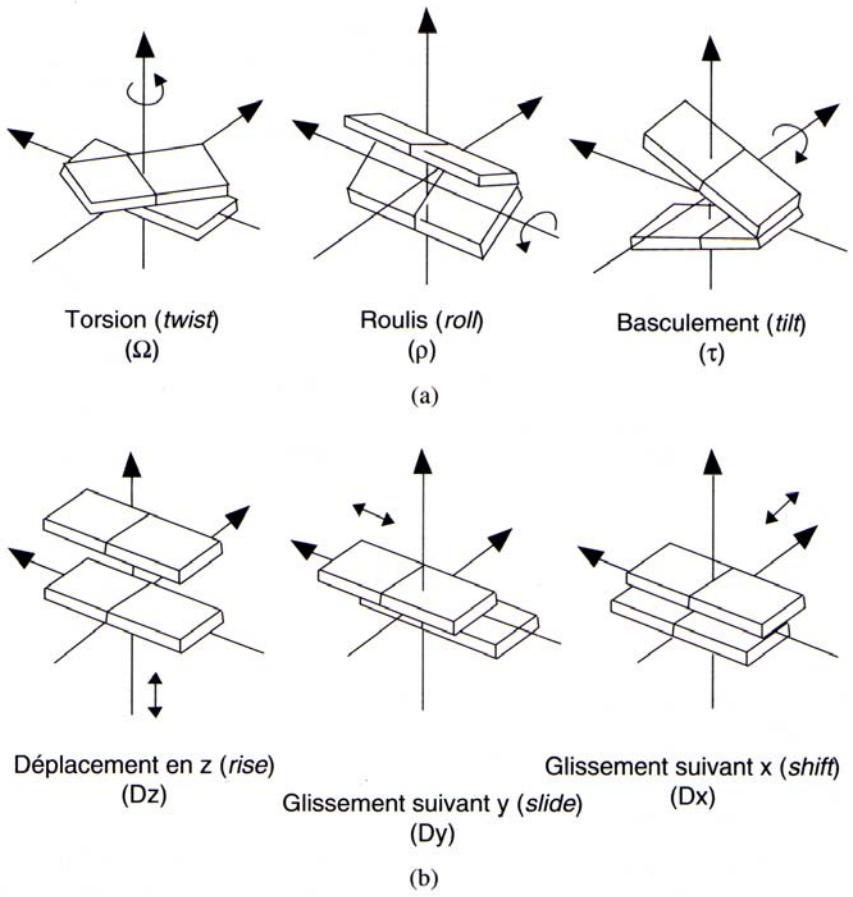


FIG. A.2.1 : ***Définition des angles de rotation et des translations de deux paires de bases l'une par rapport à l'autre suivant les trois axes x, y et z de la convention de Cambridge :*** *(a) les rotations; (b) les translations.*

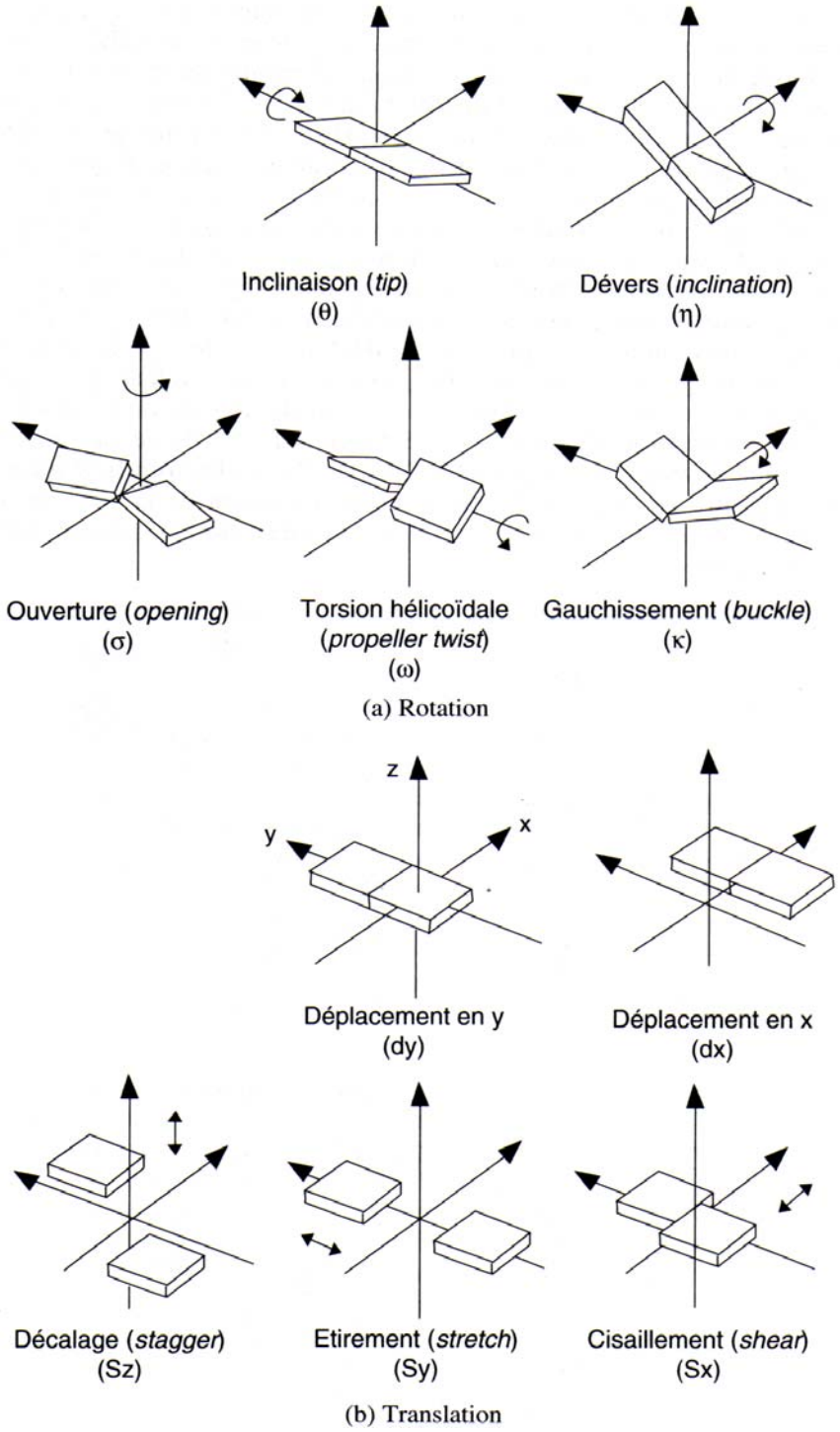
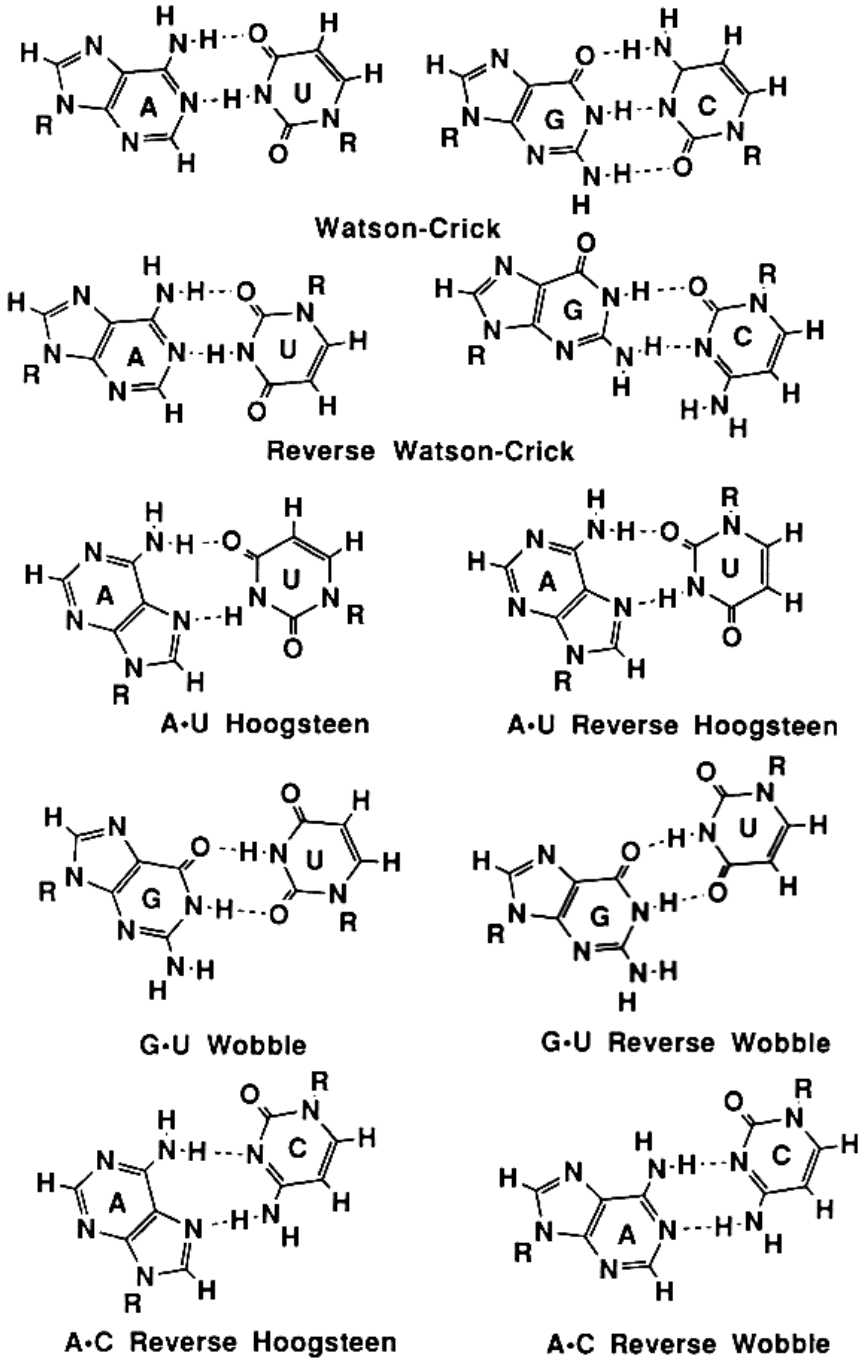


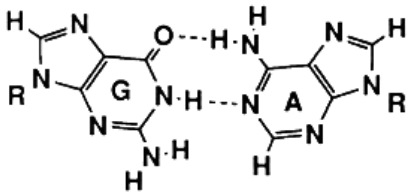
FIG. A.2.2 : ***Définition des angles de rotation et des translations d'une base par rapport à l'autre à l'intérieur d'une paire, suivant les trois axes x, y et z de la convention de Cambridge :*** *(a) les rotations; (b) les translations.*

B. Description des Appariements et des mésappariements

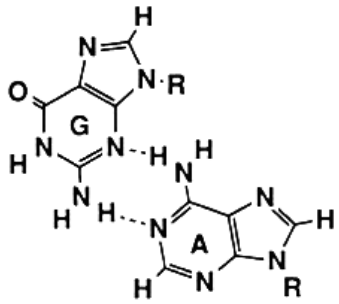
B.1 Appariements Watson-Crick et Appariements Hoogsteen et Wobble



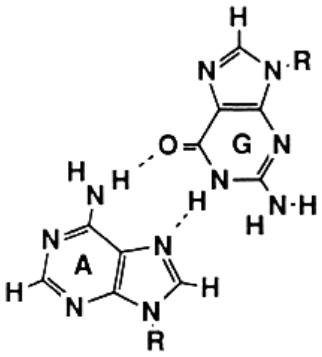
B.2 Mésappariements hétéro-puriques



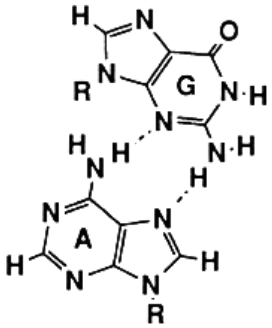
G•A N1-N1,
carbonyl-amino



G•A N3-amino,
amino-N1

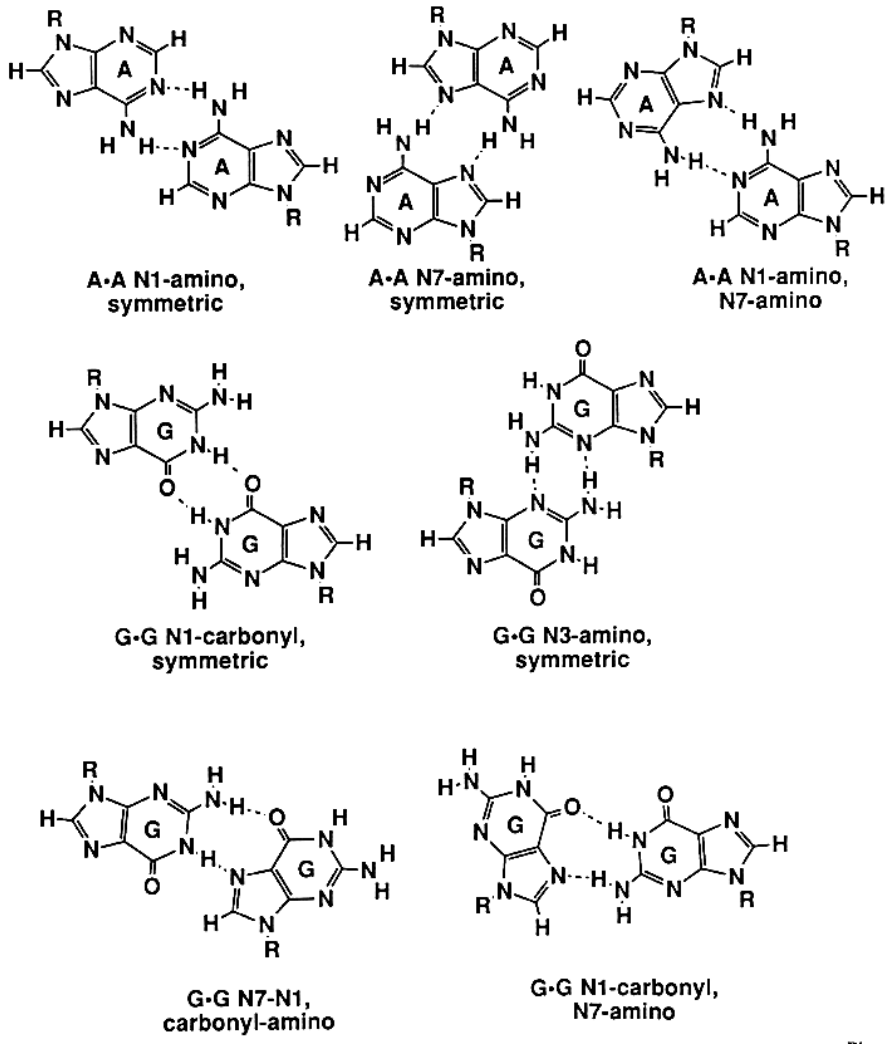


A•G N7-N1,
amino-carbonyl

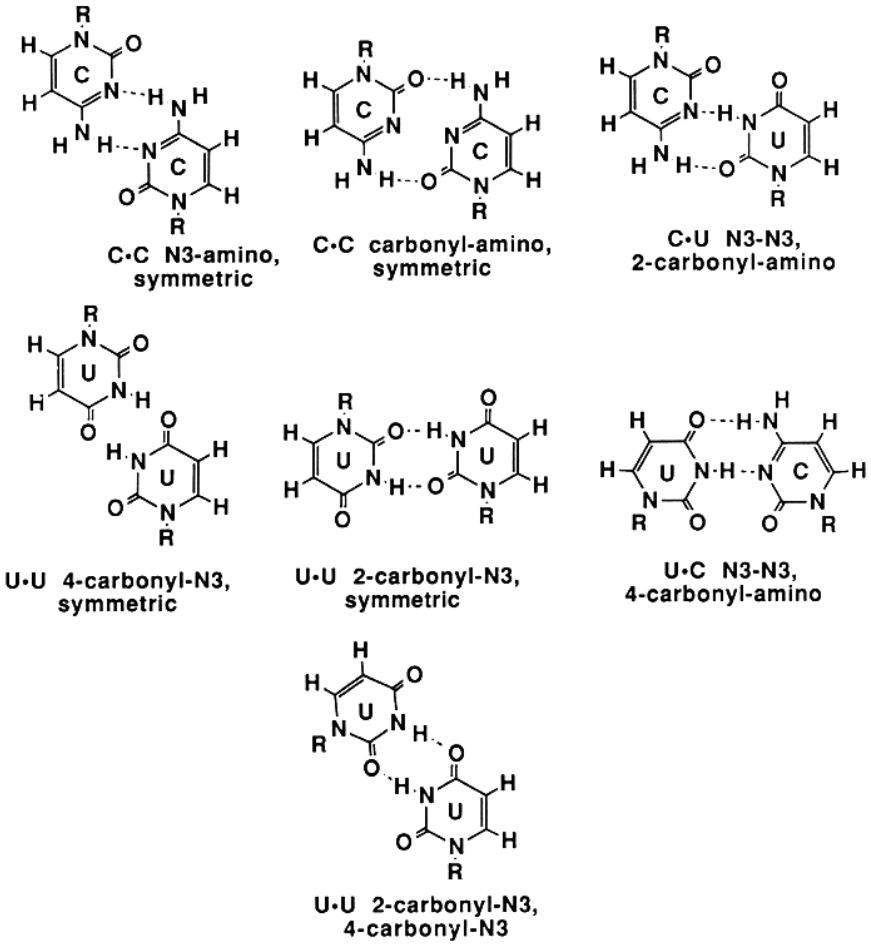


A•G N7-amino,
amino-N3

B.3 Mésappariements homo-puriques



B.4 Mésappariements pyrimidiques



Bibliographie

[1] Dickerson, R. E., Bansal, M., Calladine, C. R., Diekmann, S., Hunter, W. N., Kennard, O. & et al. (1989) Definitions and nomenclature of nucleic acid structure parameters. *EMBO J,* **8**, 1–4.

[2] Nowakowski, J. & Tinoco, I. J. (1999) Oxford Handbook of Nucleic Acid Structure - RNA structure in solution, Oxford University Press, .

[3] Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res,* **31**(13), 3406–3415.

[4] Pakleza, C. Méthodologie du repliement de l'ADN à différentes échelles : 1/ Modélisation moléculaire des épingles à cheveux à partir de l'élasticité et de contraintes RMN, 2/ Analyse et mesure de la flexibilité à partir de microscopies. PhD thesis Université Pierre et Marie Curie Paris (2002).

[5] Bissler, J. J. (1998) DNA inverted repeats and human disease. *Front Biosci,* **3**, d408–418.

[6] Sinden, R. S. (1994) DNA structure and function, Academic Press, San Diego.

[7] Todd, A., Cossons, N., Aitken, A., Price, G. B. & Zannis-Hadjopoulos, M. (1998) Human Cruciform Binding Protein Belongs to the 14-3-3 Family. *Biochemistry,* **37**, 14317–14325.

[8] Pearson, C. E., Zorbas, H., Price, G. B. & Zannis-Hadjopoulos, M. (1996) Inverted repeats, stem-loops, and cruciforms: significance for initiation of DNA replication. *J Cell Biochem,* **63**(1), 1–22.

[9] Dai, X. & Rothman-Denes, L. B. (1999) DNA structure and transcription. *Curr Opin Microbiol,* **2**(2), 126–130.

[10] Reddy, M. S., Vaze, M. B., Madhusudan, K. & Muniyappa, K. (2000) Binding of SSB and RecA protein to DNA-containing stem loop structures: SSB ensures the polarity of RecA polymerization on single-stranded DNA. *Biochemistry,* **39**(46), 14250–14262.

[11] Moore, H., Greenwell, P. W., Liu, C. P., Arnheim, N. & Petes, T. D. (1999) Triplet repeats form secondary structures that escape DNA repair in yeast. *Proc Natl Acad Sci USA,* **96**(4), 1504–1509.

[12] Ren, J., Qu, X., Chaires, J. B., Trempe, J. P., Dignam, S. S. & Dignam, J. D. (1999) Spectral and physical characterization of the inverted terminal repeat DNA structure from adenoassociated virus 2. *Nucleic Acids Res,* **27**(9), 1985–1990.

[13] Ryan, J. H., Zolotukhin, S. & Muzyczka, N. (1996) Sequence requirements for binding of Rep68 to the adeno-associated virus terminal repeats. *J Virol,* **70**(3), 1542–1553.

[14] Chou, S. H., Tseng, Y. Y. & Chu, B. Y. (2000) Natural abundance heteronuclear NMR studies of the T3 mini-loop hairpin in the terminal repeat of the adenoassociated virus 2. *J Biomol NMR,* **17**(1), 1–16.

[15] Samani, T. D., Jollès, B. & Laigle, A. (2001) Best minimally modified antisense oligonucleotides according to cell nuclease activity. *Antisense Nucleic Acid Drug Dev,* **11**(3), 129–136.

[16] Maksimenko, A. V., Gottikh, M. B., Helin, V., Shabarova, Z. A. & Malvy, C. (1999) Physico-chemical and biological properties of antisense phosphodiester oligonucleotides with various secondary structures. *Nucleosides Nucleotides,* **18**(9), 2071–2091.

[17] Hirao, I., Kawai, G., Yoshizawa, S., Nishimura, Y., Ishido, Y., Watanabe, K. & Miura, K. (1994) Most compact hairpin-turn structure exerted by a short DNA fragment, d(GCGAAGC) in solution: an extraordinarily stable structure resistant to nucleases and heat. *Nucleic Acids Res,* **22**(4), 576–582.

[18] Hermann, T. & Patel, D. J. (2000) Adaptive recognition by nucleic acid aptamers. *Science,* **287**(5454), 820–825.

[19] Boiziau, C., Dausse, E., Yurchenko, L. & Toulme, J. J. (1999) DNA aptamers selected against the HIV-1 trans-activation-responsive RNA element form RNA-DNA kissing complexes. *J Biol Chem,* **274**(18), 12730–12737.

[20] Collin, D., van Heijenoort, C., Boiziau, C., Toulmé, J. J. & Guittet, E. (2000) NMR characterization of a kissing complex formed between the TAR RNA element of HIV-1 and a DNA aptamer. *Nucleic Acids Res,* **28**(17), 3386–3391.

[21] Paddison, P. J., Caudy, A. A., Bernstein, E., Hannon, G. J. & Conklin, D. S. (2002) Short hairpin RNAs (shRNAs) induce sequence-specific silencing in mammalian cells. *Genes Dev,* **16**(8), 948–958.

[22] Hannon, G. J. (2002) RNA interference. *Nature,* **418**(6894), 244–251.

[23] Chandrasekaran, R. & Arnott, S. (1990) The Structures of DNA and RNA Helices in Oriented Fibers, Landolt-Bornstein Numerical Data and Functional Relationships in Science and TechnologySaenger, W., Springer-Verlag, .

[24] Fuller, W., Wilkins, M. H. F., Wilson, H. R. & Hamilton, L. D. (1965) The molecular configuration of deoxyribonucleic acid. IV. X-ray diffraction study of the A form. *J Mol Biol,* **12**, 60–80.

[25] Arnott, S. & Hukins, D. W. (1973) Refinement of the structure of B-DNA and implications for the analysis of x-ray diffraction data from fibers of biopolymers. *J Mol Biol,* **81**(2), 93–105.

[26] Chandrasekaran, R., Arnott, S., He, R.-G., Millane, R. P., Park, H.-S., Puigjaner, L. C. & Walker, J. K. (1985) More complex DNA structures. *J Macromol Sci Phys,* **24**(B), 1–20.

[27] Arnott, S., Hukins, D. W., Dover, S. D., Fuller, W. & Hodgson, A. R. (1973) Structures of synthetic polynucleotides in the A-RNA and A'-RNA conformations: x-ray diffraction analyses of the molecular conformations of polyadenylic acid–polyuridylic acid and polyinosinic acid–polycytidylic acid. *J Mol Biol,* **81**(2), 107–122.

[28] van Dam, L. & Levitt, M. H. (2000) BII nucleotides in the B and C forms of natural-sequence polymeric DNA: A new model for the C form of DNA. *J Mol Biol,* **304**(4), 541–561.

[29] Watson, J. D. & Crick, F. H. C. (1953) Genetic implication of the structure of deoxyribonucleic acid. *Nature,* **171**, 964–967.

[30] Watson, J. D. & Crick, F. H. C. (1953) Molecular structure of nucleic acids. *Nature,* **171**, 737–739.

[31] Saenger, W. (1984) Principles of nucleic acids structure, Springer advanced text in chemistrySpringer-Verlag, Berlin, .

[32] Leach, A. R. (2001) Molecular Modelling : Principles and Applications, Prentice-Hall, New Upper Saddle River, New Jersey, .

[33] Weissig, H. & Bourne, P. E. (2002) Protein structure resources. *Acta Crystallogr D Biol Crystallogr,* **58**(Pt 6 No 1), 908–915.

[34] Gralla, J. & Crothers, D. M. (1973) Free energy of imperfect nucleic acid helices. 3. Small internal loops resulting from mismatches. *J Mol Biol,* **78**(2), 301–319.

[35] Uhlenbeck, O. C., Borer, P. N., Dengler, B. & Tinoco, I. J. (1973) Stability of RNA hairpin loops: A 6 -C m -U 6. *J Mol Biol,* **73**(4), 483–496.

[36] Varani, G. (1995) Exceptionally stable nucleic acid hairpins. *Annu Rev Biophys Biomol Struct,* **24**, 379–404.

[37] Hilbers, C. W., Heus, H. A., van Dongen, M. J. P. & Wijmenga, S. S. (1994) The hairpin elements of nucleic acid structure : DNA and RNA folding. In Eckstein, F. and Lilley, D. (eds). *Nucleic Acids and Molecular Biology. Springer-Verlag, Berlin,* **8**, 56–104.

[38] Auffinger, P., Louise-May, S. & Westhof, E. (1996) Molecular dynamics simulations of the anticodon hairpin of tRNAAsp : Structuring effects of C-H...O hydrogen bonds and of long-range hydration forces. *J Am Chem Soc,* **118**, 1181–1189.

[39] Auffinger, P., Bielecki, L. & Westhof, E. (2004) Anion binding to nucleic acids. *Structure (Camb),* **12**(3), 379–388.

[40] Auffinger, P., Bielecki, L. & Westhof, E. (2004) Symmetric K+ and Mg2+ ion-binding sites in the 5S rRNA loop E inferred from molecular dynamics simulations. *J Mol Biol,* **335**(2), 555–571.

[41] Auffinger, P., Bielecki, L. & Westhof, E. (2001) Hydrophobic Groups Stabilize the Hydration Shell of 2'-O-Methylated RNA Duplexes. *Angew Chem Int Ed Engl,* **40**(24), 4648–4650.

[42] Pearlman, D. A., Case, D. A., Caldwell, J. W., Ross, W. R., T.E. Cheatham, T. E., DeBolt, S., Ferguson, D., Seibel, G. & Kollman, P. (1995) AMBER, a

computer program for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to elucidate the structures and energies of molecules. *Comp Phys Commun,* **91**, 1–41.

[43] Brooks, B. R., Bruccoleri, R. E., D, O. B., States, D. J., Swaminathan, S. & Karplus, M. (1983) A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J Comp Chem,* **4**, 187–217.

[44] MacKerell, A. D. J., Brooks, B., Brooks, C. L. I., Nilsson, L., Roux, B., Won, Y. & Karplus, M. (1998) The Energy Function and Its Parameterization with an Overview of the ProgramVol. 1, of The Encyclopedia of Computational Chemistry John Wiley and Sons, Chichester, .

[45] D, C. W., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M. J., Ferguson, D M end Spellmeyer, D. C., Fox, T., Caldwell, J. W. & Kollman, P. A. (1995) A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J Am Chem Soc,* **117**, 5179–5197.

[46] Cheatham, T. E. I. & Young, M. A. (2001) Molecular dynamics simulation of nucleic acids: Successes, limitations and promise. *Biopolymers,* **56**, 232–256.

[47] Chou, S. H., Zhu, L., Gao, Z., Cheng, J. W. & Reid, B. R. (1996) Hairpin loops consisting of single adenine residues closed by sheared A.A and G.G pairs formed by the DNA triplets AAA and GAG: solution structure of the d(GTACAAAGTAC) hairpin. *J Mol Biol,* **264**(5), 981–1001.

[48] Güntert, P., Braun, W. & Wüthrich, K. (1991) Efficient computation of three-dimensional protein structures in solution from nuclear magnetic resonance data using the program DIANA and the supporting programs CALIBA, HABAS and GLOMSA. *J Mol Biol,* **217**(3), 517–530.

[49] Lavery, R., Zakrzewska, K. & Sklenar, H. (1995) JUMNA: Junction Minimisation of Nucleic Acids. *Comp Phys Comm,* **91**, 135–158.

[50] Guntert, P., Mumenthaler, C. & Wuthrich, K. (1997) Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J Mol Biol,* **273**(1), 283–298.

[51] Lu, X.-J. & Olson, W. K. (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res,* **31**(17), 5108–5121.

[52] Wing, R., Drew, H., Takano, T., Broka, C., Tanaka, S., Itakura, K. & Dickerson, R. E. (1980) Crystal structure analysis of a complete turn of B-DNA. *Nature,* **287**(5784), 755–758.

[53] Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S. H., Srinivasan, A. R. & Schneider, B. (1992) The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys J,* **63**(3), 751–759.

[54] van Dongen, M. J., Mooren, M. M., Willems, E. F., van der Marel, G. A., van Boom, J. H., Wijmenga, S. S. & Hilbers, C. W. (1997) Structural features of the DNA hairpin. *Nucleic Acids Res,* **25**(8), 1537–1547.

[55] Zhu, L., Chou, S. H. & Reid, B. R. (1996) A single G-to-C change causes human centromere TGGAA repeats to fold back into hairpins. *Proc Natl Acad Sci USA,* **93**(22), 12159–12164.

[56] Zhu, L., Chou, S. H., Xu, J. & Reid, B. R. (1995) Structure of a single-cytidine hairpin loop formed by the DNA triplet GCA. *Nat Struct Biol,* **2**(11), 1012–1017.

[57] Allain, F. H., Howe, P. W., Neuhaus, D. & Varani, G. (1997) Structural basis of the RNA-binding specificity of human U1A protein. *EMBO J,* **16**(18), 5764–5772.

[58] Butcher, S. E., Allain, F. H. & Feigon, J. (1999) Solution structure of the loop B domain from the hairpin ribozyme. *Nat Struct Biol,* **6**(3), 212–216.

[59] Colmenarejo, G. & Tinoco, I. J. (1999) Structure and thermodynamics of metal binding in the P5 helix of a group I intron ribozyme. *J Mol Biol,* **290**(1), 119–135.

[60] Allain, F. H. & Varani, G. (1995) Structure of the P1 helix from group I self-splicing introns. *J Mol Biol,* **250**(3), 333–353.

[61] Gunderson, S. I., Vagner, S., Polycarpou-Schwarz, M. & Mattaj, I. W. (1997) Involvement of the carboxyl terminus of vertebrate poly(A) polymerase in U1A autoregulation and in the coupling of splicing and polyadenylation. *Genes Dev,* **11**(6), 761–773.

[62] Ulyanov, N. B., Bauer, W. R. & James, T. L. (2002) High-resolution NMR structure of an AT-rich DNA sequence. *J Biomol NMR,* **22**(3), 265–280.

[63] Padrta, P., Stefl, R., Kralik, L., Zidek, L. & Sklenar, V. (2002) Refinement of d(GCGAAGC) hairpin structure using one- and two-bond residual dipolar couplings. *J Biomol NMR,* **24**(1), 1–14.

[64] Chin, K.-H. & Chou, S.-H. (2003) Sheared-type G(anti).C(syn) base-pair: a unique d(GXC) loop closure motif. *J Mol Biol,* **329**(2), 351–361.

[65] Amir-Aslani, A., Mauffret, O., Sourgen, F., Neplaz, S., Maroun, R. G., Lescot, E., Tevanian, G. & Fermandjian, S. (1996) The hairpin structure of a topoisomerase II site DNA strand analyzed by combined NMR and energy minimization methods. *J Mol Biol,* **263**(5), 776–788.

[66] Amir-Aslani, A., Mauffret, O., Bittoun, P., Sourgen, F., Monnot, M., Lescot, E. & Fermandjian, S. (1995) Hairpins in a DNA site for topoisomerase II studied by 1H- and 31P-NMR. *Nucleic Acids Res,* **23**(19), 3850–3857.

[67] Chou, S. H., Tseng, Y. Y. & Wang, S. W. (1999) Stable sheared A.C pair in DNA hairpins. *J Mol Biol,* **287**(2), 301–313.

[68] Yoshizawa, S., Kawai, G., Watanabe, K., Miura, K. & Hirao, I. (1997) GNA trinucleotide loop sequences producing extraordinarily stable DNA minihairpins. *Biochemistry,* **36**(16), 4761–4767.

[69] Olson, W. K., Bansal, M., Burley, S. K., Dickerson, R. E., Gerstein, M., Harvey, S. C., Heinemann, U., Lu, X. J., Neidle, S., Shakked, Z., Sklenar, H., Suzuki, M., Tung, C. S., Westhof, E., Wolberger, C. & Berman, H. M. (2001) A standard reference frame for the description of nucleic acid base-pair geometry. *J Mol Biol,* **313**(1), 229–237.

[70] Daune, M. (1997) Biophysique moléculaire, Interéditions UDunod, .

[71] Saenger, W. (1984) Principles of nucleic acids structure, Springer advanced text in chemistrySpringer-Verlag, Berlin, .

[72] Rao, S. T. & Sundaralingam, M. (1970) Stereochemistry of nucleic acids and their constituents. 13. The crystal and molecular structure of 3'-O-acetyladenosine. Conformational analysis of nucleosides and nucleotides with syn glycosidic torsional angle. *J Am Chem Soc,* **92**(16), 4963–4970.

[73] Leonard, G. A., McAuley-Hecht, K., Brown, T. & Hunter, W. N. (1995) Do C-H...O hydrogen bonds contribute to the stability of nucleic acid base pairs?. *Acta Crystallogr D Biol Crystallogr,* **51**(Pt 2), 136–139.

[74] Chou, S. H. & Reid, B. R. (1999) Oxford Handbook of Nucleic Acid Structure - DNA mismatches in solution, Oxford University Press, .

[75] Cognet, J. A., Gabarro-Arpa, J., Le Bret, M., van der Marel, G. A., van Boom, J. H. & Fazakerley, G. V. (1991) Solution conformation of an oligonucleotide containing a G.G mismatch determined by nuclear magnetic resonance and molecular mechanics. *Nucleic Acids Res,* **19**(24), 6771–6779.

[76] Cognet, J. A., Boulard, Y. & Fazakerley, G. V. (1995) Helical parameters, fluctuations, alternative hydrogen bonding, and bending in oligonucleotides containing a mistmatched base-pair by NOESY distance restrained and distance free molecular dynamics. *J Mol Biol,* **246**(1), 209–226.

[77] Fazakerley, G. V. & Boulard, Y. (1995) DNA mismatches and modified bases. *Methods Enzymol,* **261**, 145–163.

[78] Boulard, Y., Cognet, J. A. & Fazakerley, G. V. (1997) Solution structure as a function of pH of two central mismatches, C . T and C . C, in the 29 to 39 K-ras gene sequence, by nuclear magnetic resonance and molecular dynamics. *J Mol Biol,* **268**(2), 331–347.

[79] Chou, S.-H., Chin, K.-H. & Wang, A. H.-J. (2003) Unusual DNA duplex and hairpin motifs. *Nucleic Acids Res,* **31**(10), 2461–2474.

[80] Pakleza, C. & Cognet, J. A. H. (2003) Biopolymer Chain Elasticity: A novel concept and a least deformation energy principle predicts backbone and overall folding of DNA TTT hairpins in agreement with NMR distances. *Nucleic Acids Res,* **31**(3), 1075–1085.

[81] Santini, G. P. H., Pakleza, C. & Cognet, J. A. H. (2003) DNA tri- and tetra-loops and RNA tetra-loops hairpins fold as elastic biopolymer chains in agreement with PDB coordinates. *Nucleic Acids Res,* **31**(3), 1086–1096.

[82] Antao, V. P., Lai, S. Y. & Tinoco, I. J. (1991) A thermodynamic study of unusually stable RNA and DNA hairpins. *Nucleic Acids Res,* **19**(21), 5901–5905.

[83] Prive, G. G., Yanagi, K. & Dickerson, R. E. (1991) Structure of the B-DNA decamer C-C-A-A-C-G-T-T-G-G and comparison with isomorphous decamers C-C-A-A-G-A-T-T-G-G and C-C-A-G-G-C-C-T-G-G. *J Mol Biol,* **217**(1), 177–199.

[84] Yanagi, K., Prive, G. G. & Dickerson, R. E. (1991) Analysis of local helix geometry in three B-DNA decamers and eight dodecamers. *J Mol Biol,* **217**(1), 201–214.

[85] Love, A. E. H. (1994) A treatise on the mathematical theory of elasticity, New York Dover, 4th edition.

[86] Landau, L. & Lifchitz, E. (1990) Physique théorique - Théorie de l'élasticité, MIR, Moscou 2ème edition.

[87] Miller, J. L., Cheatham, T. E. I. & Kollman, P. A. (1999) Oxford Handbook of Nucleic Acid Structure - Simulation of nucleic acid, Oxford University Press, .

[88] Wang, W., Donini, O., Reyes, C. M. & Kollman, P. A. (2001) Biomolecular simulations: recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions. *Annu Rev Biophys Biomol Struct,* **30**, 211–243.

[89] Kollman, P. A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., Donini, O., Cieplak, P., Srinivasan, J., Case, D. A. & Cheatham, T. E. r. (2000) Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc Chem Res,* **33**(12), 889–897.

[90] Pan, Y. & MacKerell, A. D. J. (2003) Altered structural fluctuations in duplex RNA versus DNA: a conformational switch involving base pair opening. *Nucleic Acids Res,* **31**(24), 7131–7140.

[91] Noy, A., Perez, A., Lankas, F., Javier Luque, F. & Orozco, M. (2004) Relative flexibility of DNA and RNA: a molecular dynamics study. *J Mol Biol,* **343**(3), 627–638.

[92] Auffinger, P. & Vaiana, A. C. (2005) Handbook of RNA Biochemistry Molecular Dynamics of RNA Systems, Wiley WCH, Weinheim, .

[93] Shi, Y. & Hearst, J. E. (1994) The Kirchhoff elastic rod, the nonlinear Schrödinger equation, and DNA supercoiling. *J. Chem. Phys.,* **101**(6), 5186–5200.

[94] Tobias, I., Coleman, B. D. & Olson, W. K. (1994) The dependence of DNA tertiary structure on end conditions : Theory and implications for topological transitions. *J. Chem. Phys.,* **101**(12), 10990–10996.

[95] Allain, F. H. & Varani, G. (1997) How accurately and precisely can RNA structure be determined by NMR?. *J Mol Biol,* **267**(2), 338–351.

[96] Flory, P. (1969) Statistical Mechanics of Chain Molecules, Wiley, New York, .

[97] Williams, L. D. & Maher, L. J. r. (2000) Electrostatic mechanisms of DNA deformation. *Annu Rev Biophys Biomol Struct,* **29**, 497–521.

[98] Kuznetsov, S. V., Shen, Y., Benight, A. S. & Ansari, A. (2001) A semiflexible polymer model applied to loop formation in DNA hairpins. *Biophys J,* **81**(5), 2864–2875.

[99] Eisenberg, H. & Felsenfeld, G. (1967) Studies of the temperature-dependent conformation and phase separation of polyriboadenylic acid solutions at neutral pH. *J Mol Biol,* **30**(1), 17–37.

[100] Inners, L. D. & Felsenfeld, G. (1970) Conformation of polyribouridylic acid in solution. *J Mol Biol,* **50**(2), 373–389.

[101] Smith, S. B., Cui, Y. & Bustamante, C. (1996) Overstretching B-DNA: the elastic response of individual double-stranded and single-stranded DNA molecules. *Science,* **271**(5250), 795–799.

[102] Rivetti, C., Walker, C. & Bustamante, C. (1998) Polymer chain statistics and conformational analysis of DNA molecules with bends or sections of different flexibility. *J Mol Biol,* **280**(1), 41–59.

[103] Tinland, B., Pluen, A., Sturm, J. & Weill, G. (1997) Persistence lenght of single-stranded DNA. *Macromolecules,* **30**, 5763–5765.

[104] Mills, J. B., Vacano, E. & Hagerman, P. J. (1999) Flexibility of single-stranded DNA: use of gapped duplex helices to determine the persistence lengths of poly(dT) and poly(dA). *J Mol Biol,* **285**(1), 245–257.

Résumé

Biopolymer Chain Elasticity (BCE) est une nouvelle méthodologie de modélisation moléculaire qui assimile le squelette de l'ADN ou de l'ARN à un fil flexible au moyen de la théorie de l'élasticité des barres minces. Avec BCE, nous avons construit les structures de plusieurs familles de molécules en épingle à cheveux comportant un appariement dans la boucle : les tri-boucles d'ADN AAA, GCA, GAA, AGC, ATC et GAC, la tétra-boucle d'ADN GTTA et quatre tétra-boucles d'ARN UUCG. Cette approche reproduit avec succès les différentes trajectoires des squelettes, et décrit les positions de chaque nucléotide de la boucle au moyen de trois degrés de liberté. Ceci démontre les propriétés structurantes en torsion et flexion du squelette à ces échelles. En explorant exhaustivement les espaces conformationnels des bases appariées définis par ces petits nombres de degrés de liberté, nous obtenons par raffinement d'énergie les appariements et les structures complètes de ces molécules à l'échelle atomique.

Mots Clefs : ADN, ARN, RMN, conformations des épingles à cheveux, modélisation moléculaire, BCE, théorie de l'élasticité, appariements, liaisons hydrogène, empilement, exploration conformationnelle.

Abstract

Biopolymer Chain Elasticity (BCE) is a new molecular modeling methodology which identifies the DNA or RNA backbone as a flexible wire by means of the theory of elasticity of thin bars. With BCE, we built the structures of several families of hairpin molecules comprising a base pairing in the loop: the DNA tri-loops AAA, GCA, GAA, AGC, ATC and GAC, the DNA tetra-loop GTTA and four RNA tétra-loops UUCG. This approach reproduces with success the different trajectories of the backbones, and describes the positions of each nucleotide in the loop by means of three degrees of freedom. This demonstrate the structuring properties in terms of torsion and flexion of the backbone at different scales. By exploring exhaustively conformational spaces of paired bases defined by these small numbers of degrees of freedom, we obtain by energy refinement the complete structures of these molecules and their base pairing at the atomic scale.

Key words : DNA,RNA, NMR, hairpins structures, molecular modeling, BCE, theory of elasticity, base pairing, hydrogen bonds, stacking, conformational exploration.