



**HAL**  
open science

# Prédiction et modélisation in silico des oligonucléotides simple brin

Thomas Binet

► **To cite this version:**

Thomas Binet. Prédiction et modélisation in silico des oligonucléotides simple brin. Bio-informatique [q-bio.QM]. Université de Technologie de Compiègne, 2023. Français. NNT : 2023COMP2781 . tel-04664245

**HAL Id: tel-04664245**

**<https://theses.hal.science/tel-04664245v1>**

Submitted on 30 Jul 2024

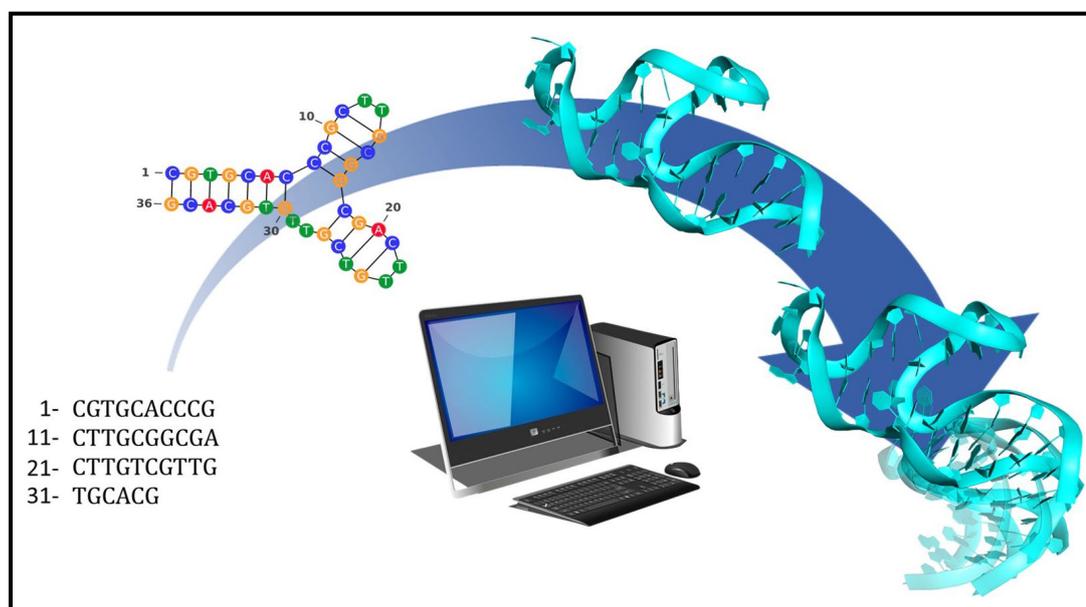
**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Par **Thomas BINET**

*Prédiction et modélisation in silico  
des oligonucléotides simple brin*

Thèse présentée  
pour l'obtention du grade  
de Docteur de l'UTC



Soutenu le 18 décembre 2023

**Spécialité :** Ingénierie Moléculaire et Interaction : Unité de  
recherche Génie Enzymatique et Cellulaire - GEC (UMR-7025)

D2781



## THESE DE DOCTORAT

Présentée pour l'obtention du grade de

DOCTEUR DE L'UNIVERSITE DE TECHNOLOGIE DE COMPIEGNE

*Génie enzymatique et cellulaire (GEC) – UMR CNRS 7025*

Spécialité : Ingénierie Moléculaire et Interaction

# **Prédiction et modélisation *in silico* des oligonucléotides simple brin**

Par **Thomas BINET**

Soutenue le 18 Décembre 2023

Composition du jury :

Pr. Bernard Offmann	Nantes Université	Rapporteur
Dr. Yann Ponty	Ecole Polytechnique de Paris	Rapporteur
Pr. Catherine Etchebest	Univ. Paris Cité	Examinatrice
Pr. Séverine Padiolleau-Lefèvre	Univ. De Technologie de Compiègne	Examinatrice
Pr. Bérangère Bihan-Avalle	Univ. De Technologie de Compiègne	Co-directrice
Dr. Irene Maffucci	Univ. De Technologie de Compiègne	Co-directrice



## Remerciements

Au terme de ma thèse, je me dois d'adresser un grand merci à mes deux directrices : Irene et Bérangère.

Irene, tu as été une vraie guide dans ce projet, une directrice sur laquelle on peut compter, jamais à court d'idées et de conseils pour me faire avancer là où seul j'aurai pu perdre mes repères. Une personnalité forte qui m'a motivé dans la réussite comme dans les difficultés capables de révéler le meilleur de moi-même. Tu m'as enseigné la rigueur scientifique, nos échanges étaient toujours constructifs et ta patience m'a toujours poussé à ne jamais te décevoir. Toujours présente, même lorsque les circonstances nous ont amené à échanger à distance. Il n'existe pas 1000 mots pour dire merci, mais à défaut il sera sincère et dans ta langue natale « Grazie mille ! ».

Bérangère, je te remercie car tu as toujours été d'un très grand secours également, prête à répondre ou s'interroger avec moi sur tout type de question. Chaque discussion était enrichissante et grâce à cela j'ai énormément appris auprès de toi.

Pour ce projet j'ai également eu l'occasion de travailler avec Miraine avec qui j'ai eu le plaisir de partager quelques discussions enrichissantes et de bon conseils mathématiques en plus de m'avoir aidé à soumettre mon premier article et pour cela je la remercie également.

J'adresse également mes remerciements à la Professeure Catherine Etchebest, le Professeur Bernard Offmann, le docteur Yann Ponty et la Professeure Séverine Padiolleau-Lefèvre pour avoir accepté de prendre part à mon jury de thèse. Je me réjouis de pouvoir discuter de mon travail avec vous.

Le GEC fut aussi un lieu de rencontre, avec de nombreuses personnes à remercier. Je commencerai donc par celle et celui qui m'ont accueilli dans leur bureau, le temps d'une année : Nicolas et Melissa. Les petites soirées pizza, instants cafés, et les picnics dans le parc resteront les petites joies de ma première année. J'y ajoute Mickael, camarade d'infortune pour ces deux dernières années, notre râleur Breton, avec qui j'ai pu bien rire sans filtre. Also, Meihua with whom I shared my first international congress in "the city of Church everywhere"

aka Oxford. Un joyeux quatuor avec qui j'ai partagé mon travail, mes rires et mes ressentis. Merci à vous.

Je remercie également les autres membres de la team Quiche, Séverine, Adeline, Stéphane et Geneviève pour leur suggestions, encouragements et leur soutien pendant ma thèse. Je garderai ce petit bout d'équipe soudé, riche de conseils et de sourires dans ma mémoire pour le reste de ma vie.

Je pense aussi à mes camarades doctorants qui ont été là également pour partager la vie de labo et hors labo (à défaut de comprendre tout ce que je faisais) et je les remercie. Un petit mot pour Tiffany la marseillaise qui me suivra bientôt avec qui j'aurai fini jusqu'au bout de la thèse. J'ajoute Camille, Christos, Noel, Alessia, Lucas, Thomas, Cyrian, Elise, Salim Rémi le petit stagiaire bioinfo, Annick, Vivi, Carmen, Clémence, Océane et tous les autres qui étaient de passage. J'ajoute un petit mot pour Selma qui reprend le flambeau pour ce projet, à qui je souhaite le meilleur et qui s'épanouira autant que j'ai pu m'épanouir durant cette thèse.

Un grand merci aussi à ma famille : Brigitte, François et Vincent. Le noyau dur de ma vie, mon tremplin partout où je vais, où que je sois, où que j'aille, ils m'aiment, me soutiennent et m'encouragent. C'est ce qu'ils ont fait pendant ces 3 ans. Je ne serais pas là sans eux. C'est tout naturellement que je leur adresse ce petit mot et j'espère vous avoir rendu fiers.

Le dernier mot sera pour mes amis de longues dates, qu'ils aient répondu présents ou non, ils ont été mon moyen d'évasion et sans le savoir ont été d'un grand support ces trois dernières années : Romain, Clément, Rémi, Tatiana, Lucile, Perrine et Thomas, Loïc, merci.

# Sommaire

<b>Introduction Générale .....</b>	<b>1</b>
<b>Partie 1. Les oligonucléotides simple brin .....</b>	<b>3</b>
<b>1. Composition et caractéristiques des oligonucléotides simple brin.....</b>	<b>3</b>
<b>2. Modifications de la composition des oligomères.....</b>	<b>7</b>
<b>3. Rôle et application des oligonucléotides .....</b>	<b>11</b>
3.1. Sondes.....	12
3.2. Régulation de l'expression des gènes.....	13
3.3. Aptamères .....	16
<b>4. Structure des oligonucléotides.....</b>	<b>17</b>
4.1. Structure primaire .....	17
4.2. Structure secondaire.....	17
4.2.1 Appariements et repliements .....	18
4.2.2 Rôle et implication de la structure secondaire .....	20
4.3. Structure tertiaire des oligonucléotides .....	22
4.4. Obtention de la structure des oligonucléotides.....	24
4.4.1 Cristallographie par rayon X.....	25
4.4.2 Spectroscopie à Résonance Magnétique Nucléaire .....	26
4.4.3 Cryo-microscopie électronique.....	26
4.4.4 Techniques pour la détermination de la structure secondaire .....	27
4.5. Ressources disponibles et variété de structures dans les bases de données.....	28
<b>5. Approche <i>in silico</i>.....</b>	<b>31</b>
5.1. Prédiction de structure secondaire et tertiaire.....	31
5.2. Dynamique moléculaire appliquée aux oligonucléotides.....	32
5.3. Complémentarité des outils.....	33
<b>Partie 2. Prédiction de structures secondaires des oligonucléotides.....</b>	<b>35</b>
<b>1. Sélection des données structurales des oligonucléotides .....</b>	<b>35</b>
1.1.1 Extraction et tri des données .....	35
1.1.2 Distribution, caractéristiques et statistiques .....	36
<b>2. Annotation informatique de la structure secondaire .....</b>	<b>38</b>
<b>3. Méthodes <i>in silico</i> pour la structure secondaire des oligonucléotides .....</b>	<b>40</b>

3.1.	Les approches basées sur la Minimisation d'Énergie Libre (MEL) .....	41
3.2.	Les approches <i>Knowledge based</i> ou basées sur l' <i>Intelligence Artificielle</i> .....	43
3.2.1	<i>Machine Learning</i> .....	43
3.2.2	Algorithmes hybrides.....	44
3.2.3	<i>Deep Learning</i> .....	44
<b>4.</b>	<b>Comparaison de structures secondaires .....</b>	<b>45</b>
4.1.	Comparaison de chaînes de caractères .....	46
4.2.	Classificateurs binaires.....	47
4.3.	Distances entre structures secondaires.....	47
4.4.	Distance d'édition des arbres.....	49
4.5.	Autres approches.....	50
4.6.	AptaMat : un outil de comparaison de structures secondaires efficace.....	51
4.6.1	Présentation de l'algorithme .....	53
4.6.2	Comparaison des d'AptaMat face aux autres distances .....	56
4.6.3	Calcul d'AptaMat appliqué aux ensembles de repliements .....	59
4.6.4	Performances d'AptaMat dans le regroupement d'oligonucléotides en familles Rfam.....	61
<b>5.</b>	<b>Performances des outils de prédiction .....</b>	<b>65</b>
5.1.	Génération et classification des structures.....	65
5.2.	Performances des approches de Minimisation d'Énergie Libre (MEL) .....	68
5.2.1	Comparaison sur l'ensemble des données .....	68
5.2.2	Effet du modèle thermodynamique sur la prédiction des structures d'ADN simple brin .....	70
5.2.3	Analyse des prédictions sous-optimales.....	71
5.3.	Comparaison globale des outils testés .....	73
5.3.1	Performances des approches Machine Learning et Deep Learning .....	73
5.3.2	Performances des approches mixtes .....	77
5.4.	Les difficultés de prédictions des motifs structuraux particuliers .....	78
5.4.1	Petits oligonucléotides et « Minidumbbell » .....	78
5.4.2	Pseudonœuds.....	79
5.4.3	Structures obtenues de complexes .....	80
5.5.	Prédictions de G-quadruplexes .....	83
<b>6.</b>	<b>Discussion .....</b>	<b>84</b>
<b>Partie 3. Échantillonnage conformationnel des oligonucléotides d'ADN à simple brin ...</b>		<b>89</b>
<b>1.</b>	<b>Méthodes de prédiction de structure tridimensionnelle .....</b>	<b>89</b>
<b>2.</b>	<b>Prédiction d'une structure tridimensionnelle .....</b>	<b>92</b>
2.1.	Sélection des données .....	92

2.2.	Outils et protocole de prédiction .....	92
2.3.	Optimisation des modèles tridimensionnels obtenus.....	94
2.4.	Mesure de la déviation Quadratique Moyenne des prédictions face à l'expérimental.....	95
<b>3.</b>	<b>Comparaison des prédictions .....</b>	<b>97</b>
3.1.	Prédictions structures secondaires de référence.....	97
3.2.	Prédictions à partir des structures secondaires obtenues par mfold .....	103
<b>4.</b>	<b>Exploration de l'espace conformationnel des oligonucléotides par dynamique moléculaire à échantillonnage intensif.....</b>	<b>111</b>
4.1.	Fondamentaux de la dynamique moléculaire.....	112
4.2.	Dynamiques moléculaires à échantillonnage intensif.....	116
4.3.	Protocoles de simulation .....	120
4.3.1	Préparation du système.....	120
4.3.1.1	Choix et préparation des structures des oligonucléotides .....	120
4.3.1.2	Choix du champ de force.....	124
4.3.1.3	Choix du modèle d'eau.....	125
4.3.1.4	Présence des ions.....	126
4.3.2	Protocole de dynamique moléculaire appliqué.....	127
4.3.2.1	Minimisation, chauffe et équilibration .....	127
4.3.2.2	Phase de production .....	128
4.3.2.3	Dynamiques moléculaires accélérées et dynamiques moléculaires accélérées gaussiennes 128	
4.3.2.4	Obtention des conformations majoritaire par clustering.....	131
4.4.	Etude conformationnelle des oligonucléotides par dynamique moléculaire.....	132
4.4.1	1NGO.....	133
4.4.2	1EZN.....	135
4.4.2.1	Simulations effectuées à partir de la structure expérimentale.....	135
4.4.2.2	Simulations effectuées à partir du modèle de RNAdenovo depuis la structure secondaire expérimentale.....	139
4.4.2.3	Simulations effectuées à partir du modèle de RNAdenovo depuis la structure secondaire prédite par mfold .....	142
4.4.3	3HXO .....	144
4.4.3.1	Simulations effectuées à partir de la structure expérimentale.....	145
4.4.3.2	Simulations effectuées à partir du modèle RNAdenovo obtenu depuis la structure secondaire expérimentale.....	146
4.4.3.3	Simulations effectuées à partir du modèle RNAdenovo obtenu depuis la structure secondaire prédite par mfold .....	151
4.4.4	3THW .....	152

4.4.4.1	Simulations effectuées à partir de la structure expérimentale.....	153
4.4.4.2	Simulations effectuées à partir du modèle RNAde novo généré depuis la structure secondaire expérimentale.....	155
4.4.5	SHTO.....	158
4.4.5.1	Simulations effectuées à partir de la structure expérimentale.....	159
4.4.5.2	Simulations effectuées à partir du modèle RNAde novo généré depuis la structure secondaire expérimentale.....	159
4.4.5.3	Simulations effectuées à partir du modèle RNAde novo généré depuis la structure secondaire prédite par mfold.....	163
4.5.	Discussion.....	165
<b>Conclusion Générale.....</b>		<b>169</b>
<b>Références .....</b>		<b>173</b>
<b>Annexes .....</b>		<b>187</b>

## Liste des illustrations

Figure 1-1. Structures des acides nucléiques .....	4
Figure 1-2. Clivage d'une portion d'ARN médié par la ribonucléase A.....	5
Figure 1-3. Appariements des acides nucléiques.....	7
Figure 1-4. Modifications de la chaîne nucléotidique.....	9
Figure 1-5. Applications cliniques possibles des oligonucléotides.....	12
Figure 1-6. Structures secondaires des acides nucléiques.....	18
Figure 1-7. Appariement de pseudonœuds type H, K, L et M.....	19
Figure 1-8. Formation d'un G-quadret et des appariement Hoogsteen.....	20
Figure 1-9. Impact d'une structure secondaire sur le processus de transcription de l'ADN en ARN messenger .....	21
Figure 1-10. Visualisation schématique des hélices d'ADN .....	23
Figure 1-11. Echelles de taille et compatibilité des techniques d'acquisition de structure ...	24
Figure 1-12. Evolution du nombre de structures disponibles publiés sur la PDB.....	29
Figure 1-13. Niveaux de complexité de repliement et de mobilité des oligonucléotides .....	33
Figure 2-1. Distribution des oligonucléotides du jeu de données par longueur de séquence	37
Figure 2-2. Méthodes de représentation de la structure secondaire.....	40
Figure 2-3. Représentation schématique des arbres associés à la structure secondaire.....	49
Figure 2-4. Comparaison d'une structure « épingle à cheveux » (a) et 3 structures alternatives avec la distance de Hamming, RNAdistance, BP, le score RBP, le score F1, le Coefficient de corrélation de Mathews (CCM), DoPloCompare et AptaMat.....	52
Figure 2-5. Exemple appliqué d'AptaMat sur deux structures hypothétiques.....	54
Figure 2-6. Comparaison d'une structure replié « hélice-bourgeon-hélice-boucle » et 4 structures alternatives avec la distance de Hamming, RNAdistance, BP, le score RBP, le score F1, le Coefficient de corrélation de Mathews (CCM) et AptaMat.....	58
Figure 2-7. Comparaison d'une structure type jonction à 3 voies et 2 structures alternatives avec la distance de Hamming, RNAdistance, BP, le score RBP, le score F1, le Coefficient de corrélation de Mathews (CCM), DoPloCompare et AptaMat .....	59
Figure 2-8. Clustering effectué sur 14 familles de structures de RFAM en utilisant la distance AptaMat et l'algorithme de propagation d'affinité .....	62

Figure 2-9. Représentation de la structure secondaire modèle des petites sous unité des ARN ribosomaux des bactéries ou des archées .....	63
Figure 2-10. Pseudocode du calcul du coefficient de Tanimoto modifié.....	67
Figure 2-11. Proportion de prédiction exactes suggéré par mfold et RNAfold par comparaison à la référence expérimentale. Les proportions sont mesurées sur les données ADN, ARN et ADN+ARN .....	69
Figure 2-12. Prédiction des structures secondaires de 6IY5, 6FKE, 1ECU, 2L5K et 2VIC en utilisant les modèles ADN (SantaLucia) et ARN (Mathews) de mfold .....	71
Figure 2-13. Pourcentage de structures prédites par différentes méthodes et la proportion de prédictions exactes, prédictions similaires, et mauvaises prédictions mesurés sur les données ADN, ARN et ADN+ARN .....	74
Figure 2-14. Qualité des prédictions par structure ADN et par outil.....	78
Figure 2-15. Représentation de la structure d'un « minidumbbell » .....	78
Figure 2-16. Représentation linéaire de la structure secondaire de 5HTO expérimentale et prédite par mfold ADN, RNAfold ADN, MXfold2 et Ufold.....	79
Figure 2-17. Exemples de 4 oligonucléotides en complexes avec la structure de référence et les différences de prédictions entre Mxfold2 et MC-fold.....	81
Figure 2-18. Représentation dot-bracket des structures secondaires expérimentales de 2XXA, 3WC1 et 5TF6 ainsi que les prédictions obtenues avec Ufold, RNAfold et mfold.....	82
Figure 3-1. Comparaison de l'épingle à cheveux G19-C26 de 1EZN ARN généré par RNAdenovo et ADN après correction avec PyMOL et Amber.....	94
Figure 3-2. Sélection des atomes du squelette phosphate d'une chaîne nucléotidique.....	96
Figure 3-3. Diagrammes en barre de la valeur de RMSD pour les prédictions obtenues avec RNAComposer, SimRNA et RNAdenovo à partir de la structure secondaire expérimentale .....	98
Figure 3-4. Portion structurée du squelette phosphate de 9 structures d'énergie minimale obtenues avec RNAdenovo à partir de la structure secondaire expérimentale alignées avec la structure de référence .....	100
Figure 3-5. Prédiction à partir de la structure secondaire expérimentale de 5HTO et 5HRU obtenues avec RNAdenovo, SimRNA et RNAComposer alignées avec la structure de référence .....	101

Figure 3-6. Prédiction à partir de la structure secondaire de mfold de 5HTO et 5HRU obtenues avec RNAdenovo, SimRNA et RNAComposer alignées avec la structure de référence .....	104
Figure 3-7 Diagrammes en barre de la valeur de RMSD pour les prédictions obtenues avec RNAComposer, SimRNA et RNAdenovo à partir de la structure secondaire de mfold .	106
Figure 3-8. Prédiction à partir de la structure secondaire expérimentale de 2CDM obtenues avec RNAdenovo, SimRNA et RNAComposer alignées avec la structure de référence .	107
Figure 3-9. Prédiction à partir de la structure secondaire de mfold de 5F55 obtenues avec RNAComposer, SimRNA et RNAdenovo alignées avec la structure de référence .....	109
Figure 3-10. Etapes du protocole de dynamique moléculaire .....	112
Figure 3-11. Type d'interactions entre atomes illustrant les termes du calcul de l'énergie potentielle .....	114
Figure 3-12. Profil énergétique d'un système simulé en dynamique moléculaire classique et l'effet de lissage induit par le facteur $\alpha$ appliqué en aMD .....	118
Figure 3-13. Etapes de préparation de la dynamique accélérée Gaussienne.....	119
Figure 3-14. Représentation tridimensionnelle sur PyMOL des structures 1NGO, 1EZN, 3HXO, 5HTO et 3THW obtenues soit de la PDB, soit prédit par RNAdenovo en utilisant la structure secondaire expérimentale ou suggérée par mfold .....	121
Figure 3-15. Représentation sur PyMOL de 3THW .....	123
Figure 3-16. Courbe de RMSD des simulations en dynamique moléculaire classique de 500ns effectuée sur les structures expérimentales de 3THW et 3HXO .....	129
Figure 3-17. Alignement des clusters majoritaires des différentes simulations aMD et GaMD effectuées en partant de la structure résolue expérimentalement de 1NGO.....	133
Figure 3-18. RMSD des simulations de GaMD sur 1NGO modélisé avec RNAdenovo basé sur la structure secondaire expérimentale en TIP3P .....	134
Figure 3-19. Alignement des clusters des différentes simulations aMD effectuées en partant de la structure résolue expérimentalement de 1EZN.....	136
Figure 3-20. Courbe de RMSD et distribution des clusters au cours des GaMD simulés dans un système concentré à 0,1M en NaCl avec différents modèles d'eau appliqués sur la structure expérimentale de 1EZN .....	137

Figure 3-21. RMSD calculée entre les différentes conformations de 1EZN expérimentale et le cluster 1.2 de la GaMD n°1 dans un système incluant une concentration de 0,1 M en NaCl et modèle d'eau TIP3P .....	138
Figure 3-22. RMSD calculée entre les différentes conformations de 1EZN expérimentale et le cluster 2.2 du modèle RNAdenovo de 1EZN du modèle RNAdenovo de 1EZN obtenue à partir de la structure secondaire expérimentale simulé en aMD dans un système neutralisé sans NaCl et modèle d'eau TIP4P-Ew .....	139
Figure 3-23. RMSD calculée entre les différentes conformations de 1EZN expérimentale et le cluster 1.3 du modèle RNAdenovo de 1EZN obtenue à partir de la structure secondaire expérimentale simulé en aMD dans un système intégrant une concentration de 0,1 M de NaCl et modèle d'eau TIP3P .....	140
Figure 3-24. RMSD calculée entre les différentes conformations de 1EZN expérimentale et le cluster 2.2 du modèle RNAdenovo de 1EZN obtenue à partir de la structure secondaire expérimentale simulé en GaMD dans un système incluant une concentration de 0,1 M de NaCl et le modèle d'eau TIP3P .....	141
Figure 3-25. Mesure de la RMSD entre clusters suggéré par TTclust provenant des simulations de GaMD en système incluant une concentration de 0,1 M de NaCl pour les simulations effectuées sur 1EZN expérimentale et 1EZN modélisé avec RNAdenovo en utilisant la structure secondaire expérimentale .....	141
Figure 3-26. Distribution des clusters au cours des GaMD simulés avec le modèle d'eau TIP3P sur le modèle RNAdenovo de 1EZN utilisant la structure secondaire de mfold.....	143
Figure 3-27. Alignement des clusters des différentes simulations aMD effectuées en partant de la structure résolue expérimentalement de 3HXO .....	145
Figure 3-28. Distribution des clusters au cours des GaMD simulées avec différents modèles d'eau appliqués sur la structure expérimentale de 3HXO .....	147
Figure 3-29. Alignement des clusters des différentes simulations aMD effectuées en partant du modèle RNAdenovo de 3HXO obtenu à partir de la structure secondaire expérimentale. ....	148
Figure 3-30. Distribution des clusters au cours des GaMD simulés avec le modèle d'eau TIP3P sur le modèle RNAdenovo de 3HXO utilisant la structure secondaire expérimentale..	149

Figure 3-31. Mesure de la RMSD entre clusters suggéré par TTclust de 3HXO expérimentale et 3HXO modélisé avec RNAdenovo en utilisant la structure secondaire expérimentale simulés en GaMD incluant une concentration de 0,1 M de NaCl avec le modèle d'eau TIP3P .....	150
Figure 3-32. Alignements de la structure de 3HXO résolue expérimentalement avec les clusters X.1 obtenues avec TTclust provenant des GaMD sur modèle RNAdenovo de 3HXO utilisant la structure secondaire de mfold .....	152
Figure 3-33. RMSD de 3HXO modélisé avec RNAdenovo utilisant la structure secondaire de mfold face à la référence expérimentale mesurée pour les modèles d'eau TIP3P et TIP4P-Ew en GaMD n°1 et GaMD n°2 .....	152
Figure 3-34. Alignement des clusters des différentes simulations aMD effectuées en partant de la structure résolue expérimentalement de 3THW en système neutralisé avec et sans concentration de 0,1 M de NaCl avec les modèles d'eau TIP3P ou TIP4P-Ew.....	153
Figure 3-35. Alignement des clusters des différentes simulations GaMD effectuées en partant de la structure résolue expérimentalement ou le modèle RNAdenovo obtenu à partir de la structure secondaire expérimentale de 3THW. Toutes ces simulations ont été effectuée en système avec concentration de 0,1 M de NaCl testé avec les modèles d'eau TIP3P ou TIP4P-Ew.....	154
Figure 3-36. Mesure de la RMSD entre clusters suggéré par TTclust de 3THW modélisé avec RNAdenovo en utilisant la structure secondaire expérimentale simulé en GaMD incluant une concentration de 0,1 M de NaCl avec le modèle d'eau TIP3P .....	156
Figure 3-37. Courbe de RMSD et distribution des clusters et des conformations associées au cours des GaMD simulés dans un système concentré à 0,1 M en NaCl avec différents modèles d'eau (TIP3P ou TIP4P-Ew) appliqués sur le modèle RNAdenovo de 3THW en utilisant la structure secondaire expérimentale .....	157
Figure 3-38. Mesure de la RMSD entre clusters suggéré par TTclust de 3THW expérimentale et 3THW modélisé avec RNAdenovo en utilisant la structure secondaire expérimentale simulés en GaMD incluant une concentration de 0,1 M de NaCl en TIP3P.....	158
Figure 3-39. Alignement des clusters des différentes simulations aMD effectuées en partant du modèle RNAdenovo de 5HTO obtenu à partir de la structure secondaire expérimentale en système neutralisé avec et sans concentration de 0,1 M de NaCl testé avec les modèles d'eau TIP3P ou TIP4P-Ew .....	160

Figure 3-40. Distribution des clusters au cours de la GaMD n°1 simulés dans un système concentré à 0,1 M en NaCl avec le modèle d'eau TIP3P sur le modèle RNAdenovo de 5HTO utilisant la structure secondaire expérimentale. Chaque couleur représente un cluster différent.....	161
Figure 3-41. Distribution des clusters au cours de la GaMD n°1 simulés dans un système concentré à 0,1 M en NaCl avec le modèle d'eau TIP4P-Ew sur le modèle RNAdenovo de 5HTO utilisant la structure secondaire expérimentale. Chaque couleur représente un cluster différent.....	162
Figure 3-42. Distribution des clusters associées au cours des GaMD simulés dans un système concentré à 0,1 M en NaCl avec différents modèles d'eaux (TIP3P ou TIP4P-Ew) appliqués sur le modèle RNAdenovo de 5HTO utilisant la structure secondaire de mfold .....	164
Figure C-0-1. Protocole complet de prédiction de structure des oligonucléotides.....	170

## Liste des tableaux

Tableau 1. Présentation des oligonucléotides commercialisés, avec leur nom commercial, le type d'action, la compagnie qui a produit le composé, la cible et la pathologie traitée.	15
Tableau 2. Distribution des données structurales retenues pour le jeu de données d'oligonucléotides obtenues à partir de la PDB.....	36
Tableau 3. Présentation des outils de prédiction de structures secondaires utilisés, avec le principe de fonctionnement, les avantages et les inconvénients pour chaque méthode. ....	41
Tableau 4. Méthodes de calculs proposés par RNAdistance, utilisable via le paramètre -D et la lettre associée dans Méthode. ....	50
Tableau 5. Comparaison de l'ensemble conformationnel de l'ARN TAR à l'état libre face à la structure secondaire adoptée en forme lié au peptide ligand .....	60
Tableau 6. Tableau résumé des différences majeures entre les 3 approches Deep Learning étudiées.....	75
Tableau 7. Paramètres appliqués aux outils de prédiction RNAComposer, RNAdenovo et SimRNA.....	94
Tableau 8. Différences de paramètres entre les modèles d'eau en 3 points TIP3P et en 4 points TIP4P et TIP4P-Ew .....	125
Tableau 9. Résumé de l'ensemble des simulations effectuées sur les différents oligonucléotides soit 1EZN, 1NGO, 3HXO, 3THW ou 5HTO .....	131



## Liste des équations

Équation 1: Tanimoto.....	46
Équation 2: Distance de Hamming.....	46
Équation 3: Distance Base Pair.....	47
Équation 4: Distance RBP.....	47
Équation 5: Score F1.....	47
Équation 6: Coefficient de Corrélacion de Matthews.....	48
Équation 7: DoPloCompare.....	50
Équation 8: RMS distance (Distance Quadratique Moyenne).....	50
Équation 9: Normalisation de DoPloCompare.....	51
Équation 10 : Distance de Manhattan.....	53
Équation 11 : $Apta_D$ .....	55
Équation 12: $Apta_D$ sur ensemble.....	56
Équation 13 : Matrice d'affinité de l'ensemble de structure Rfam comparé avec AptaMat...	61
Équation 14 : RMSD (Deviation Quadratique Moyenne).....	95
Équation 15 : Equation du mouvement.....	112
Équation 16 : Equation du mouvement selon la dérivée de l'accélération.....	112
Équation 17 : Gradient négatif de potentiel.....	113
Équation 18 : Gradient négatif de potentiel selon la dérivée de l'accélération.....	113
Équation 19 : Calcul de l'énergie potentielle $V$ .....	113
Équation 20 : Modification de l'énergie potentielle.....	116
Équation 21 : Modification de l'énergie potentielle appliquée en aMD.....	116
Équation 22 : Biais $\alpha P$ .....	117
Équation 23 : Biais $\alpha D$ .....	117
Équation 24 : Seuil d'énergie $E_p$ .....	117
Équation 25 : Seuil d'énergie $E_d$ .....	117
Équation 26 : Modification de l'énergie potentielle appliquée en GaMD.....	118
Équation 27 : Intervalle du seuil d'énergie $E$ .....	118
Équation 28 : Calcul de $k_0$ en fonction des limites inférieures ou supérieures de $E$ .....	118
Équation 29 : Equation des termes électrostatique de TIP4P-Ew.....	125
Équation 30 : Calcul du nombre d'atome en fonction de la concentration.....	127



## Liste des abréviations

ADN : acide désoxyribonucléique

ARN : acide ribonucléique

ARNi : ARN interférent

miARN : micro ARN interférents

ARNm:ARN messager

ARNt: ARN de transfert

G : guanine

C : cytosine

A : adénine

T : thymine

U : uracile

WC : Watson-Crick

MD : dynamique moléculaire

aMD : dynamique moléculaire accélérée

GaMD : dynamique moléculaire accélérée

Gaussienne

PCR : Polymerase Chain Reaction

RMN : résonance magnétique nucléaire

cryo-ME : cryo-microscopie électronique

cryo-TE : cryo-tomographie électronique

DMS : diméthylsulfate

PDB : Protein Data Bank

VEGF : facteur de croissance de l'endothélium vasculaire (*Vascular endothelial growth factor*)

RBS : site de fixation du ribosome (*Ribosome binding site*)

CCM : coefficient de corrélation de Mathews

MEL : minimisation de l'énergie libre

MutS $\beta$  : Mutator S $\beta$



# Introduction Générale

Les oligonucléotides simple brin (que nous nommerons ici 'oligonucléotides' pour simplifier) sont des fragments d'acides nucléiques simple brin (acide désoxyribonucléique, ADN, ou acide ribonucléique, ARN) capables de régir certaines fonctions cellulaires, comme la régulation de l'expression des gènes (R. C. Lee et al., 1993), et pour lesquels ont été identifiés des champs d'application dans le domaine de la thérapie et du diagnostic. En effet, ces molécules sont capables non seulement de lier des séquences oligonucléotidiques spécifiques, mais elles disposent aussi de la capacité de reconnaître tout type de molécule ou structure (petite molécule, protéine, cellule, etc.), ce qui les rend exploitables dans les domaines thérapeutique ou diagnostique.

Les oligonucléotides sont caractérisés par la diversité des conformations qu'ils peuvent adopter : les nucléotides complémentaires de la chaîne peuvent former des appariements responsables de l'apparition de repliements caractéristiques. On identifie trois niveaux d'organisation structurale : la structure primaire, correspondant à la séquence de nucléotides, la structure secondaire, caractérisée par le simple appariement des nucléotides complémentaires, et la structure tertiaire déterminée par les appariements à longue distance responsables de repliements plus complexes en trois dimensions. L'intérêt porté aux acides nucléiques monocaténaire, notamment les ARN, a permis d'acquérir les connaissances nécessaires à la compréhension de ce type de structure, le mécanisme d'action ainsi que le mode d'interaction avec les biomolécules. Cependant, l'enrichissement des connaissances sur la structure des oligonucléotides reste limité à cause des techniques d'acquisition relativement lentes et coûteuses, impactant directement le développement d'oligonucléotides à visées thérapeutique ou diagnostique. L'évolution des approches *in silico* appliquées à la modélisation des acides nucléiques permet désormais de les envisager comme support aux techniques plus conventionnelles utilisées dans la recherche d'oligonucléotides d'intérêt.

Les travaux de cette présente thèse ont permis d'étudier les méthodes computationnelles associées à la modélisation moléculaire d'oligonucléotides et ils s'insèrent dans un projet visant à concevoir *de novo* des oligonucléotides capables de reconnaître des molécules d'intérêt biotechnologique. Les méthodes disponibles permettent de prédire les deux niveaux

de structures des oligonucléotides : la structure secondaire et la structure tertiaire. Différents outils de prédiction de structures secondaires ont été évalués, permettant de déterminer leurs avantages et inconvénients pour en désigner le plus performant selon les critères. Un bon outil se définit par sa capacité à s'approcher de la structure résolue expérimentalement, et nécessite l'utilisation de métriques de comparaison pour estimer les différences entre prédiction et référence. Pour cela, un jeu de données contenant des oligonucléotides libres ou en complexe avec une protéine dont la structure tridimensionnelle expérimentale est connue a été créé et une nouvelle approche de comparaison des structures secondaires, AptaMat, a été développée. Ceux-ci ont été utilisés pour la comparaison de 9 outils de prédiction de la structure secondaire des oligonucléotides, pour déterminer l'outil le plus fiable pour l'obtention de l'appariement correct des nucléotides, capable de guider la prédiction de la structure tridimensionnelle des oligonucléotides. Pour cette dernière, les approches disponibles ont également été évaluées en comparant les performances de 3 outils (RNAde novo, SimRNA et RNAComposer). Le manque d'exploration de la flexibilité structurale des oligonucléotides a ensuite été complété par l'utilisation des approches de dynamiques moléculaires à échantillonnage intensif (dynamique moléculaire accélérée ou accélérée-gaussienne), qui ont permis de confirmer la stabilité des repliements, d'explorer d'autres conformations favorables ou retrouver la structure résolue.

# Partie 1. Les oligonucléotides simple brin

## 1. Composition et caractéristiques des oligonucléotides simple brin

Les oligonucléotides simple brin sont des molécules d'acides nucléiques, soit d'acide désoxyribonucléique (ADN) soit d'acide ribonucléique (ARN) simple brin, composées d'un nombre limité de nucléotides, en moyenne 25 mais pouvant aller jusqu'à 200 (Dias & Stein, 2002). Chaque nucléotide est constitué d'une molécule de sucre à 5 atomes de carbone (D-ribose pour l'ARN ou D-désoxyribose pour l'ADN), d'un groupement phosphate sur le carbone C5' du sucre, et d'une base azotée sur le carbone C1' du sucre (Figure 1-1a).

L'assemblage de plusieurs nucléotides pour former un oligonucléotide se fait grâce à des liaisons phosphodiester entre le phosphate d'un nucléotide et le groupement hydroxyle du carbone C3' du nucléotide précédent, ce qui confère à l'oligonucléotide une direction de progression de 5' à 3' (Figure 1-1d). Ces liaisons sont sujettes à l'activité des nucléases qui dégradent les acides nucléiques. L'enchaînement de groupements phosphate et de sucres constitue le squelette sucre-phosphate des oligonucléotides, qui joue un rôle direct dans leur mobilité structurale. Dans ce contexte, le désoxyribose des oligonucléotides d'ADN se différencie du ribose par l'absence de l'hydroxyle sur le carbone C2' (Figure 1-1b). La stabilité de la structure tridimensionnelle des oligonucléotides d'ARN s'en retrouve améliorée par rapport aux ADN grâce à l'hydratation de la molécule possible par des interactions électrostatiques du 2' hydroxyle avec les molécules d'eau (Gyi et al., 1998). La présence de ce groupement dans les oligonucléotides constitués d'ARN impacte également la mobilité des angles du cycle osidique, provoquant un effet de torsion, notamment au niveau du carbone C3' ce qui impacte également les angles dièdres des atomes composant le squelette sucre-phosphate. L'ensemble de cette mobilité permet de renforcer les liaisons hydrogène impliquées dans l'appariement des bases (Fohrer et al., 2006). En contrepartie ce groupement nucléophile supplémentaire est impliqué dans la réaction de substitution nucléophile de type 2 avec le phosphate de la liaison phosphodiester. Cette réaction est favorisée par le mécanisme d'action des nucléases, présenté en Figure 1-2, qui provoque la séparation en deux sous-chaînes de nucléotides (Hale et al., 1993). L'absence de groupement 2' hydroxyle dans les molécules d'ADN les rend naturellement plus résistantes aux nucléases, expliquant en partie leur temps de demi-vie plus grand par rapport à l'ARN, qui est plus sujet à la dégradation

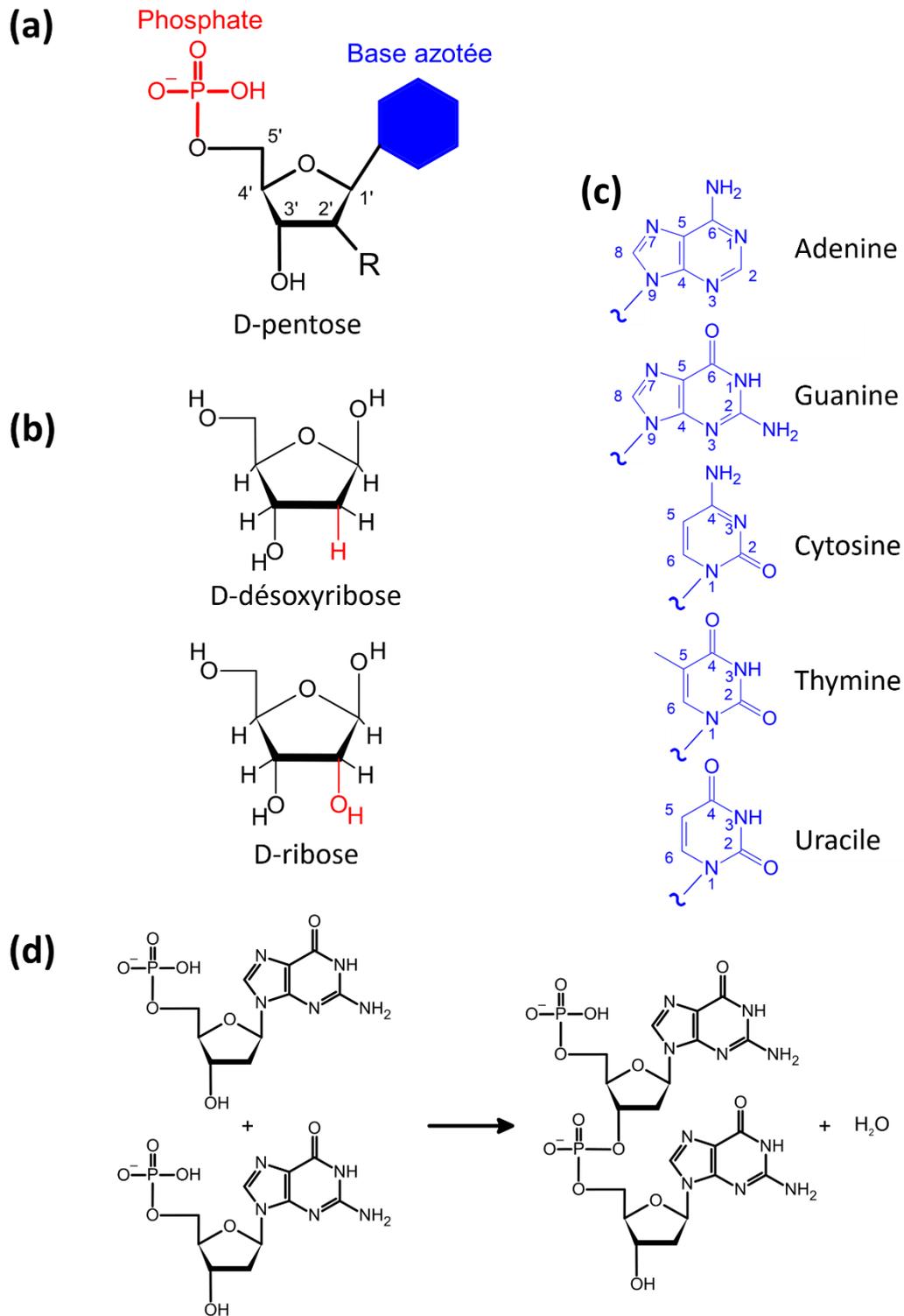
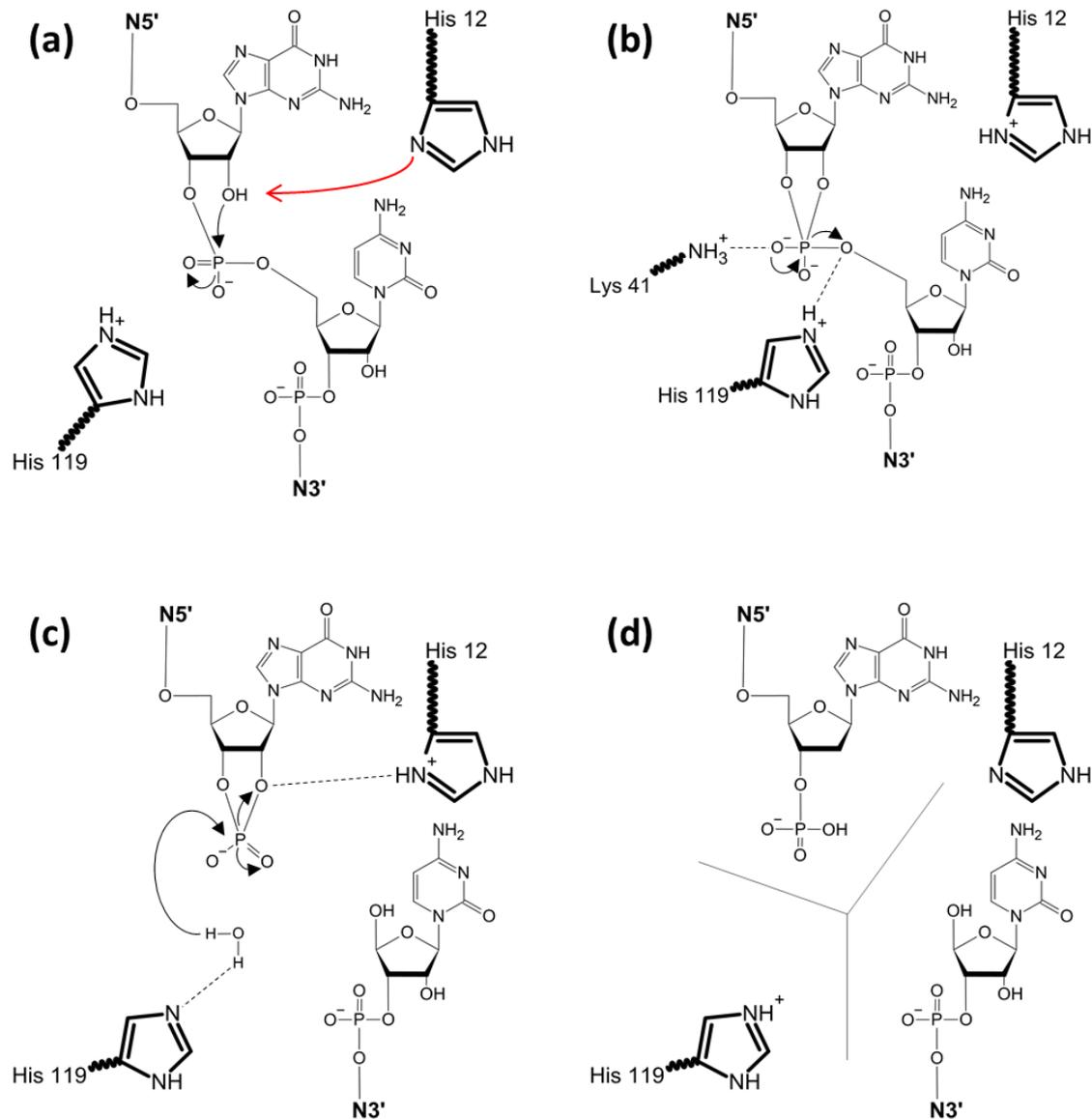


Figure 1-1. Représentation schématique des structures d'acides nucléiques avec l'unité nucléotides représentée en (a) avec le D-pentose lié à la base azotée en C1' et le groupement phosphate  $\text{H}_2\text{PO}_4^-$ . Les différences localisées sur la partie osidique sont montrées en (b). Les différentes bases azotées sont montrées en (c). La réaction de formation de la liaison phosphodiester entre 2 nucléotides est montrée en (d). Figure réalisée avec ChemSketch.

dans le sérum (quelques dizaines de minutes pour l'ADN versus quelques minutes pour l'ARN) (Houseley & Tollervey, 2009).

Cependant, le repliement tridimensionnel de l'ARN est impacté par la présence du 2' hydroxyle, avec une diversité de conformation amoindrie. De fait, l'encombrement stérique de ce groupement limite la flexibilité globale du polymère ARN ce qui stabilise les interactions responsables du repliement tridimensionnel (Nowakowski & Tinoco, 1997)



La diversité de séquences et de structures des oligonucléotides est directement associée aux bases azotées qui les composent. Dans la nature, les molécules d'ADN et ARN peuvent contenir 4 bases différentes : adénine, cytosine, guanine, et thymine pour l'ADN et uracile pour l'ARN (A, C, G, T/U). Cytosine, thymine et uracile sont les trois bases pyrimidiques, organisées autour d'un hétérocycle aromatique azoté (N1 et N3), avec un groupement carbonyle en C2 et, en C6, un groupement amine pour la cytosine ou carbonyle pour la thymine et l'uracile. Les bases puriques adénine et guanine se composent d'un hétérocycle aromatique azoté de type pyrimidique accolé à un cycle imidazole. On y retrouve, donc, les atomes d'azote N1 et N3, et en C6 : un groupement amine pour les adénines et un groupement carbonyle pour les guanines qui comprennent également un groupement amine en C2.

Les bases azotées ont la capacité de s'apparier en formant des liaisons hydrogène impliquant les groupements polaires. Les appariements Watson-Crick (WC) sont canoniques et impliquent des liaisons G-C, et A-T ou A-U respectivement dans les molécules d'ADN et d'ARN (Figure 1-3a, b et c). D'autres appariements sont possibles, comme les *Wobble pair* (appariement bancal) (Figure 1-3d) chez les ARN, impliquant des G-U, qui contribuent à stabiliser la structure de l'oligonucléotide. Le degré de stabilité des oligonucléotides est relié à son énergie libre ( $\Delta G$ ) : plus le  $\Delta G$  est négatif, plus la molécule sera stable. Chaque liaison hydrogène contribue à abaisser le  $\Delta G$  de l'oligonucléotide de -3,6/6 kcal/mol pour les interactions de type N – H  $\cdots$  O et de -4 kcal/mol pour les interactions de type N – H  $\cdots$  N (Figure 1-3) (Haynie, 2008). Les appariements A-T comportent une interaction N – H  $\cdots$  O et une N – H  $\cdots$  N, les appariements G-C sont formés par deux interactions N – H  $\cdots$  O et une interaction N – H  $\cdots$  N, et un appariement G-U consiste en deux interactions N – H  $\cdots$  O. Les appariements canoniques sont plus fréquemment rencontrés, car les liaisons hydrogènes impliqués apportent une plus grande cohésion et stabilisent le repliement des oligonucléotides bien que d'autres liaisons hydrogènes entre bases soit possible. La présence ou l'absence de ces interactions impacte la structure secondaire des oligonucléotides et apporte une grande diversité de repliements structuraux.

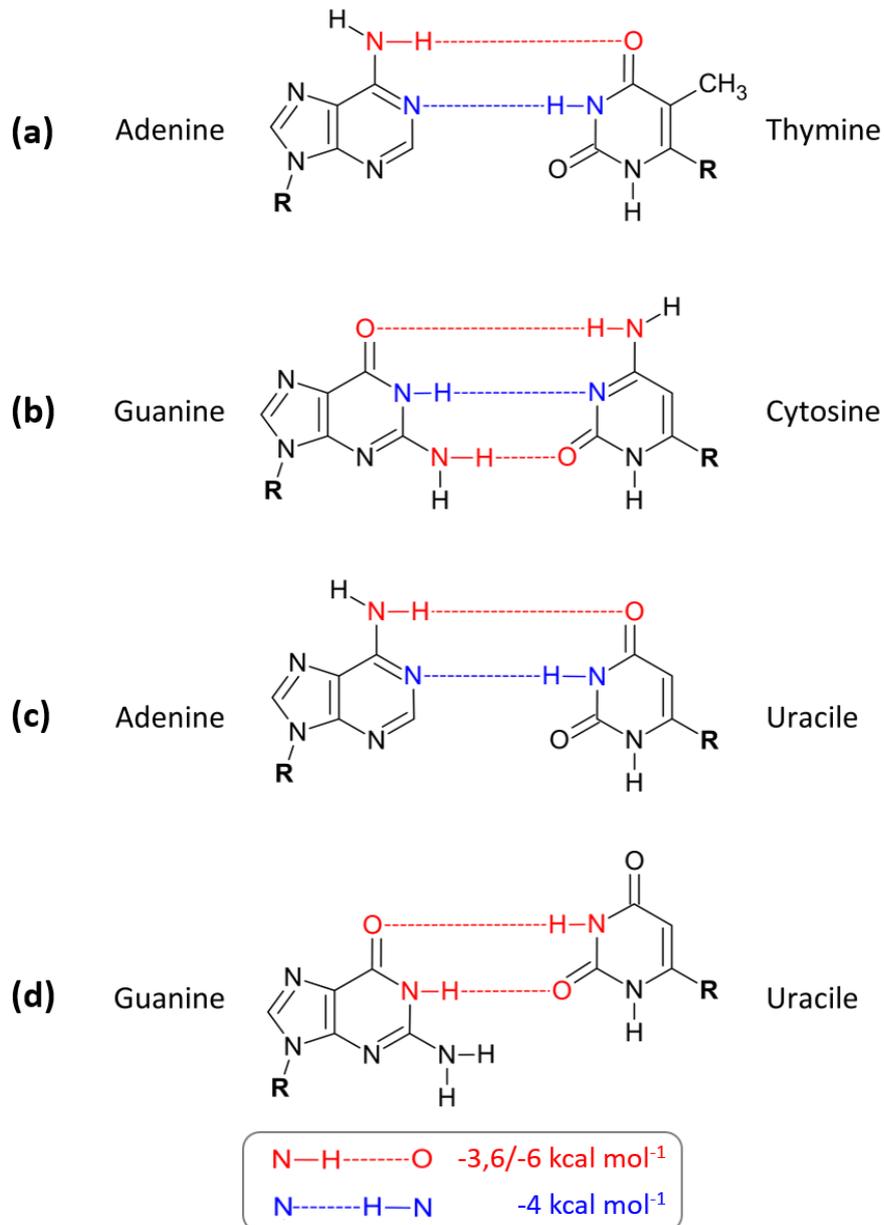


Figure 1-3. Représentation des appariements d'acides nucléiques. (a), (b) et (c) représentent les appariements canoniques type Watson Crick, et (d) un appariement particulier appelé *Wobble pair*. Les hydrogènes impliqués sont colorés en rouge ou bleu selon leur type et leur contribution énergétique est indiquée en bas de la figure.

## 2. Modifications de la composition des oligomères

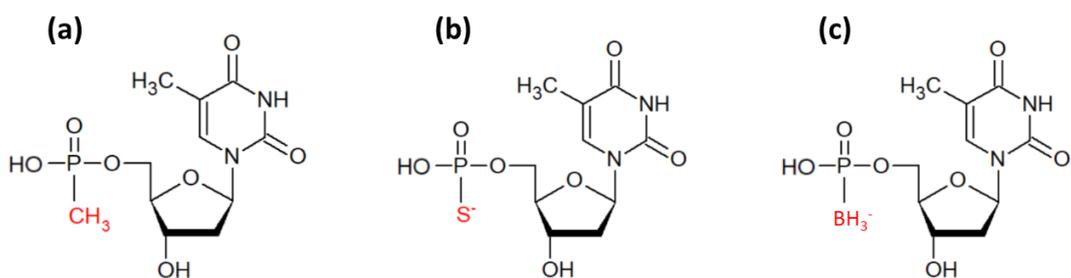
La composition des oligonucléotides simple brin influence leur comportement en solution, notamment leur flexibilité, leur biodisponibilité, leurs propriétés physico-chimiques ou leur temps de demi-vie. Par exemple, il a été précédemment décrit que les nucléases étaient capables de dégrader les oligonucléotides, avec un effet négatif sur leur temps de demi-vie. Il a été identifié dans la nature des oligonucléotides disposant de propriétés améliorées, cependant limitées à la modification de =O en =S ou de -OH en -SH localisée sur les bases

azotées ou sur le groupement phosphate (Zheng et al., 2021). Sur la base de ce type de modification il a été envisagé d'inclure des modifications sur les nucléotides, introduites chimiquement pour des applications thérapeutiques et diagnostiques. La plupart des modifications existantes vont dans le sens d'une protection vis à vis de l'action des endo/exonucléases.

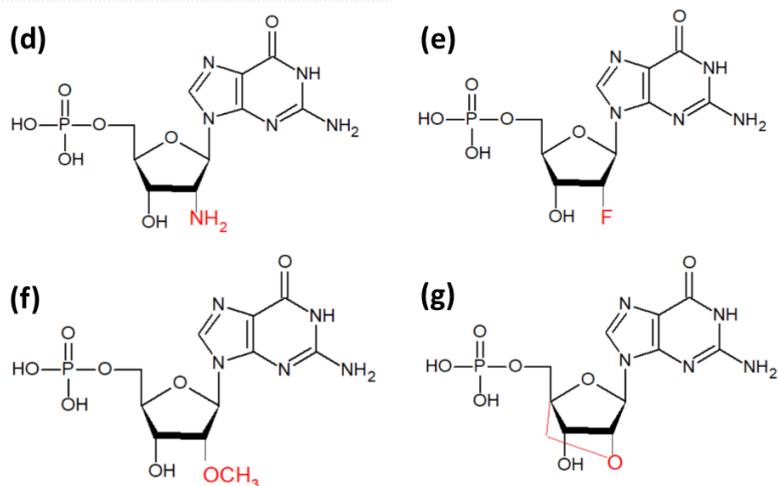
Tout d'abord, les extrémités 5' et/ou 3' peuvent présenter un groupement additionnel protecteur localisé au niveau du groupement phosphodiester. En ajoutant un monomère, de nature nucléotidique ou non, la chaîne est coiffée et protégée de l'action des exonucléases, améliorant significativement son temps de demi-vie de quelques minutes à plusieurs heures (Lacroix et al., 2017). Les types de modifications sont variés, mais il est possible d'utiliser certains conjugués selon l'extrémité. Notamment, à l'extrémité 3' il est commun d'ajouter une base inversée, en général une thymine inversée (Figure 1-4i) (Floege et al., 1999). Ainsi, l'oligomère ne dispose plus de son extrémité 3', ce qui le protège de l'action des exonucléases clivant par l'extrémité 3'. La biotinylation de l'extrémité 3' a également démontré une bonne capacité d'amélioration de la stabilité des oligonucléotides (Figure 1-4h) (Dougan et al., 2000), de 10 à 20 fois supérieure à celle de la molécule non modifiée en milieu complexe. Dans le cas des extrémités 5', il est plus fréquent de conjuguer des molécules de grande taille. Ceci offre l'avantage de diminuer la clairance rénale et améliorer leur biodisponibilité car la taille des conjugués impacte le poids moléculaire et donc le processus de réabsorption. Par exemple, l'ajout d'un groupement alkylamine permet de bloquer l'action des exonucléases 5' en agissant comme site nucléophile (Figure 1-4j). D'autres composés moléculaires jouent également ce rôle, comme des acides gras, des protéines, des polycations, des stérols (Figure 1-4k) ou des polyéthylène glycols (PEG) (Manoharan, 2002).

D'autres modifications chimiques directes des groupements phosphates intervenant dans la liaison phosphodiester peuvent également aider à améliorer la résistance vis-à-vis de l'action d'enzymes. La substitution du phosphate par un méthylphosphonate, un phosphorothioate ou boranephosphonate (Figure 1-4a, b et c) protège le polymère de l'action des endonucléases (Reynolds et al., 1996). En plus de l'augmentation du temps de demi-vie, en fonction du groupement introduit, la biodistribution peut être facilitée (Kaur et al., 2013).

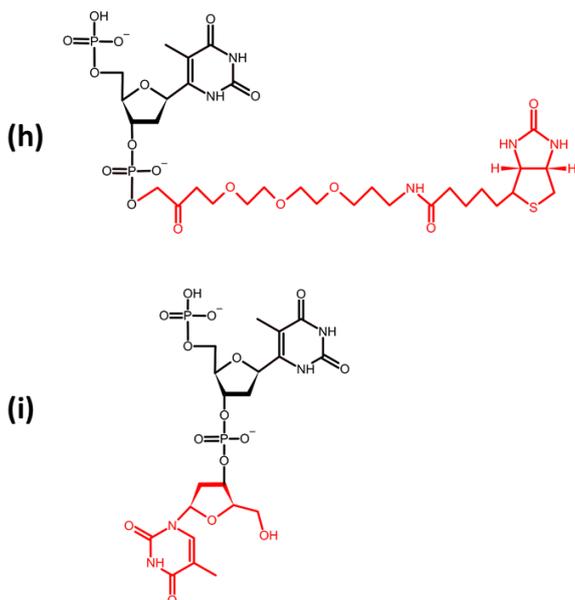
## Modifications sur phosphodiester



## Modifications sur ose



## Coiffes 3'



## Coiffes 5'

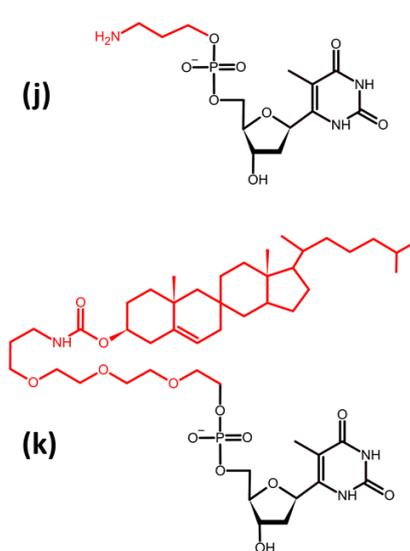


Figure 1-4. Exemples de modifications de la chaîne nucléotidique localisées sur les phosphodiesters (a, b, c), sur les oses (d, e, f, g) ou sur les extrémités 3' (h, i) et 5' (j, k).

La partie osidique peut également être modifiée enfin d'augmenter le temps de demi-vie des ribonucléotides dans le sérum (Healy et al., 2004). La substitution de l'hydroxyle en 2' par une amine, un fluor ou un O-méthyl (Figure 1-4d, e et f) a pour conséquence de diminuer la sensibilité aux nucléases (Khvorova & Watts, 2017). D'autres modifications du cycle D-pentose sont possibles. L'utilisation d'un pontage entre le groupement en 2' et le C4' (Figure 1-4g) permet d'augmenter la stabilité du cycle en échange d'une mobilité réduite, et favorise les interactions et l'appariement des bases. De plus, le pontage augmente la résistance aux nucléases grâce à la fois à la modification de la distribution des charges et à l'impossibilité d'initier l'attaque nucléophile (Elayadi et al., 2002). Enfin, il est également possible de modifier directement le cycle en substituant l'oxygène en 4' par un autre atome, par exemple un atome de soufre, ou encore de substituer le pentose en hexose (Hendrix et al., 1997; Hoshika et al., 2004).

Enfin, il est possible d'intégrer, à la séquence, des bases azotées dites non-naturelles. Ces variantes sont obtenues par mutagenèse et sont fréquemment utilisées pour i) améliorer l'affinité de liaison avec les cibles potentielles, ii) renforcer les appariements entre bases par l'ajout de sites donneur ou accepteur de liaisons hydrogène iii) empêcher des appariements indésirés. Les modifications apportées sur les nucléobases peuvent être de nature variée et de nombreux exemples sont présentés dans la revue de Lee and Berdis (2010). Certaines bases azotées non-naturelles se différencient fortement des bases naturelles, avec l'incorporation d'avantage de cycles aromatiques ou de groupements encombrants qui modifient la stéréochimie des bases azotées limitant ainsi les interactions entre deux bases (naturelles ou non). L'impossibilité de former des appariements augmente la disponibilité des bases azotées pour créer des interactions avec la cible, avec la possibilité d'apporter une meilleure complémentarité de surface, des contacts hydrophobes ou des empilements  $\pi - \pi$  (Davies et al., 2012; X. Zhang et al., 2005). Certaines modifications ne changent pas profondément la nature hétérocyclique des bases azotées et contribuent au renforcement des appariements par des schémas de liaisons hydrogène différents pouvant enrichir l'alphabet génétique conventionnel (Johnson et al., 2004). Plus simplement, les isomères des bases azotées naturelles apparaissent comme une alternative possible aux appariements WC. Les liaisons hydrogène impliquées et les paramètres de liaisons ne diffèrent pas par rapport à celles créées

par les bases naturelles, mais peuvent affecter la configuration tridimensionnelle (Hoshika et al., 2004; Yang et al., 1998).

En définitive, la modification des nucléobases apporte une grande diversité de propriétés aux oligonucléotides bien que l'objectif soit de conserver les caractéristiques pour lesquelles la séquence a été sélectionnée. Toutes ces modifications de la chaîne nucléotidique peuvent être associées pour aboutir à la synthèse de nouveaux oligomères aux propriétés favorables pour les applications envisagées. (Floege et al., 1999)

### 3. Rôle et application des oligonucléotides

Les oligonucléotides possèdent la capacité de s'hybrider à une séquence nucléotidique complémentaire de par leur nature mais sont également capables de lier spécifiquement un grand nombre de molécules. Ces facultés leur permettent, en fonction de la cible, d'exercer une activité catalytique, d'inhiber ou stimuler l'expression de certains gènes et parfois de réguler la fonction de macromolécules ou de cellules (Roberts et al., 2020).

Les oligonucléotides, dans la cellule, sont avant tout impliqués dans les processus cellulaires de conservation et d'expression de l'information génétique et sont présents sous la forme d'ARN monocaténaire. Parmi les oligonucléotides les plus connus, on retrouve ceux de la famille des microARN, qui sont de courts fragment d'ARN actifs dans la régulation de l'expression post-transcriptionnelle des gènes (Bartel, 2009). D'autres oligonucléotides, très importants d'un point de vue biologique, sont les ribozymes, qui sont des ARN structurés ayant la capacité de catalyser des réactions chimiques, comme des réactions redox, des transméthylations, la formation de liaisons carbone-carbone et des réactions de phosphorylation (Alonso & Mondragón, 2021). En outre, les sous-unités ribosomales qui sont impliquées dans la traduction des ARN messager (ARNm), les ribonucléases qui catalysent le clivage des molécules d'ARN, ou le splicéosome qui catalyse le processus d'épissage dans le noyau sont aussi des oligonucléotides. Bien que localisés sur des portions non codantes de l'ARNm, les riboswitchs peuvent être considérés comme des oligonucléotides naturels. Ils ont une activité autolytique et réagissent à la présence d'un ligand activateur pour cliver certaines parties de l'ARN et bloquer ou empêcher son expression. Les thermorégulateurs (ou thermosenseurs) n'ont pas d'activité catalytique, mais leur expression est régulée par les changements de température. Il existe près de 4100 familles d'ARN non codant naturels qui

présentent des fonctions variées. Ces familles sont regroupées dans la *RNA families database* (Rfam) <sup>1</sup> (Kalvari et al., 2021).

La capacité de ces familles d'acides nucléiques simple brin de réguler les fonctions cellulaires a motivé leur développement pour des applications dans le domaine thérapeutique ou diagnostique, applications rendues possibles par la synthèse chimique des oligonucléotides (Khvorova & Watts, 2017) (Figure 1-5)

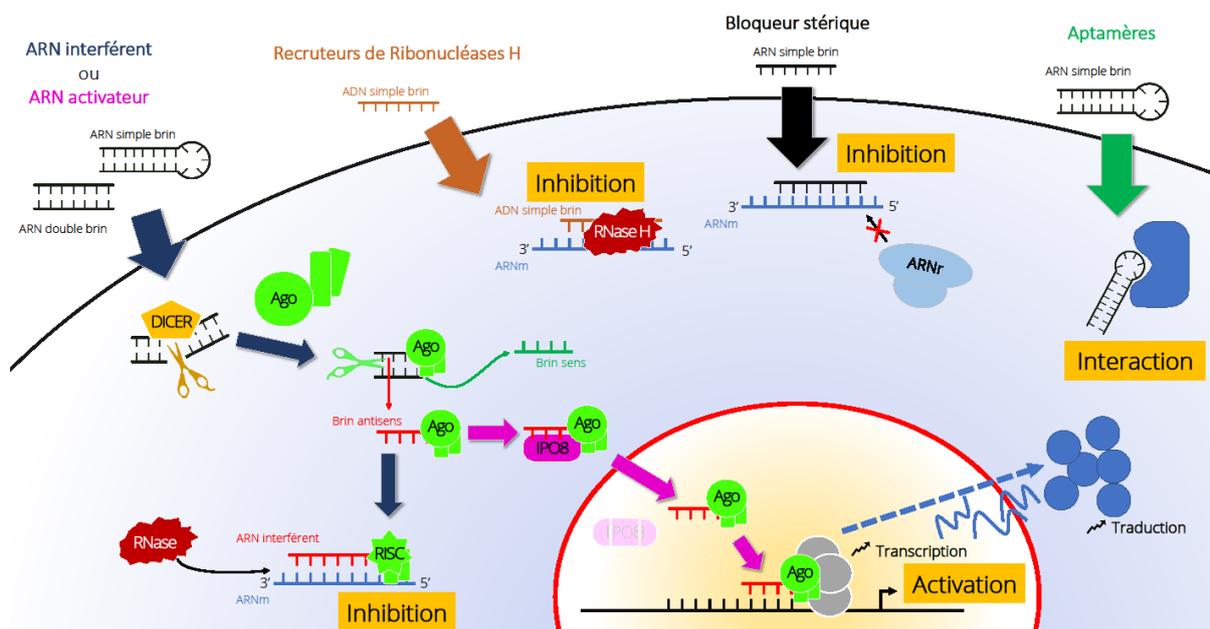


Figure 1-5. Schéma récapitulatif des applications cliniques possibles des oligonucléotides et les cascades de réactions induisant inhibition ou activation de l'expression d'un gène cible ou interaction avec une cible protéique.

### 3.1. Sondes

La capacité d'hybridation des oligonucléotides, combinée aux modifications chimiques, permet leur utilisation comme sondes de détection des ADN ou ARN complémentaires. En effet, l'ajout d'un groupement fluorescent ou marqué radioactivement permet la détection ou la quantification *in vitro* des séquences cibles d'ADN ou ARNm après migration et séparation sur un gel d'électrophorèse, respectivement *southern blot* et *northern blot*. L'hybridation *in situ* en fluorescence (FISH en anglais) permet de réaliser ce type de détection sur des échantillons cellulaires ou tissulaires pour identifier les mutations chromosomiques

<sup>1</sup> <https://rfam.org/>

(Amann & Fuchs, 2008). Enfin, il est fréquent d'ajouter ce type de marqueur sur les amorces pour l'amplification d'ADN (par Polymerase Chain Reaction - PCR). Ainsi, pendant un processus d'amplification des acides nucléiques, le bon déroulement de l'expérience peut être facilement vérifié et le brin amplifié peut être différencié du brin guide grâce à sa fluorescence.

### 3.2. Régulation de l'expression des gènes

En thérapie, les oligonucléotides sont majoritairement utilisés dans la régulation des fonctions cellulaires en agissant sur le matériel génétique impliqué dans cette expression. Les oligonucléotides développés sont qualifiés d'antisens, car ils correspondent à de courts fragments ARN ou ADN complémentaires d'une séquence donnée. Ces fragments sont synthétisés pour moduler l'expression de certains gènes en utilisant leur capacité d'hybridation avec leur brin complémentaire cible. Le développement de ce type de traitement est ainsi dépendant de nos connaissances de la pathologie que l'on souhaite cibler, notamment les voies de biosynthèses et les gènes impliqués. Parallèlement, la complémentarité de l'oligonucléotide développé se doit d'être spécifique à un gène d'intérêt sans parasiter d'autres fonctions en s'hybridant à d'autres fragments.

Les recruteurs de Ribonucléases H font intervenir des nucléases capables de reconnaître les hybrides ARN-ADN et d'entraîner leur hydrolyse (Figure 1-5). Ces oligonucléotides ADN se lient donc aux ARNm codant pour une fonction définie. L'hybride ADN/ARNm devient donc la cible de la ribonucléase H qui dégrade l'hétéroduplex prévenant ainsi la synthèse de la protéine correspondante. (Crooke, 2017).

D'autres oligonucléotides antisens, les bloqueurs stériques, ne sont pas compétents pour le recrutement de nucléases, mais la simple hybridation à la cible affecte leur activité (Figure 1-5) car les transcrits ciblés, une fois hybridés, sont incapables d'exercer leur fonction. Parmi les applications possibles, les bloqueurs stériques peuvent bloquer les interactions des ARN liant les ARN ou les protéines, moduler l'épissage alternatif ou affecter le processus de traduction (Aartsma-Rus et al., 2017; Boiziau et al., 1991).

Les ARN interférents (ARNi) peuvent se présenter sous forme de fragments double (petits ARN interférents) ou simple brin (micro ARN interférents, miARN). Les petits ARNi sont formés à partir d'une longue chaîne d'ARN double brin. Le mécanisme nécessite l'intervention d'une

endo-ribonucléase DICER qui va fragmenter la séquence en plusieurs ARNi qui vont pouvoir s'hybrider avec l'ARN messager cible et inhiber ou sous-réguler la synthèse de la protéine correspondante. Les ARNi générés sont guidés par la protéine Argonaute 2 (Ago 2) et forment alors un duplex en s'hybridant avec l'ARN messager cible et intègre alors le *RNA-induced silencing complex* qui induira un clivage de l'ARNm (Figure 1-5).

Pour les ARN activateurs, tout comme les ARN interférents, le mécanisme repose sur le recrutement des endo-ribonucléases DICER et la protéine Ago 2 pour ne mobiliser qu'un unique brin guide. Celui-ci traverse la membrane du noyau par l'intermédiaire de la protéine importine 8 et s'hybride sur le promoteur de la région cible en recrutant d'autres protéines et former un complexe d'activation transcriptionnelle pour augmenter l'expression du gène souhaité (Figure 1-5).

Dès les années 1970, l'utilisation des oligonucléotides à des fins thérapeutiques était envisagée mais limitée par les propriétés physicochimiques de ce type de molécules, jugées peu stables, peu spécifiques et disposant d'une mauvaise biodisponibilité. Plusieurs années de recherches dans le domaine ont permis d'identifier le potentiel thérapeutique des oligonucléotides agissant sur les facteurs de transcriptions ou sur certaines protéines difficilement accessibles ou peu spécifiques pour d'autres molécules (Moumné et al., 2022). Aussi, la synthèse d'oligonucléotides aux propriétés physicochimiques améliorés a permis le développement en 1998 du premier oligonucléotide approuvé à la commercialisation par la Food and Drug Administration (FDA) sous le nom de Fomivirsen et utilisé pour lutter contre les infections du cytomégalovirus (CMV)(Roehr, 1998). Il est également le seul oligonucléotide utilisé dans le traitement d'une maladie virale. Par la suite, de nombreux oligonucléotides ont été développés, et en 2021, on dénombre 14 composés commercialisés, et près de 80 autres en phase d'essai clinique II ou III indépendamment du mode d'action, accessible sur Clinicaltrials.gov<sup>2</sup>. Tous les composés commercialisés sont impliqués dans le traitement de maladies liées à des mutations génétiques, héréditaires ou non (Tableau 1).

---

<sup>2</sup> <https://clinicaltrials.gov/>

Tableau 1. Présentation des oligonucléotides commercialisés, avec leur nom commercial, le type d'action, la compagnie qui a produit le composé, la cible et la pathologie traitée. Repris et adapté de Moumné et al. (2022)

<b>Nom commercial</b>	<b>Type</b>	<b>Compagnie Pharmaceutique</b>	<b>Cible</b>	<b>Pathologie</b>
<b>Nusinersen</b>	Oligonucléotide antisens	Biogen	ARN pre-messenger SMN2	Amyotrophie spinale
<b>Eteplirsén</b>	Oligonucléotide antisens	Sarepta Therapeutics	ARN pre-messenger DMD	Myopathie de Duchenne
<b>Inotersen</b>	Oligonucléotide antisens	Akcea Therapeutics	ARN messenger TTR	Amylose cardiaque à transthyrétine
<b>Viltolarsen</b>	Oligonucléotide antisens	NS Pharma	ARN pre-messenger DMD	Myopathie de Duchenne
<b>Casimersen</b>	Oligonucléotide antisens	Sarepta Therapeutics	ARN pre-messenger DMD	Myopathie de Duchenne
<b>Golodirsén</b>	Oligonucléotide antisens	Sarepta Therapeutics	ARN pre-messenger DMD	Myopathie de Duchenne
<b>Mipomersen</b>	Oligonucléotide antisens	Kastle Therapeutics	ARN messenger apoB-100	Hypercholestérolémie
<b>Volanesorsen</b>	Oligonucléotide antisens	Akcea Therapeutics	ARN messenger apoC-III	SHCF
<b>Fomivirsén</b>	Oligonucléotide antisens	Novartis	ARN messenger IE2	Infection à CMV
<b>Patisiran</b>	Petits ARN interférents	Alnylam Pharmaceuticals	ARN messenger TTR	Amylose cardiaque à transthyrétine
<b>Givosiran</b>	Petits ARN interférents	Alnylam Pharmaceuticals	ARN messenger ALAS1	Porphyrie hépatique aiguë
<b>Inclisiran</b>	Petits ARN interférents	Novartis	ARN messenger PCSK9	Maladie cardiovasculaire athéroscléreuse
<b>Lumasiran</b>	Petits ARN interférents	Alnylam Pharmaceuticals	ARN messenger HAO1	Hyperoxalurie Type 1
<b>Vutrisiran</b>	Petits ARN interférents	Alnylam Pharmaceuticals	ARN messenger TTR	Amylose cardiaque à transthyrétine

### 3.3. Aptamères

Certains oligonucléotides possèdent la capacité de se fixer à une molécule ou à une structure biologique de quelconque nature de manière spécifique. Dans ce contexte, les aptamères oligonucléotidiques sont des courts fragments d'acides nucléiques simple brin capables de se lier avec une haute affinité aux molécules qu'ils ciblent spécifiquement. Notons que, dans son acception large, le terme "aptamère" désigne une biomolécule structurée ayant été sélectionnée pour sa capacité de reconnaissance ou de régulation de l'activité d'une cible. Les aptamères peuvent donc être des chaînes polypeptidiques structurées (cyclisées par exemple) ou des oligonucléotides simple brin. Dans le contexte de ce manuscrit, le terme aptamère sera utilisé pour les aptamères oligonucléotidiques uniquement.

Les aptamères se caractérisent par la grande diversité des cibles avec lesquelles ils peuvent interagir : protéines, acides nucléiques, membranes cellulaires, cellules (récepteurs cellulaires), ou petites molécules chimiques. Leur capacité d'interaction dépend fortement de leur structure, notamment grâce à la variété de conformations accessibles aux oligonucléotides, qui est directement associée à la capacité de former des hélices et motifs particuliers grâce à l'appariement des bases (Figure 1-5).

L'affinité d'un aptamère pour sa cible est déterminée à partir de la constante de dissociation à l'équilibre ( $K_D$ ). On estime que l'interaction entre un aptamère et sa cible atteint des valeurs d'affinité de l'ordre du nano- au picomolaire (Jenison et al., 1994; Win et al., 2006), gamme comparable à la gamme de  $K_D$  mesurée pour des complexes anticorps/antigènes. Cette affinité est favorisée par la complémentarité de conformation avec la cible qui va dépendre de la nature de cette dernière, mais également par les interactions physiques créées entre l'aptamère et sa cible. Parmi ces interactions on retrouve les interactions hydrophobes et celles de van der Waals entre le squelette phosphate et la cible, ou encore les liaisons hydrogène entre les bases non appariées ou les groupements phosphate du squelette phosphate et les atomes de la cible. Notons que le groupement hydroxyle en 2' des aptamères à ARN peut également être impliqué dans des liaisons hydrogène. Les contacts non-polaires jouent également un rôle majeur dans l'affinité, impliquant en général le ribose (pour les ARN) ou les bases pyrimidiques (Joseph L. Kim et al., 1993).

L'utilisation d'oligonucléotides aptamères (ayant donc été sélectionnés contre une cible d'intérêt) pour réguler l'activité de certaines biomolécules a été envisagée dès les années 1990. Le premier aptamère validé à la commercialisation par la FDA est le Macugen ou Pegaptanib en 2000. Il a été développé pour cibler le facteur de croissance de l'endothélium vasculaire (*Vascular endothelial growth factor, VEGF*) impliqué notamment dans la dégénérescence maculaire liée à l'âge. En 2023, à la différence des thérapies développées à partir d'oligonucléotides ciblant les ARN messenger ou pré-messenger, peu d'aptamères ont dépassé le stade des essais cliniques (Ni et al., 2021).

## 4. Structure des oligonucléotides

Les propriétés et les applications des oligonucléotides dépendent fortement de la structure qu'ils adoptent en condition physiologique. La structure des oligonucléotides est caractérisée par trois niveaux d'arrangements qui définissent leur fonction et impactent directement leur stabilité et donc leur activité.

### 4.1. Structure primaire

La structure primaire désigne l'enchaînement des monomères nucléotidiques (la séquence), en général A, C, G, T/U, associés par des liaisons phosphodiesters au niveau du cycle pentose. Cet enchaînement se lit de l'extrémité 5' à 3'. L'ordre d'apparition des nucléotides dans la séquence influence les niveaux d'arrangement supérieurs.

### 4.2. Structure secondaire

La structure secondaire est directement dépendante de la structure primaire puisqu'elle consiste en l'appariement des bases azotées intra-chaîne nucléotidique par la formation de liaisons hydrogène entre deux bases. Ces liaisons hydrogène peuvent impliquer toute paire de nucléotides (Leontis & Westhof, 2001). Comme mentionné dans la partie 1, on considère comme « canoniques » les paires G-C et A-T/U, définies par Watson & Crick en 1954, car elles apportent une plus haute stabilité à la structure, grâce aux liaisons hydrogène et la distance séparant les atomes de chaque paire.

#### 4.2.1 Appariements et repliements

L'enchaînement de bases appariées ou non appariées induit, pour les oligonucléotides simple brin, un repliement qui peut être caractérisé par plusieurs motifs, identifiés en fonction des appariements formés et de leur position dans la chaîne (Figure 1-6). Un enchaînement de paires de nucléotides formera une structure en hélice, alors qu'un enchaînement de nucléotides non appariés formera une boucle. Une hélice qui se termine par une boucle représente un type de motif, appelé "épingle à cheveux", rencontré fréquemment dans les oligonucléotides repliés et avec une implication majeure dans la formation de structures plus complexes. Au sein d'une hélice, il est possible de trouver des paires de nucléotides qui ne sont pas attenantes aux autres. Plus en détail, les boucles internes et les bourgeons impliquent des interruptions dans l'appariement de bases au sein d'une hélice. Lorsque les nucléotides non appariés se trouvent sur un seul des brins formant l'hélice, on parle de bourgeon. Lorsque les nucléotides non appariés sont sur les deux brins de l'hélice, on parle de boucles internes. Les boucles internes peuvent être symétriques si le nombre de nucléotides non appariées est identique pour les deux brins de l'hélice. Dans le cas contraire, on aura une boucle interne asymétrique. Les boucles internes peuvent être symétriques si le nombre de nucléotides non appariées est identique pour les deux brins de l'hélice. Dans le cas contraire, on aura une boucle interne asymétrique.

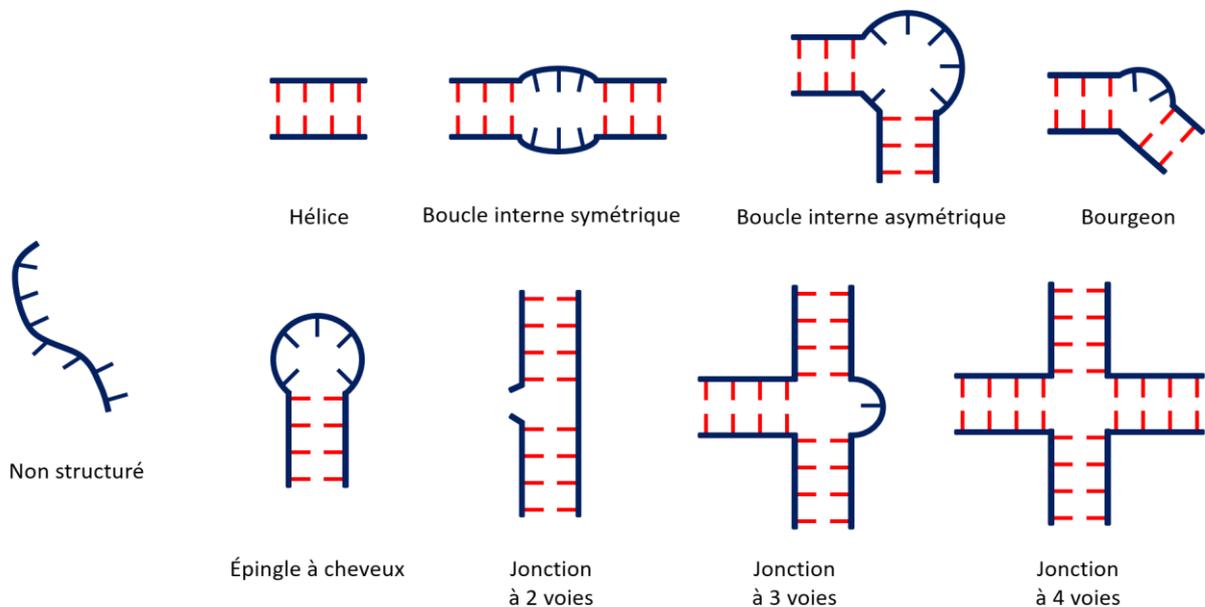


Figure 1-6. Représentation schématique des différentes structures secondaires des acides nucléiques. Repris du manuscrit de thèse de Claire Loussouarn (2014)

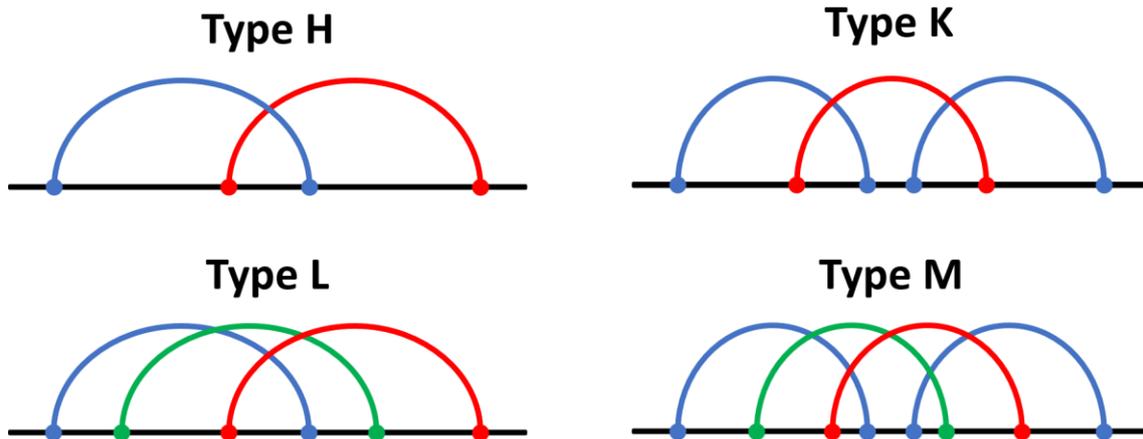
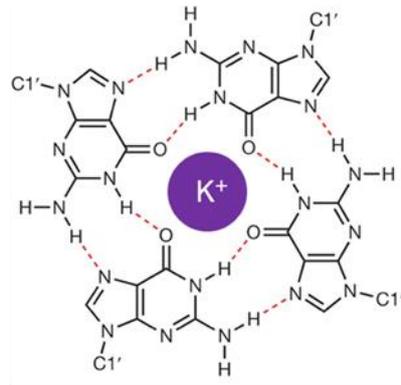


Figure 1-7. Représentation schématique linéaire de l'appariement de pseudonœuds type H, K, L et M.

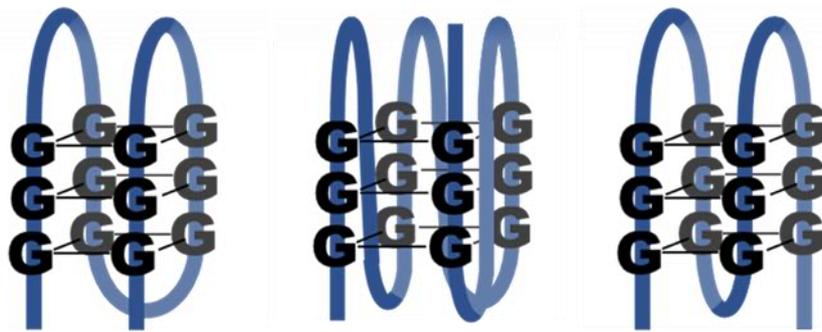
Pour les séquences les plus longues, on peut identifier un autre type de motif appelé jonctions. Les jonctions impliquent plusieurs hélices différenciées à cause d'appariements non contigus. On voit ainsi apparaître plusieurs ramifications de la structure secondaire.

Parallèlement, les longues séquences peuvent également former des interactions à longue distance, souvent limitées à des paires seules entre nucléotides éloignés dans la séquence. Dans ce contexte, les pseudonœuds désignent l'interaction d'un segment non structuré de la chaîne nucléotidiques avec une boucle. On peut les classifier selon 4 niveaux de complexité en fonction du nombre de ces motifs emboîtés : les pseudonœuds de type H indiquant un pseudonœud simple, de type K lorsqu'un second pseudonœud s'intercale, et ainsi de suite pour les types L et M (Figure 1-7).

Enfin, certains acides nucléiques riches en guanine forment un type de structure particulière appelé G-quadruplexes. La formation de ces structures est médiée par la présence de cations monovalent ( $K^+$  ou  $Na^+$ ) s'intercalant dans la cavité centrale qui se crée avec ce type de structure. Ces cations favorisent l'interaction entre quatre guanines formant un plan par un appariement de type Hoogsteen impliquant deux liaisons hydrogène entre guanines 2 à 2 entre l'hydrogène en N1 et l'oxygène 6 d'une part et l'hydrogène en N2 et l'azote 7 d'autre part (Figure 1-8). Les G-quadruplexes résultent d'un enchaînement de ce type d'interaction sur 2 à 4 niveaux (Burge et al., 2006).



G-quadret



Parallèle

Anti-parallèle

Hybride

Figure 1-8. Schéma de formation d'un G-quadret autour d'un cation monovalent potassium  $K^+$  par formation d'appariement Hoogsteen impliquant la formation de liaisons hydrogène en pointillé rouge. Les G-quadret s'enchainent selon 3 topologies possibles, parallèle, anti-parallèle et hybride pour former des G-quadruplexes. Repris de Mishra et al. 2019.

#### 4.2.2 Rôle et implication de la structure secondaire

Le caractère dynamique des oligonucléotides peut engendrer des variations de structures affectant la structure secondaire. On décrit alors la structure des oligonucléotides comme un ensemble de conformations favorables, dépendantes de la séquence. Ces fluctuations structurales sont associées à des changements spontanés et continus observables dans tout système biologique dynamique. On désigne 2 types d'états dynamiques observables : i) les états de transition conformationnelle incluant les conformations minoritaires nécessaires pour atteindre un état plus stable ii) les fluctuations à l'équilibre, observées pour une conformation donnée qui se maintient dans le temps. Certains facteurs peuvent impacter la structure secondaire en solution, les plus importants étant la température, le pH et la concentration saline du milieu.

La structure secondaire du brin oligonucléotidique est fondamentale dans certaines fonctions de régulation. Plus précisément, la structure secondaire des ARN représente un moyen de réguler le processus de transcription (Figure 1-9) : la simple présence d'une structure secondaire de type épingle à cheveux peut entraîner son arrêt par la dissociation du complexe d'élongation, lorsque l'oligonucléotide est couplé à un facteur de terminaison, ou la poursuite de l'élongation. Dans d'autres cas, l'épingle concernée peut déclencher la dissociation du complexe à elle seule si c'est un terminateur ou, à l'inverse, prolonger l'élongation s'il s'agit d'un anti-terminateur (Chetkowska-Pauszek et al., 2021).

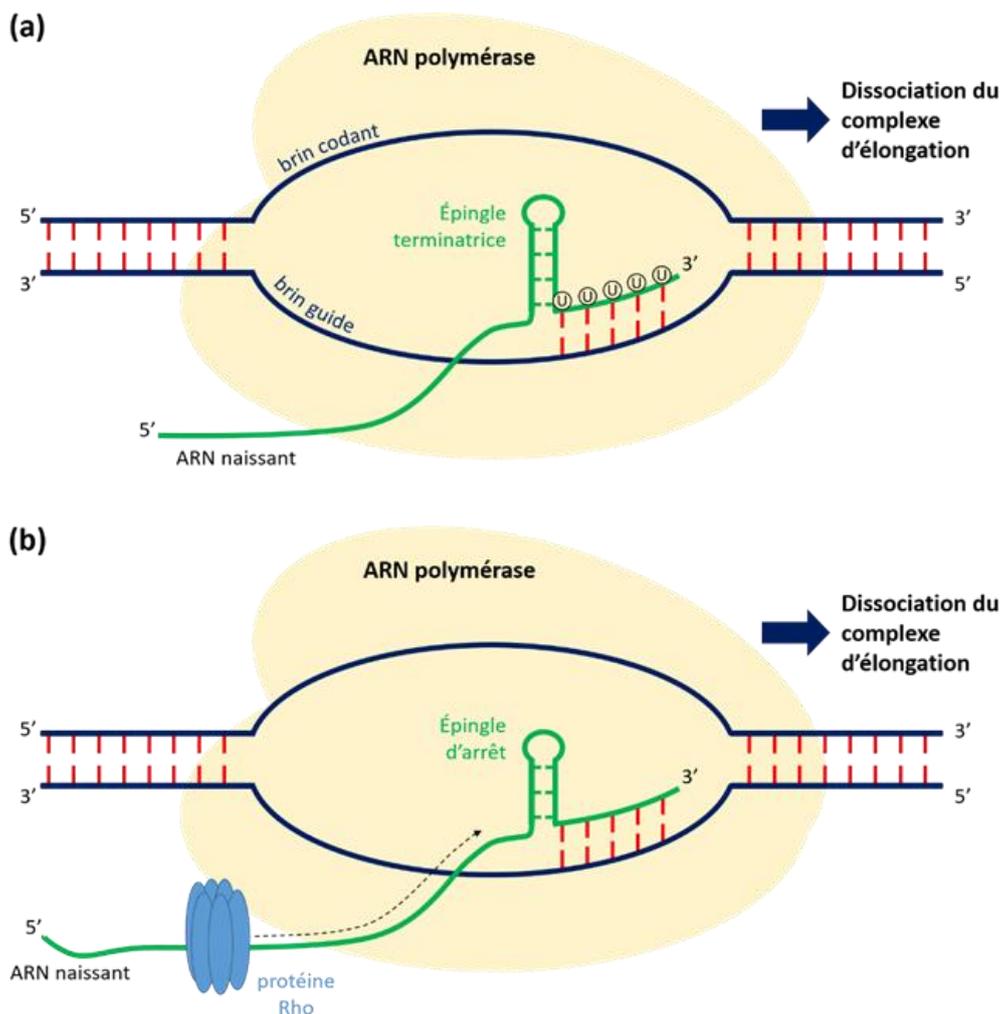


Figure 1-9. Schéma expliquant l'impact d'une structure secondaire sur le processus de transcription de l'ADN en ARN messager. En (a) est détaillé le processus sans intervention de la protéine Rho où l'épingle positionnée en aval couplée à la région riche en uracile entraîne la dissociation du complexe d'élongation. Le processus (b) où Rho intervient comme facteur de dissociation du complexe d'élongation suite à l'arrêt engendré par la présence d'une épingle. Illustration inspirée de *Chetkowska-Pauszek et al. 2021*

La structure secondaire peut également influencer sur la traduction lorsqu'elle implique l'ARNm mature. L'accès au site de fixation du ribosome (RBS) peut être perturbé par différents moyens. Les riboswitchs ou les thermosenseurs, décrits plus haut, sont des régulateur Cis, c'est-à-dire qu'ils sont localisés sur des portions de gène proches et affectent la traduction grâce à leurs propriétés structurales. Les thermosenseurs localisés dans une portion non-codante de l'ARNm réagissent aux changements de température avec une modification de leur structure secondaire, ce qui libère ou masque le site de liaison du ribosome (RBS). Les riboswitchs se décomposent en un domaine aptamère qui présente une haute spécificité pour son ligand activateur et un domaine d'expression (Breaker, 2011). La liaison au ligand entraîne d'importants changements structuraux bloquant l'accès au RBS. De la même façon, les éléments régulateurs Trans, qui sont des éléments localisés sur des gènes éloignés (en opposition aux régulateurs Cis), peuvent provoquer une disparition d'éléments de structure secondaire au sein de l'ARNm pour libérer le site de fixation du ribosome et déclencher la traduction.

La structure secondaire joue également un rôle important dans le processus d'épissage alternatif (Bartys et al., 2019; Buratti & Baralle, 2004). Les régulateurs de l'épissage se positionnent le long de l'ARN pour exercer leur activité. La présence de structures secondaires sur des positions clé de l'épissage peut entraîner des décalages. En conséquence, l'ARN messenger mature peut inclure ou exclure des exons selon un schéma différent de ce qui est attendu à l'état natif.

### 4.3. Structure tertiaire des oligonucléotides

La structure tertiaire d'un oligonucléotide désigne ses repliements et conformations adoptés dans l'espace à trois dimensions. Ces repliements sont directement responsables de la capacité de ces molécules à catalyser des réactions ou reconnaître des sites de liaison sur les ligands ciblés.

La notion de structure tertiaire des oligonucléotides est liée à celle de la structure secondaire et influence le positionnement des nucléotides permettant la formation des paires de bases. Cependant, la structure secondaire ne tient pas compte de la flexibilité globale des oligomères. En effet, l'appariement des bases implique certaines contraintes d'angles dièdres sur les atomes composant le squelette phosphate (Murray et al., 2003). L'enchaînement de

nucléotides appariés structure une hélice en raison des dièdres formés par le squelette phosphate dans la chaîne. Les oligonucléotides structurés les plus complexes seront composés en majorité de ce type de structure, car elle engendre une forte stabilisation, grâce à la succession de liaisons hydrogène entre les paires de bases. Les hélices, au sein des acides nucléiques simple brin, disposent des mêmes caractéristiques que les hélices d'acides nucléiques double brin puisque ceux-ci se construisent autour d'appariements de nucléotides. La configuration spatiale des hélices est influencée par plusieurs facteurs importants. Le premier est le « pli » formé par le cycle ribose dans le nucléotide. Trois géométries d'hélices peuvent alors se former : Forme A, Forme B ou Forme Z (Figure 1-10b), les deux premières étant les géométries les plus fréquemment rencontrées (Eichhorn & Al-Hashimi, 2014; Ussery, 2002). Ensuite, les liaisons hydrogène entre les paires de bases, notamment leur nombre et leur stabilité, vont fortement impacter la structure de ces hélices en modifiant la taille du grand ou du petit sillon (Figure 1-10a). Ainsi, l'enchaînement des nucléotides aura son importance au même titre que les conditions du milieu dans lequel l'oligonucléotide est placé (température, concentration en ions, pression).

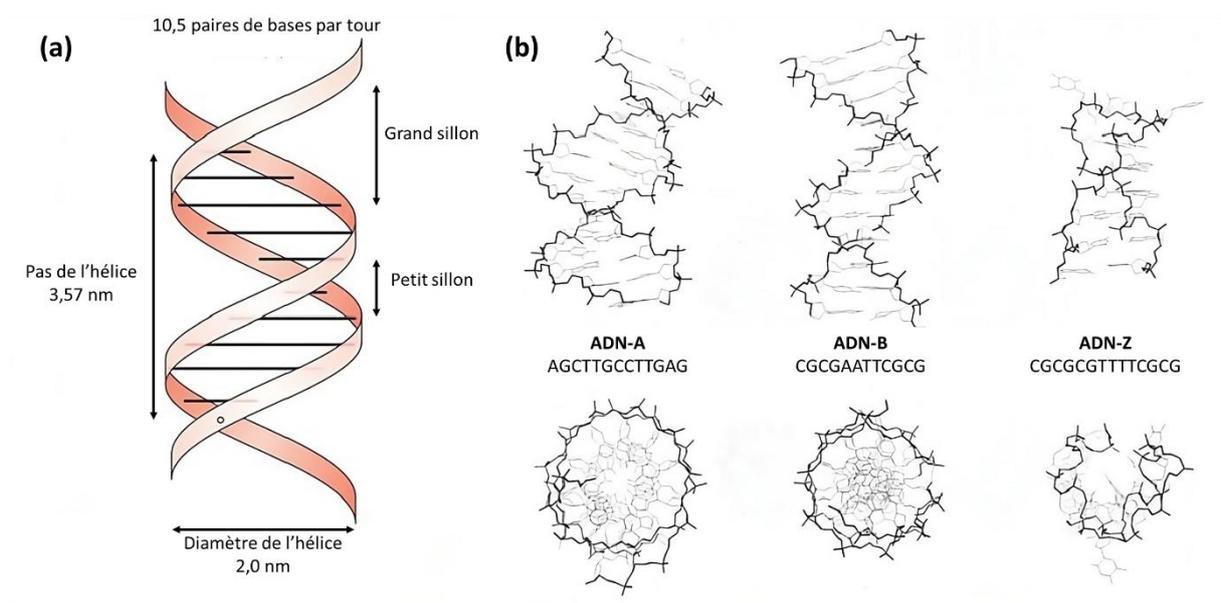


Figure 1-10. Visualisation schématique des hélices d'ADN. (a) Agencement de l'hélice d'ADN-B en trois dimensions avec la représentation des unités influençant leur structure. (b) Visualisation des hélices d'ADN-A, -B et -Z vue de face (en haut) et vue du dessus (en bas). Illustration reprise de Ussery (2002).

Les épingles à cheveux et les boucles internes sont également influencées par les paires de bases qui les précèdent mais résultent en une très grande diversité de conformations en fonction de leur taille.

Enfin, la structure tertiaire des acides nucléiques est fortement influencée par les interactions à longue distance permettant aux oligonucléotides de former une structure plus compacte. Les pseudonœuds, précédemment abordés, vont permettre l'interaction d'un ensemble de nucléotides distants entre eux, entraînant la formation d'une structure plus stable et plus compacte (Staple & Butcher, 2005). Lorsque des interactions à longue distance font intervenir des nucléotides libres issus de 2 épingles à cheveux, le motif formé est un *kissing loop*.

La détermination expérimentale des structures des oligonucléotides a permis de construire des bases de données regroupant et classifiant ces motifs qui se conservent d'une structure à l'autre malgré la diversité (Petrov et al., 2013). Ainsi, les connaissances sur ces différents niveaux d'organisation structurale ont été fortement enrichies par l'augmentation croissante des structures obtenues et l'amélioration des techniques d'acquisition de la structure.

#### 4.4. Obtention de la structure des oligonucléotides

Comme pour les protéines, la structure des acides nucléiques peut être déterminée expérimentalement. Plusieurs méthodes existent et sont appropriées pour l'obtention de la structure tertiaire de molécules de plus ou moins grande taille (Figure 1-11). La taille des molécules et les méthodes utilisées vont influencer sur la résolution atomique de la structure, définie par la capacité à distinguer deux atomes adjacents. Une résolution de bonne qualité se mesure en Å et plus la valeur est basse, plus il est aisé de différencier les atomes.

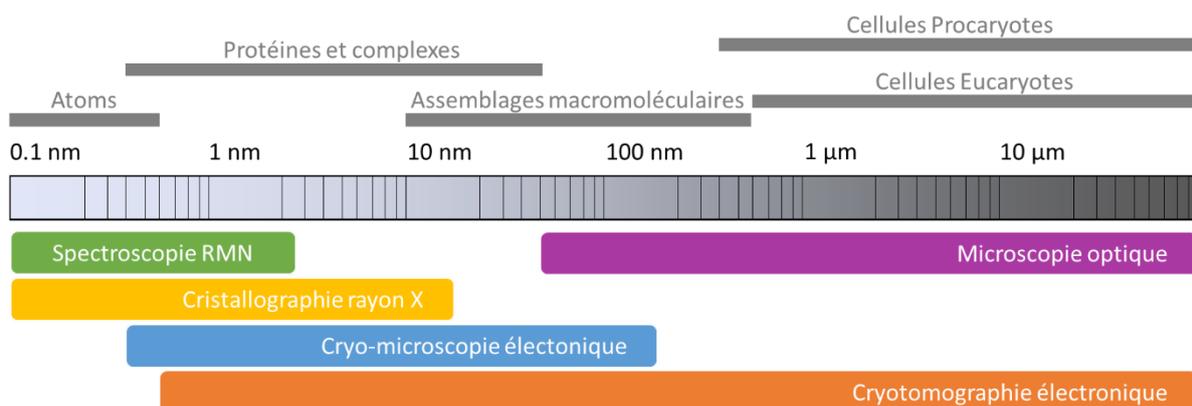


Figure 1-11. Echelles de taille de l'atome à la cellule et compatibilité des techniques d'acquisition de structure. Figure reprise de Leigh et al. 2019

#### 4.4.1 Cristallographie par rayon X

La cristallographie par rayons X est une méthode fiable d'obtention de la structure d'une molécule. C'est la plus ancienne technique de résolution de structures de cristaux à l'échelle atomique car développée au début du 20<sup>e</sup> siècle et a fortement contribué à l'évolution des connaissances de la structure à l'échelle moléculaire. En plus de fournir une résolution atomique de bonne qualité, cette méthode n'a pas de limite technique théorique pour la taille des échantillons à cristalliser (Figure 1-11). Néanmoins, des structures de grande taille aboutissent souvent à des résolutions plus hautes, entraînant une perte d'information.

La cristallographie par rayons X utilisée sur les acides nucléiques permet d'obtenir aussi bien la structure du matériel génétique de virus complet que celle de complexes protéine-aptamère (Ruigrok et al., 2012). Les inconvénients de cette méthode résident dans les propriétés de la molécule que l'on souhaite cristalliser. Elle doit être cristallisable et suffisamment pure et soluble pour produire un cristal de bonne qualité. L'inconvénient majeur pour la cristallisation des acides nucléiques est lié aux charges négatives du squelette phosphate exposé à la surface de la molécule. Cela complexifie la cristallisation car le processus repose sur la formation de contacts (interactions électrostatiques) au sein de la même molécule et le manque de points de contacts stables à cause de l'homogénéité de la structure des acides nucléique altère la qualité du cristal par rapport à une protéine plus hétérogène disposant de nombreux points de contact intermoléculaires. De plus la structure tertiaire des fragments non-appariés des oligonucléotides peut également être impactée par cette absence de contact lors de la formation du cristal. Ainsi la préparation du cristal nécessite de prendre en considération les variations dans la structure (flexibilité, forme, longueur de la chaîne, présence de sites de liaisons), ce qui rend la préparation non homogène entre échantillons. Pour ces mêmes raisons, les oligonucléotides obtenus par cristallisation sont fréquemment en complexe avec une autre macromolécule (Holbrook et al., 1991; Mooers, 2009). Enfin l'image tri-dimensionnelle produite ne donne aucune information sur la mobilité conformationnelle des oligonucléotides, ce qui reste le principal obstacle dans l'étude de la structure de ce type de molécule (Felden, 2007; Ke & Doudna, 2004).

#### 4.4.2 Spectroscopie à Résonance Magnétique Nucléaire

La spectroscopie à résonance magnétique nucléaire (RMN) est également une méthode utilisable pour la détermination de la structure des acides nucléiques (Schnieders et al., 2020). Le spectre RMN fournit les informations des liaisons entre atomes voisins pour acquérir la structure primaire, puis les différents niveaux de repliements (structures secondaire et tertiaire) sont obtenus par analyse des contraintes de distances et d'angles. Seules les conformations pertinentes sont conservées car elles disposent des contraintes d'angles et de distance compatibles. La préparation des échantillons pour la RMN est simple car ce type d'analyse peut être directement fait sur l'échantillon en solution. Il devient donc possible d'observer la structure des molécules en solution qui implique moins de contraintes que lors du processus de cristallisation (G. Wang et al., 2014). La RMN est également une méthode qui permet d'obtenir plusieurs conformations de la molécule qui peuvent exister dans l'échantillon. Cette technique est ainsi adaptée à l'étude de la mobilité de structure des oligonucléotides et il devient donc possible d'observer l'impact des mutations et de l'environnement sur la structure et la mobilité des oligonucléotides. Par exemple, il est possible d'étudier les oligonucléotides thermosensibles qui subissent des changements conformationnels après un changement brutal de température (Chowdhury et al., 2006). Néanmoins, des artefacts peuvent également apparaître durant l'acquisition du spectre RMN car le matériel est sensible aux perturbations. En outre, la spectroscopie RMN est plus difficile à réaliser sur des molécules de grande taille à cause du phénomène de superposition du spectre, qui peut rendre difficile l'acquisition de certaines informations de liaison dans la structure et une perte de structure en conséquence pendant la lecture. Ce défaut limite l'utilisation de la RMN à des ensembles de molécules relativement petites, fixant la limite autour de 100 nucléotides pour l'obtention des structures d'acides nucléiques (Felden, 2007; Marušič et al., 2023) (Figure 1-11).

#### 4.4.3 Cryo-microscopie électronique

Plus récemment la cryo-microscopie électronique (cryo-ME) est devenue une alternative pour l'obtention de structures moléculaires. Cette technique est applicable à tout type d'échantillons biologiques entre 1 et 100 nm (Figure 1-11). Elle permet d'obtenir des structures de grande taille, comme des systèmes cellulaires entiers ou des complexes protéines / acides nucléiques mais également de visualiser plusieurs conformations dans

l'échantillon cryogénisé sans la contrainte de taille de la RMN, le tout présentant une bonne résolution (Frank, 2002). De ce fait, elle présente les avantages des deux techniques précédentes.

L'approche de cryo-microscopie se divise en deux techniques : i) analyse de particule unique ii) cryo-tomographie électronique (cryo-TE). La première se fait par observation de l'échantillon cryogénisé pour en exploiter les informations bi-dimensionnelles et en déduire la conformation tri-dimensionnelle. En cryo-TE, la structure de l'échantillon est reconstruite par l'acquisition en plusieurs étapes de la tomographie sous différents angles. La taille des échantillons observables est bien supérieure en cryo-TE, mais reste plus complexe que l'analyse de particule unique. Bien que la préparation des échantillons soit plus fastidieuse, la cryogénéisation présente moins de contraintes que la cristallisation car tout échantillon est potentiellement sensible à la cryogénéisation (Noble et al., 2018). Appliquées à l'étude de la structure des oligonucléotides, les techniques de cryo-microscopie ont permis l'obtention de la structure des échantillons de ribonucléoprotéine de grande taille comme le splicéosome (Golas et al., 2005), allant jusqu'à la visualisation des changements conformationnels sur des bactériophages entiers (Gorzelnik & Zhang, 2021).

#### 4.4.4 Techniques pour la détermination de la structure secondaire

Certaines méthodes de détermination de la structure secondaire ont permis de simplifier l'obtention de la structure secondaire des oligonucléotides sans avoir recours à des processus complexes comme les 3 techniques précédemment abordés. Ces techniques expérimentales chimiques et enzymatiques permettent de se soustraire à l'utilisation de matériel complexe pour la résolution de la structure tertiaire, et par conséquent de la structure secondaire. L'approche DMS-Map (pour Dimethyl Sulfate-Sequencing Mapping en anglais) (Zubradt, M., Gupta, P., Persad, S., Lambowitz, M. A., Weissman, S. J. & Rouskin, S. (2017) DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. *Nat Methods*, 14, 75–82.

) permet de retrouver la structure secondaire de fragment d'ARN in vivo en appliquant des modifications chimiques aux bases azotées. Le diméthylsulfate (DMS) dans lequel sont incubées les cellules permet de modifier les bases azotées exposées soit celles qui ne sont pas impliquées dans des appariements. Cela permet d'obtenir après rétrotranscription des fragments d'ADN ponctuellement mutés. De la même façon, la méthode SHAPE (pour

Selective 2'-Hydroxyl Acylation and Primer Extension en anglais) (Smola, M., Rice, G., Busan, S., Siegfried, A. N. & Weeks, M. K (2015). Selective 2'-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. *Nat Protoc*, 10, 1643–1669.

) applique des modifications des bases azotées exposées dans la séquence. Après amplification et séquençage, les fragments obtenus peuvent être associés à des portions structurées ou non.

L'approche RIC-seq (pour RNA In Situ Conformation Sequencing en anglais) (Cai, Z., Cao, C., Ji, L., Ye, R., Wang, D., Xia, C., Wang, S., Du, Z., Hu, N., Yu, X., Chen, J., Wang, L., Yang, X., He, S. & Xue Y. (2020) RIC-seq for global in situ profiling of RNA–RNA spatial interactions. *Nature* 582, 432–437.

) permet d'étudier le mécanisme d'interaction des molécules d'ARN au sein des cellules. Cette approche peut ainsi être exploitée pour déterminer la structure secondaire de fragments souhaités. L'ARN est d'abord mis en condition dans des cellules afin de permettre la formation et la préservation des interactions. Les ARN sont marqués puis une ligase T4 est ensuite utilisée pour lier les ARN proches formant des brins chimériques avant de fragmenter à nouveau cette banque. Les fragments marqués conservés sont amplifiés puis séquencés, permettant alors de retrouver la structure secondaire

#### 4.5. Ressources disponibles et variété de structures dans les bases de données

Les structures des oligonucléotides, libres ou en complexe avec des autres molécules, obtenues expérimentalement, peuvent être déposées et consultées dans des bases de données publiques dédiées. Notamment, les bases de données Protein Data Bank (PDB) et Nucleic Acid Database (NDB) (Berman et al., 1992, 2002) répertorient les données structurales pour de nombreuses molécules. Les structures et leurs annotations sont disponibles en libre accès et ont fortement contribué à l'enrichissement et au partage des connaissances sur les biomolécules. A ce jour, la PDB compile les informations structurales expérimentales de

210 868 molécules, dont 182 735 protéines seules, 4 320 acides nucléiques seuls et 12 219 acides nucléiques complexés avec une protéine<sup>3</sup> (Figure 1-12).

L'accroissement des données disponibles sur les acides nucléiques, comme pour les autres molécules, est lié au développement des méthodes de résolution de structures toujours plus fiables et plus précises et il est facilité par notre compréhension de ce type de molécules. Ainsi, depuis le début des années 2000, le nombre de structures d'acides nucléiques libres disponibles sur la PDB a été multiplié par 5, et pour les structures présentant au moins une unité de type acide nucléique c'est jusqu'à 12 fois plus (ADN, ARN ou hybride).

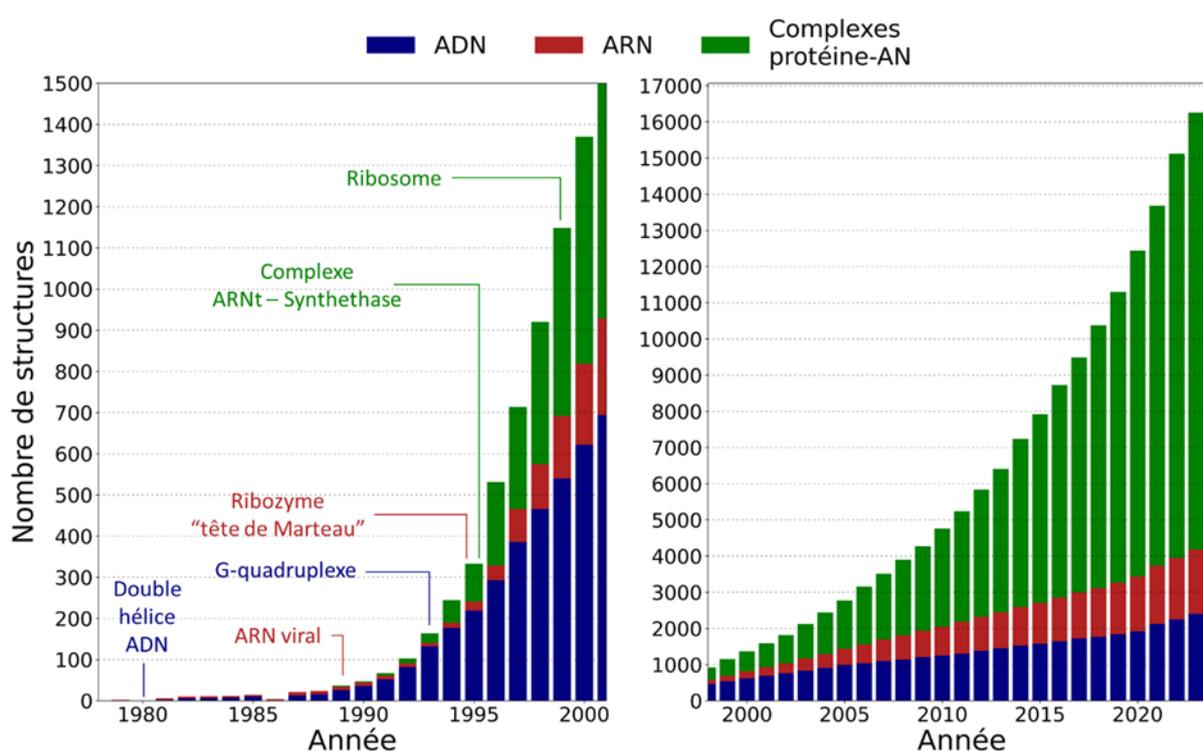


Figure 1-12. Evolution du nombre de structures disponibles publiées sur la PDB inspiré de la revue de Westhof (2021) et complété à partir des données de la PDB. A gauche, à l'échelle 0-1500 structures, sont indiqués les structures ayant impacté la recherche sur les acides nucléiques avant 2000. A droite, est indiquée l'évolution croissante du nombre de structures disponibles entre 2000 et 2023.

Au fil des découvertes, les bases de données ont été enrichies des connaissances sur ces biomolécules au travers des structures tridimensionnelles obtenues. Les principes de base de composition chimiques, de rotations des angles dièdres dans les pentoses qui composent les nucléotides, ou encore les règles d'appariement des bases ont pu être définis très tôt entre

<sup>3</sup> <https://www.rcsb.org/stats/summary>

les années 50 à 70. Jusqu'aux années '90, l'étude de structures d'ARN de transfert (ARNt) a fortement contribué à comprendre le rôle des appariements de nucléotides sur la formation des doubles hélices et des motifs de la structure secondaire. La complexité des structures des ARNt contribue à l'étude des repliements tridimensionnels avec l'identification des interactions à courte et longue distance. De ces interactions résultent différents repliements complexes influencés par la structure secondaire et les interactions à longue distance divisant la molécule en différents domaines. Ces domaines ont été étudiés et il a été démontré qu'ils étaient associés à certains motifs agissant comme sites accepteurs pour les interactions et la catalyse de réactions, ce qui appuie l'hypothèse d'une relation structure-fonction des oligonucléotides et souligne l'importance des repliements bi/tri-dimensionnels (Grasby & Gait, 1994; Kim, 1978). C'est également durant cette période que le contexte de formation des appariements non canoniques a été observé et étudié pour répondre à certaines interrogations, notamment concernant les G-quadruplexes, identifiés depuis 1910, mais dont la disposition spatiale demeurait inexplicée. Ces découvertes ont également révélé l'implication des ions métalliques dans certaines interactions avec les composants nucléotidiques et leur rôle dans la fonction des acides nucléiques (Swaminathan et al., 1979).

Dans les années '90, le premier complexe protéine – ARN a été caractérisé (Rould et al., 1989), marquant ainsi le début de l'accroissement de la complexité des structures obtenues associée à l'amélioration des méthodes de résolution de structures. Le déblocage de ces verrous a permis d'envisager l'étude d'oligonucléotides plus complexes et a engendré une forte croissance du nombre de structures résolues. La première structure de ribozyme « en tête de marteau » a pu être caractérisée en 1995, devenant la première structure d'ARN avec une jonction entre 3 brins connue et répondant à de nombreuses interrogations sur les repliements tridimensionnelles. La caractérisation du premier ribosome en 1999 a renforcé la compréhension du mécanisme de traduction et les fonctions biochimiques impliquées. Par la suite, de nombreuses autres structures ont été obtenues et, en 2002, le partage des informations a permis de regrouper les données structurales sur une plateforme unique, la PDB (Berman et al., 2002) qui est constamment alimentée depuis sa création et contribue fortement à faire progresser la recherche par l'intermédiaire des nombreuses structures publiées annuellement permettant de mieux comprendre les interactions possibles avec les acides nucléiques, le mode d'action et la relation structure-fonction. Bien que la plupart des

connaissances sur les structures d'oligonucléotides simple brin ait été acquise au travers de l'étude des ARN, il a rapidement été admis que les ADN disposaient des mêmes propriétés de repliement (Neidle 2021; Westhof et Leontis 2021).

## 5. Approche *in silico*

Nous avons vu qu'il était possible d'utiliser les techniques de caractérisation de structures comme la cristallographie rayon X, la spectroscopie par RMN ou la cryo-EM. Ces méthodes ont cependant des limites :

- Leur mise en place implique l'utilisation de matériel parfois coûteux, nécessitant une expertise et souvent du personnel dédié, voire de faire appel à des prestataires.
- L'obtention d'une structure résolue est une démarche expérimentale longue, qui peut être compliquée par le comportement de la molécule dans la solution d'analyse (stabilité, solubilité, cristallisabilité...).
- Les quantités de molécules nécessaires à leur exploitation peuvent être difficiles à obtenir.

Par conséquent, pour permettre une progression des connaissances des structures et du comportement des oligonucléotides, les méthodes computationnelles peuvent être une bonne alternative

Les données disponibles dans les bases de données ont donc été exploitées pour la création d'outils de modélisation et de prédiction de la structure secondaire et tertiaire des oligonucléotides. De plus, les approches de dynamiques moléculaires offrent la possibilité de modéliser le comportement des oligonucléotides naturels ou modifiés grâce à l'ajout de paramètres ajustés. L'étude de la mobilité des oligonucléotides peut ainsi être étendue à des acides nucléiques optimisés pour des applications thérapeutiques.

### 5.1. Prédiction de structure secondaire et tertiaire

Notre compréhension des oligonucléotides a permis d'identifier un lien important entre leur structure et leur fonction. De bonnes connaissances de la structure sont donc essentielles pour bien comprendre la spécificité des interactions, la flexibilité et la stabilité de l'oligonucléotide dans les systèmes biologiques ou en solution. Contrairement aux méthodes

de résolution de la structure tridimensionnelle des oligonucléotides, il n'existe pas de méthode expérimentale permettant de déterminer leur structure secondaire, bien qu'il soit possible d'identifier des changements d'état conformationnel grâce au dichroïsme circulaire (Sosnick et al., 2000). Le développement des approches bioinformatiques appliquées à la modélisation des acides nucléiques a apporté de nombreux moyens de prédire la structure secondaire ou tertiaire. Il est alors possible d'anticiper l'intérêt d'une séquence oligonucléotidique à partir des repliements identifiés par les prédictions. Les connaissances sur la structure secondaire ont permis le développement de nombreux algorithmes de prédictions. Ces algorithmes varient par leurs approches, par leurs performances et leur applicabilité. En général, ils permettent d'obtenir une structure secondaire putative de la séquence nucléotidique. Le choix de l'oligonucléotide s'adapte donc en fonction de la présence de structures secondaires qui apportent une meilleure stabilité de l'oligonucléotide.

Les méthodes de prédiction de la structure tertiaire des oligonucléotides sont moins diversifiées en raison d'un manque de structures disponibles. Néanmoins, plusieurs approches ont été développées, basées sur i) l'analyse directe des séquences pour prédire le meilleur repliement ou ii) l'assemblage de fragment basé sur des données expérimentales connues. La structure tridimensionnelle est un fort indicateur de son potentiel d'affinité avec une cible d'intérêt. Il devient possible d'identifier les points clé de la flexibilité d'un oligonucléotide grâce à sa conformation spatiale et sert ainsi de point de départ pour d'autres analyses *in silico* orientées sur l'identification de complexe acide nucléique – cible.

Les approches de prédiction de la structure secondaire et tertiaire seront approfondies dans les Partie 2 et Partie 3, respectivement.

## 5.2. Dynamique moléculaire appliquée aux oligonucléotides

En plus des méthodes explicitement développées pour la prédiction de la structure tertiaire des oligonucléotides, les approches de dynamique moléculaire (MD), qui permettent d'étudier l'évolution d'un système moléculaire en fonction du temps, trouvent leur place dans l'investigation des structures tridimensionnelles des oligonucléotides. En effet, la dynamique moléculaire a évolué pour s'adapter à tout type de molécule et combinaison de molécules pour simuler leur comportement dynamique. La dynamique moléculaire peut ainsi être appliquée strictement pour l'étude du comportement dynamique d'une molécule seule, ou à

des systèmes plus complexes contenant protéines et ligands pour simuler la formation ou la dissociation du complexe (Hu et al., 2017) par des interactions non covalentes. La dynamique moléculaire peut donc apporter énormément d'informations sur le comportement d'une molécule, sur sa mobilité, sur sa stabilité et sur les interactions faibles intra- ou intermoléculaires qu'elle peut former. L'inconvénient de la dynamique réside dans les temps de calcul relativement long en fonction de la taille des systèmes simulés et la quantité de données générées. Cependant, l'évolution du matériel informatique en termes de puissance de calcul et de stockage relègue cette problématique au second plan. De plus, contrairement aux outils précédemment abordés, les approches de dynamiques moléculaires sont compatibles avec l'utilisation des oligonucléotides modifiés pour lesquels des paramètres appropriés peuvent être ajoutés. Ainsi, les simulations ne se limitent pas aux oligonucléotides naturels et l'ensemble conformationnel peut être obtenu pour tout type de molécule.

### 5.3. Complémentarité des outils

Le développement des approches *in silico* pour la modélisation des oligonucléotides permet d'envisager l'utilisation de ces outils comme guide dans le domaine du développement de thérapies à base d'oligonucléotides ou de sondes à visée diagnostique. Dans l'optique de la prédiction de la structure d'oligonucléotides, les approches de prédictions de structures secondaires, tertiaires et de MD permettent d'obtenir à différent niveaux les informations sur la complexité des repliements et sur la mobilité des oligonucléotides (Figure 1-13). Le potentiel de ces outils pour la prédiction de novo d'ARN a été récemment exploité (Yan et al., 2022).

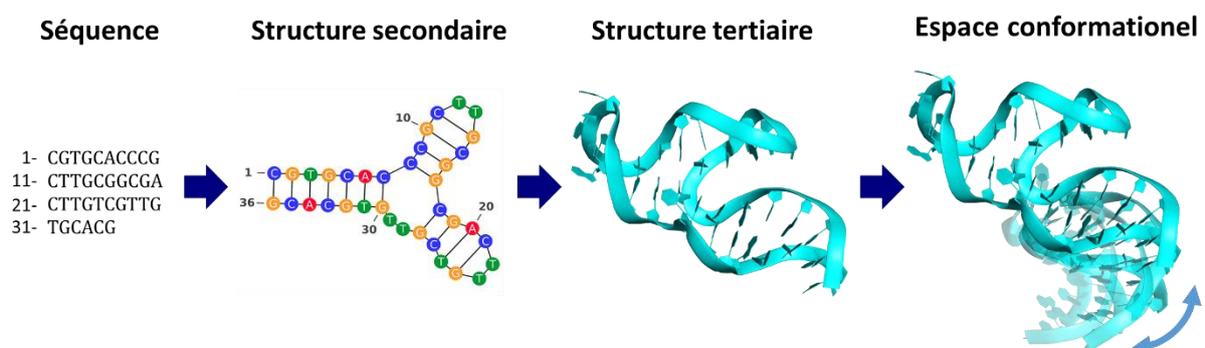


Figure 1-13. Niveaux de complexité de repliement et de mobilité des oligonucléotides.

La littérature présente également quelques travaux compilant toutes ces approches. Une approche envisageable est de combiner la prédiction de structure tridimensionnelle avec la

dynamique moléculaire pour l'exploration de l'espace conformationnel des oligonucléotides, appliquée notamment dans les travaux de Lindert, et al. 2013 qui procède en un enchaînement d'étapes de dynamique moléculaire puis de reconstruction avec RNAde novo (Leaver-Fay et al., 2011). Zhang et al. (2019) ont également testé un protocole de prédiction de structure d'aptamères, amarré sur des composés perturbateurs endocrinien par *docking* puis testé en MD afin de déterminer la stabilité du complexe, critère important pour la conception de biocapteurs.

Les travaux ici présentés s'insèrent dans ce contexte. En partant d'une séquence d'oligonucléotides, les différents niveaux de complexité sont obtenus d'abord par la prédiction de structure secondaire, qui permet de contraindre la prédiction de la structure tertiaire avant de procéder à l'échantillonnage intensif des conformations grâce aux approches de MD accélérées sur de longues durées de simulations (1  $\mu$ s). La diversité des approches disponible a d'abord motivé la réalisation d'un benchmarking des méthodes pour la prédiction des structures secondaire et tertiaire notamment afin de disposer de résultats fiables pour l'établissement d'un protocole.

## Partie 2. Prédiction de structures secondaires des oligonucléotides

### 1. Sélection des données structurales des oligonucléotides

La construction d'une librairie d'oligonucléotides dont la structure expérimentale est connue représente la première étape de ce travail. Les structures de référence extraites seront utilisées dans les essais de prédiction de structure et nous permettront d'évaluer les différences entre le modèle prédictif et les données expérimentales. Le jeu de données s'est donc construit autour des données disponibles entre Octobre et Décembre 2020.

#### 1.1.1 Extraction et tri des données

Afin de pouvoir réaliser l'étude comparative des méthodes de prédictions de la structure secondaire, les données des oligonucléotides simple brin d'ADN et d'ARN ont été extraites de la PDB et de la NDB. Sur un total de 12 736 structures d'acides nucléiques disponibles en 2020 sur la PDB, on dénombre un total de 7962 structures contenant une molécule d'ADN, et 4774 contenant une molécule d'ARN.

La plupart des structures d'ADN sont bicaténaires. Une étude comparative des méthodes de prédiction de structures disponibles nécessite des données expérimentales de référence correspondant à l'objet de l'étude. Ainsi, seuls les acides nucléiques monocaténaires ont été conservés car leur repliement ne résulte que de l'appariement des bases intra-chaînes.

Ensuite, les données récupérées ont été filtrées selon différents critères. Tout d'abord, les structures présentant des nucléotides manquants sont écartées afin de pouvoir disposer des coordonnées atomiques complètes pour la suite. Après sélection, les ADN ou ARN monocaténaires obtenus proviennent d'une structure d'oligonucléotide seul, d'un aptamère en complexe avec une protéine ou d'un aptamère en complexe avec un composé chimique. Ces derniers ont été exclus du jeu de données car la structure des oligonucléotides peut être fortement impactée par l'interaction avec le composé (Lin & Patel, 1997; Xu et al., 2019). Les oligonucléotides redondants, dont la séquence et la structure secondaire sont identiques à d'autres ont été filtrés afin de ne conserver qu'un unique représentant. Enfin, les oligonucléotides non-structurés ont également été écartés car l'intérêt de ce projet est de travailler avec des séquences ayant la capacité de former une structure secondaire et tertiaire.

L'ensemble des données conservées représente un total de 562 structures extraites de la PDB (Tableau 2).

Tableau 2. Distribution des données structurales retenues pour le jeu de données d'oligonucléotides obtenues à partir de la PDB.

Type d'oligonucléotide	Caractéristiques	Nombre de structures	Répartition par méthode de résolution	
			RMN	Rayon X
ADN	Oligonucléotide	47	47	0
	Oligonucléotide-Protéine	47	5	42
ARN	Oligonucléotide	274	267	7
	Oligonucléotide-Protéine	197	39	158
Total		565	358	207

### 1.1.2 Distribution, caractéristiques et statistiques

L'ensemble des données des ARN et ADN simples brins conservés après filtration est répertorié en Annexes 1 et 2 et un sommaire de sa composition est indiqué dans le Tableau 2. Les oligonucléotides libres représentent ~57 % des structures du jeu de données, ils sont majoritairement issus d'une résolution par spectroscopie RMN (~98 % des oligonucléotides libres), permettant de disposer d'un éventail de conformations différentes pour mieux appréhender la mobilité des oligonucléotides. A l'inverse, les oligonucléotides obtenus en complexe avec des protéines sont majoritairement des structures résolues par cristallographie aux rayons X, à l'exception de 44 structures qui sont des oligonucléotides en complexe avec de petites protéines.

L'ensemble des données regroupe des oligonucléotides de taille variée (Figure 2-1), la plupart des oligonucléotides étant d'une longueur inférieure à 50 nucléotides. Les structures de plus grande taille se distribuent entre 3 et 8 structures par classe, avec un pic de populations concernant les oligonucléotides longs de 70 à 80 nucléotides. Le jeu de données est majoritairement dominé par les ARN, avec un total de 471 structures d'oligonucléotides à ARN sur 565.

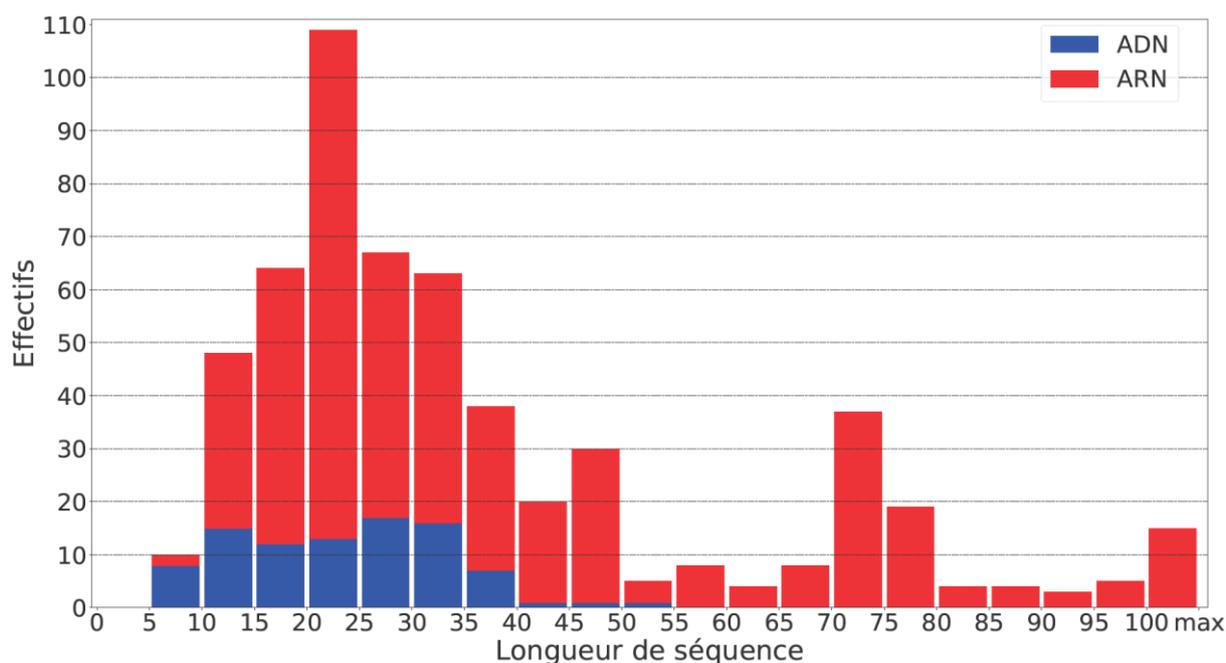


Figure 2-1. Distribution des oligonucléotides réparties dans des intervalles de longueur de séquence par incrément de 5 nucléotides.

Un total de 94 oligonucléotides ADN est contenu dans le jeu de données. La taille des oligonucléotides ADN varie de 7 à 53 nucléotides. Environ 25 % des oligonucléotides ADN sont relativement courts, avec une longueur inférieure à 15 nucléotides. Les oligonucléotides longs ( $\geq 30$  nucléotides) sont représentés dans des proportions similaires avec  $\sim 29$  % de structures. Enfin  $\sim 46$  % ont une séquence longue de 15 à 30 nucléotides.

Parmi les motifs structuraux présents dans le jeu de données filtré, on retrouve 67 oligonucléotides ADN et 396 oligonucléotides ARN disposant de structures composées d'enchaînements hélices-boucles exclusivement, 27 structures avec un motif type G-quadruplexe et 77 pseudonœuds. Les G-quadruplexes sont, dans la totalité des cas observés, des oligonucléotides à ADN, soit  $\sim 30$  % des structures à ADN. La plupart des G-quadruplexes ont été obtenus en forme libre, dans le but de caractériser ce type de motif nouvellement étudié (Kusi-Appauh et al., 2023). Ces motifs sont courts et ont été obtenus majoritairement par RMN. Seuls 4 G-quadruplexes (1HAO, 6EVV, 5CMX et 5VHE) ont été obtenus par cristallographie aux rayons X et sont extraits de complexes avec une protéine, et parmi ces structures 3 sont des aptamères dirigés contre l'alpha-thrombine humaine (1HAO, 6EVV, 5CMX). 2 autres complexes G-quadruplexe-protéine ont été obtenus en RMN (2N21 et 4I7Y). 14 séquences (7CV4, 6EVV, 2M8Z, 7CV3, 6H1K, 2M91, 5CMX, 2M90, 2M93, 7CLS, 2M92, 6ZL9, 6ZL2 et 6ZTE) caractérisent des structures plus complexes, composées de G-quadruplexes et

d'autres motifs, notamment des hélices et des épingles à cheveux. Les structures contenant des pseudonœuds, sont peu représentées parmi les oligonucléotides à ADN étudiés. Seules deux structures (5HRU et 5HTO) portent une structure arrangée en pseudonœuds de type H. Pour les structures à ARN, 75 structures présentent des interactions à longue distance, dont 17 structures présentant des pseudonœuds type H. La structure secondaire de tous les oligonucléotides retenus dans notre jeu de données a été extraite grâce à x3DNA (Lu, 2020) pour être utilisée dans les étapes suivantes.

## 2. Annotation informatique de la structure secondaire

La structure secondaire des oligonucléotides peut être représentée de plusieurs façons. Tout d'abord, la représentation qui permet une interprétation immédiate et la visualisation rapide des motifs présents est celle schématisée en Figure 2-2a. Dans ce type de représentation chaque nucléotide est indiqué en utilisant le code à une lettre et les bases appariées sont reliées par une ligne. Cette représentation permet la comparaison des structures via des métriques bien définies s'inspirant des comparaisons d'arbres phylogéniques évolutifs.

Un autre type de représentation visuelle de la structure secondaire est celle en arbre (Figure 2-2c). Dans un arbre représentatif d'une structure secondaire d'oligonucléotides, les nœuds représentent l'information liée à la structure secondaire locale (bases seules, bases appariés, motifs structuraux) et les branches la progression le long de la structure. Cette notation n'est pas privilégiée car elle reste peu accessible à la lecture pour les utilisateurs, manque de détails pour désigner la position des bases appariés et est plutôt utilisée pour l'encodage et l'utilisation informatique. Les arbres sont majoritairement utilisés dans la comparaison des structures secondaires, notamment dans RNAdistance (Hofacker, 2003). Une structure secondaire est ainsi représentée sous la forme d'un arbre enraciné qui démarre d'un nœud "racine" et qui résulte des sous-arbres qui dépendent de leur propre nœud "racine". Chaque nœud racine est représentatif d'une paire de base et chaque branche désigne une base. La représentation en arbre a ensuite été adaptée pour représenter les motifs structuraux qu'on appelle des représentations en gros grains.

Parmi les différents formats d'annotation de la structure secondaire compatibles avec une utilisation purement informatique, on retrouve le format CT, qui reporte les informations dans une table de connectivité. Cette dernière inclut les informations pour chaque nucléotide de la

séquence : le numéro d'indexage  $n$ , la base concernée (A, C, G, T, U, X), le numéro d'indexage  $n-1$ , le numéro d'indexage  $n+1$ , l'appariement avec une autre base indiqué par le numéro d'indexage de cette dernière, et la numérotation naturelle du nucléotide dans la séquence. Ce format possède l'avantage de pouvoir reporter facilement les informations essentielles pour une ou plusieurs structures secondaires. L'inconvénient de ce format est sa lisibilité par l'utilisateur, mais reste adapté à l'utilisation informatique et l'automatisation.

Le format dot-bracket présente un bon compromis entre facilité d'interprétation et manipulation informatique. Selon ce format, une structure secondaire est annotée selon un système de points et de parenthèses. Les points désignent les nucléotides de la séquence qui ne sont pas appariés et une paire de parenthèses ouverte et fermée désigne une paire de bases entre les nucléotides correspondants (Figure 2-2b). La version étendue permet également d'inclure des paires de bases croisées, qui est caractéristique des pseudonœuds. L'utilisation de paires de caractères additionnels "[ ]", "{ }", "< >" ou des caractères alphanumériques minuscules et majuscules évite les conflits avec des paires de parenthèses déjà associées. Cette notation est fréquemment utilisée comme entrée par les outils de comparaison de structures, les outils de prédiction de structure tertiaire ou ceux qui déterminent les séquences qui correspondent à une structure secondaire désignée. C'est pourquoi, au cours de ce travail, l'utilisation de cette notation a été privilégiée pour faciliter une intégration dans un pipeline.

Pendant les travaux ici présentés, un autre type de représentation des structures secondaires a été utilisé, notamment le format dotplot. Le dotplot est une matrice symétrique qui contient un point (dot) dans la case  $(x,y)$  si les nucléotides en position  $x$  et  $y$  de la séquence sont appariés, comme indiqué en Figure 2-2d. Cette représentation est utilisée notamment comme entrée pour certaines approches de comparaison de structures, comme le score F1 ou le Coefficient de corrélation de Mathews (CCM) (Chicco & Jurman, 2020; Tang et al., 2018). De plus, il peut contenir des informations supplémentaires, comme la probabilité d'appariement de 2 nucléotides en ajoutant un gradient de couleur ou de taille des points en fonction de la probabilité d'appariement.

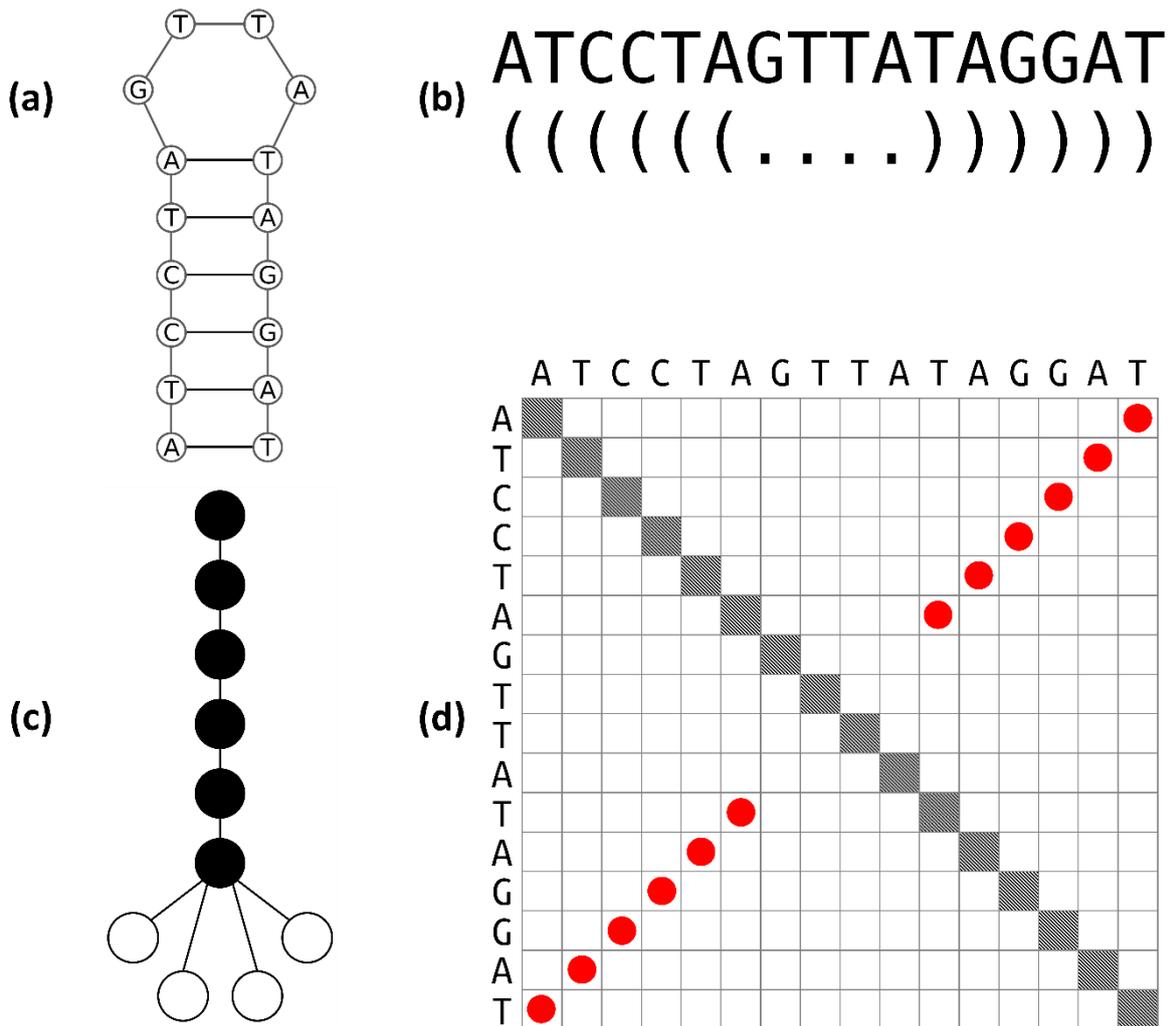


Figure 2-2. Différentes représentations de la structure secondaire avec (a) la représentation schématique dessinée avec VARNA (Darty et al., 2009) où les motifs sont facilement identifiables et les différentes représentations informatiques : (b) la représentation dot-bracket, (c) un arbre complet et (d) le format dotplot.

### 3. Méthodes *in silico* pour la structure secondaire des oligonucléotides

Bien que les bases de données aient connu un grand enrichissement du nombre de structures disponibles pour les oligonucléotides, les méthodes d'obtention des structures restent complexes, nécessitant du matériel et une bonne maîtrise de la technique. Avec le développement des approches bioinformatiques, il est devenu possible de prédire la structure secondaire *in silico* pour pallier les limitations des expériences *in vitro*. Les outils de prédiction des structures se distinguent par l'algorithme qu'ils implémentent (Tableau 3).

Tableau 3. Présentation des outils de prédiction de structures secondaires utilisés, avec le principe de fonctionnement, les avantages et les inconvénients pour chaque méthode.

<b>METHODE</b>	<b>ALGORITHME</b>	<b>AVANTAGES</b>	<b>INCONVENIENTS</b>
Mfold	Minimisation d'énergie libre	Paramètres ADN et ARN Solutions sous-optimales Température et ions pris en comptes	Aucune prédiction de pseudonœuds
RNAfold	Minimisation d'énergie libre	Paramètres ADN et ARN Solutions sous-optimales Détection des G-quadruplexes	Aucune prédiction de pseudonœuds
CentroidFold	Minimisation d'énergie libre ou apprentissage, avec estimateur de $\gamma$ -centroïde	Prédiction par approche alternative	Uniquement pour ARN
LinearFold	Minimisation d'énergie libre ou apprentissage, avec linéarisation du calcul et recherche en faisceau aléatoire	Complexité linéaire	Uniquement pour ARN
CONTRAFold	<i>Machine Learning</i>	Données d'entraînements modulables	Uniquement pour ARN
MC-fold	<i>Machine Learning</i>	Prédiction de pseudonœuds	Uniquement pour ARN Complexité élevée
MXfold2	<i>Deep Learning &amp; thermodynamique</i>	Basé sur des paramètres thermodynamiques	Uniquement pour ARN
UFold	<i>Deep Learning</i>	Prédiction de pseudonœuds	Uniquement pour ARN
SPOT-RNA	<i>Deep Learning</i>	Prédiction de pseudonœuds	Uniquement pour ARN

### 3.1. Les approches basées sur la Minimisation d'Énergie Libre (MEL)

Une même séquence a potentiellement plusieurs repliements possibles dans des mêmes conditions expérimentales, chacun ayant associé une énergie libre ( $\Delta G$ ). Plus cette énergie est faible, plus le repliement sous les conditions fixées est stable et donc plus probable selon une

distribution de Boltzmann. Le développement des connaissances sur la relation entre la structure secondaire des oligonucléotides et l'énergie libre a rendu possible la prédiction de structure secondaire basée sur l'estimation et la minimisation de l'énergie libre. Ce type d'approche prédit donc la structure secondaire des oligonucléotides en se basant sur des règles associant les appariements/non appariements à une valeur d'énergie libre. Mfold (Zuker, 2003; Zuker et al., 1991) est le premier algorithme de prédiction de structure secondaire de ce type à avoir été décrit. Le principe repose sur l'estimation du minimum d'énergie libre (MFE ou MEL) en se basant sur les règles thermodynamiques et d'estimation d'énergies pour les ARN et ADN introduits dans les travaux de Mathews et al. 1999 et SantaLucia 1998. L'algorithme génère un grand nombre de structures secondaires probables en utilisant la récurrence rendue possible grâce à la programmation dynamique de son algorithme. Les structures suggérées sont ainsi d'autres solutions explorées mais qui n'ont pas été estimées comme structure d'énergie libre minimale. Mfold permet de modifier différentes conditions, notamment la température et la concentration saline en ajustant la concentration d'ions  $\text{Na}^+$  ou  $\text{Mg}^{2+}$  qui peut impacter le repliement en solution.

RNAfold (Gruber et al., 2008) repose également sur un algorithme de prédiction de minimum d'énergie libre. Il se démarque par les paramètres d'estimation d'énergie plus récents décrits à la fois pour les ARN et les ADN (Mathews et al., 2004) et la possibilité d'inclure des paramètres d'énergie alternatifs sélectionnés par l'utilisateur. En outre, il est possible d'ajouter des contraintes pour la prédiction de structure. L'algorithme de RNAfold peut limiter la prédiction aux paires de bases Watson-Crick, et forcer ou bloquer l'apparition d'appariements sur certains nucléotides dépendant des consignes utilisateurs. La prédiction peut également être guidée par des paramètres de forme, qui ajoutent comme critère de prédiction la présence de motifs structuraux stabilisant (hélices, boucles). Avec ce paramètre, le choix de la structure optimale accorde une importance moindre aux critères thermodynamiques et privilégie la présence de motifs structuraux (Deigan et al., 2009). Si indiqué, RNAfold peut également travailler avec des séquences à caractère cyclique, ou rechercher d'éventuels G-quadruplexes au sein de la séquence.

Etant basées sur l'estimation de l'énergie libre minimale, les deux approches incluent la possibilité de générer les structures secondaires dites sous-optimales, qui présentent un  $\Delta G$  moins favorable par rapport à la structure d'énergie libre minimale.

### 3.2. Les approches *Knowledge based* ou *basées sur l'Intelligence Artificielle*

Les deux dernières décennies ont connu l'émergence de méthodes alternatives à l'approche de minimisation de l'énergie libre appliquée à la prédiction de structures secondaires. L'enrichissement des bases de données des structures a permis le développement de techniques avancées de prédiction qui utilisent ces connaissances. L'inconvénient majeur de ces techniques provient de leur capacité à prédire exclusivement les structures secondaires des ARN : aucun d'eux n'inclut de paramètres différenciés pour les ADN en raison du manque de données disponibles en apprentissage pour ce type d'acide nucléique. Il est alors admis que ces modèles sont avant tout appropriés pour les ARN. Parmi les outils disponibles, seuls ceux qui sont en libre accès et utilisables localement sur une station de travail, avec un système opérationnel de type Unix, ont été sélectionnés.

#### 3.2.1 *Machine Learning*

L'accumulation de données structurales sur les acides nucléiques a permis l'émergence des méthodes basées sur le *Machine Learning* (ML) pour la prédiction des structures secondaires. CONTRAfold (Do et al., 2006) fait partie des premiers algorithmes appartenant à cette catégorie d'approche. L'algorithme repose sur un modèle probabiliste dépendant d'une grammaire algébrique qui lui est propre, ce qui le définit comme un algorithme log-linéaire conditionnel. Le choix de la structure secondaire appropriée se base ensuite sur un schéma simplifié d'évaluation de l'énergie.

L'algorithme de MC-Fold (Parisien & Major, 2008) se base également sur les données connues des structures secondaires pour effectuer les prédictions. MC-Fold utilise une librairie de motifs structuraux extraits d'ARN de la PDB pour la prédiction de la structure secondaire. Cette librairie répertorie toutes les associations de nucléotides possibles et génère la structure secondaire par association récursive de fragments. Plusieurs solutions sous-optimales peuvent être générées. Ces solutions sont classées selon un modèle de pseudo-estimation d'énergie potentielle combiné à la probabilité de formation d'un motif dans la séquence. Grâce à cette approche, il est possible pour MC-Fold de prédire la position de pseudonœuds de type H en utilisant le paramètre dédié "*pseudoknotted*".

### 3.2.2 Algorithmes hybrides

Certains algorithmes intègrent les approches d'apprentissage automatique et celles d'estimation du minimum énergétique, mais se démarquent plutôt par leur mode de décision pour le choix de la structure secondaire finale. CentroidFold (Sato et al., 2009) se base sur les modèles de Vienna RNAfold McCaskill (critère thermodynamique) ou CONTRAfold (critère probabilistique), au choix, pour construire la structure secondaire d'une séquence. La prédiction la plus appropriée est déterminée par un estimateur de type  $\gamma$ -centroïde comme alternative aux estimateurs du minimum d'énergie libre ou *Maximum Expected Accuracy*. LinearFold (L. Huang et al., 2019) est également couplé aux modèles de Vienna RNAfold McCaskill ou CONTRAfold pour la prédiction de structure secondaire. Son algorithme repose sur la linéarisation du temps de calcul en i) parcourant progressivement la séquence de l'extrémité 5' à 3' et marquant chaque nucléotide pour un potentiel appariement ii) réduisant l'espace de recherche grâce à une recherche heuristique en faisceau (*beam search*). Ceci permet d'atteindre une complexité d'algorithme de l'ordre de  $O(n)$  valorisable sur la prédiction de structure de grande taille.

### 3.2.3 Deep Learning

Les méthodes d'apprentissage avancé ont permis de développer des nouveaux modèles qui bénéficient des connaissances en matière de structure secondaire et des données de structures disponibles. En conséquence, plusieurs algorithmes sont construits sur des approches "*Deep Learning*". Parmi eux, SPOT-RNA (Singh et al., 2019), Ufold (Fu et al., 2022), et MXfold2 (Sato et al., 2021) sont trois outils qui implémentent des algorithmes intégrant des réseaux de neurones et présentent donc une structure de fonctionnement similaire dans leur façon d'apprendre à partir des données disponibles. Ils se distinguent par l'architecture de leur réseau et notamment par les données structurales utilisées pour entraîner et tester leur modèle de prédiction.

Le modèle de prédiction de SPOT-RNA se base sur une partie des informations contenues de la base de données bpRNA-1m (Danaee et al., 2018), soit 13 419 séquences ARN et leur structure secondaire associée. En complément, 226 structures de la PDB ont été intégrées pour l'apprentissage sur des données expérimentales tridimensionnelles.

UFold dispose du plus imposant jeu de données, provenant de plusieurs bases de données de structures secondaires : RNAStralign (Tan et al., 2017), Archive II (Sloma & Mathews, 2016), bpRNA-1m (Danaee et al., 2018), bpRNA-new dérivées des données de Rfam 14.2 (Kalvari et al., 2021; Sato et al., 2021), et les données de la PDB, compilant un total d'environ 140 000 séquences d'ARN simple brin divisées en différents jeux de données d'entraînement ou de test.

MXfold2 (Sato et al., 2021) est une amélioration de MXfold (Akiyama et al., 2018) qui est un algorithme basé sur l'apprentissage automatique. Le modèle a été entraîné sur des données de structure secondaire obtenues de la littérature ainsi que de la base de données Rfam. La fonction de décision de la meilleure structure se base sur des critères thermodynamiques équivalents à ceux utilisés dans les algorithmes de prédiction par minimisation de l'énergie libre.

Ces différents algorithmes de prédiction ont été développés pour enrichir la compréhension de la structure secondaire des oligonucléotides, chacun d'eux ayant apporté son lot d'innovations ou d'améliorations. Un grand nombre d'outils de prédiction est aujourd'hui disponible mais cette diversité engendre une nouvelle complexité dans le choix du meilleur outil. Comparer les performances de ces différentes approches semble donc pertinent.

#### 4. Comparaison de structures secondaires

Chaque outil de prédiction de structure secondaire dispose de son propre algorithme et critère de prise de décision. Par conséquent, il n'est pas rare de rencontrer des différences de prédiction entre deux outils. En général, les outils sont choisis selon des critères de fiabilité ou de pertinence face à une situation particulière, comme la prédiction de pseudonœuds, de structures longues, ou de G-quadruplexes. Pour les besoins du projet, nous avons essayé de déterminer les performances des outils sélectionnés pour reproduire les structures secondaires expérimentales des oligonucléotides de notre banque. Pour ce faire, un jeu de données expérimentales et une méthode de comparaison entre la structure secondaire prédite et celle expérimentale sont nécessaires. Le jeu de données utilisé a été précédemment décrit (Partie 2.1). Pour ce qui concerne la méthode de comparaison des structures secondaires, une nouvelle métrique, AptaMat (Binet et al., 2023), a été développée pour pallier les limites des méthodes de comparaison existantes.

#### 4.1. Comparaison de chaînes de caractères

Plusieurs méthodes existent pour comparer de façon quantitative les structures secondaires des oligonucléotides. Parmi celles-ci, les méthodes de comparaison des chaînes de caractères sont très utilisées. Par exemple, le coefficient de Tanimoto (Chung et al., 2019) peut être utilisé dans de multiples contextes impliquant un codage de l'information par chaîne de caractères, comme par exemple la notation *dot-bracket* pour la structure secondaire des oligonucléotides. Le principe repose sur l'identité de position des caractères entre deux chaînes de même longueur. En partant de deux structures secondaires A et B de même longueur  $L$  représentées en utilisant la notation dot-bracket,  $S_A$  et  $S_B$ , le coefficient de Tanimoto est défini dans l'équation 1 :

$$C_{Tanimoto}(S_A, S_B) = N_{sim}/L \text{ (Équation 1)}$$

où  $N_{sim}$  représente le nombre de caractères identiques sur une même position entre  $S_A$  et  $S_B$ .

La distance de Hamming est une alternative au coefficient de Tanimoto, qui, au contraire de ce dernier, comptabilise le nombre de caractères qui varient  $N_{diff}$  entre  $S_A$  et  $S_B$  selon l'équation 2 :

$$D_{Hamming}(S_A, S_B) = N_{diff}/L \text{ (Équation 2)}$$

où  $N_{diff}$  représente le nombre de caractères qui varient entre les deux dot-bracket  $S_A$  et  $S_B$  et  $L$  la longueur de la chaîne de caractère ou du dot-bracket.

Les deux approches donnent des valeurs entre 0 et 1, mais leur interprétation est opposée. Pour le coefficient de Tanimoto, une valeur proche de 1 signifie une identité totale de  $S_A$  et  $S_B$ . À l'inverse pour la distance de Hamming, plus les valeurs sont élevées moins il y a de ressemblance. Ainsi, une valeur proche de 1 symbolisera une absence totale d'identité. Malgré leur simplicité d'application, ces approches présentent plusieurs défauts i) la structure secondaire n'est pas analysée dans son ensemble, ii) La position des paires de bases n'est pas prise en considération, iii) les différences de caractères sur une position est fortement pénalisée, et iv) la comparaison se limite à des chaînes de caractères de même longueur.

## 4.2. Classificateurs binaires

Les classificateurs binaires utilisent la relation existante entre le nombre de vrais positifs (TP), faux positifs (FP), vrais négatifs (TN) et faux négatifs (FN) afin d'estimer la véracité d'un modèle de prédiction. Cette façon de procéder peut ainsi être appliquée pour la comparaison de descripteurs représentés de manière binaire comme les descripteurs moléculaires, ou également la structure secondaire comme appliqué dans les travaux sur Ufold (Fu et al., 2022; Ji et al., 2020). Dans ce cas, les TP correspondent aux paires de bases correctement prédites, les FP au nombre de paires de bases prédites n'ayant pas de correspondance dans la structure de référence et les FN correspondent aux paires de bases présentes dans la structure de référence et pas prédites.

Parmi les classificateurs binaires, le score F1 et le CCM sont fréquemment utilisés dans la comparaison de structures secondaires. Ils sont définis selon les équations 3 et 4:

$$F1score = \frac{2TP}{2TP+FP+FN} \text{ (Équation 3)}$$

$$CCM = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \text{ (Équation 4)}$$

Le score F1 peut prendre des valeurs entre 0 et 1. Plus on s'approche de 0, moins il y a de ressemblance entre la référence et la prédiction. Les valeurs de CCM se distribuent entre -1 et 1. Des valeurs qui tendent vers le négatif supposent donc que le modèle prédit moins bien que l'aléatoire et donc que la prédiction est mauvaise.

## 4.3. Distances entre structures secondaires

Pour améliorer la sensibilité de la comparaison et permettre une adaptation plus fine aux oligonucléotides, des mesures de distance spécialement conçues pour la comparaison de structures secondaires oligonucléotidiques ont été développées. Le principe le plus simple est appliqué dans la distance Base Pair (BP) introduite par Zuker (1989). Pour deux structures  $S_1, S_2 \in S_n$ , on définit la distance entre deux paires de bases  $b_1 = i \cdot j \in S_1$  et  $b_2 = i' \cdot j' \in S_2$  comme :

$$d_0(i \cdot j, i' \cdot j') := \max\{|i - i'|, |j - j'|\} \text{ (Équation 5)}$$

L'ensemble des distances  $d$  entre les paires de bases  $b_1$  et  $b_2$  est calculé. La distance  $D_{S_1.S_2}$  entre  $S_1$  et  $S_2$  correspond à la distance minimale parmi toutes les distances entre les paires de bases  $i \cdot j \in S_1$  et  $i' \cdot j' \in S_2$ .

Cette formule de distance a ensuite servi de base à des améliorations car elle présentait des défauts de sensibilité majeurs. Le premier problème rencontré étant le manque de sensibilité aux changements mineurs de la structure, une première correction a été apportée deux ans plus tard (Zuker et al., 1991) et intègre un paramètre de relaxation. On définit donc l'ensemble de distance entre paires de bases  $b_1 \in S_1$  et  $b_2 \in S_2$  si, pour toute paire de bases  $b_1$  en excluant celles de distances  $d$ , il existe une paire  $b_2$  distante de  $d$  au plus par rapport à  $b_1$ . En prenant en compte le calcul symétrique de la distance  $D_{S_1.S_2}$  représente la distance maximale entre les deux distances possibles entre  $S_1, S_2$  ou  $S_2, S_1$ . Cette méthode de calcul affinée a été intégrée à mfold comme paramètre de génération des structures sous-optimales et reste une mesure fiable de la distance.

Le *Relaxed Base Pair score* (RBP) (Agius et al., 2010) s'inspire de la distance BP puisqu'il en utilise la distance  $d$  entre les paires de bases. Il inclut un paramètre de relaxation  $t \in \mathbb{R}^+$  ajustable. Ainsi, pour deux structures  $S_1$  et  $S_2$  et l'ensemble des distances  $b_1 \in S_1$  et  $b_2 \in S_2$ , celles-ci sont assemblées dans un nouvel ensemble  $b_M \in S_1, S_2$ . Les distances sont classées par ordre décroissant (sans exclure les distances redondantes) ce qui donne  $\{\Delta_1, \Delta_2, \dots, \Delta_{M_1+M_2}\}$  où  $\Delta_i \geq \Delta_j$  lorsque  $i < j$ , et le score RBS est calculé selon l'équation 6:

$$RBP(S_1, S_2) = \min \{ m \in \mathbb{Z} \mid m \geq 0, \Delta_k \leq tm \text{ si } k > m \} \text{ (Équation 6)}$$

En résumé, la RBP est l'entier naturel minimal  $m$  pour lequel la distance  $\Delta_k$  est inférieure ou égale à cet entier multiplié par le paramètre de relaxation  $t$  et si l'index dans  $\{\Delta_1, \Delta_2, \dots, \Delta_{M_1+M_2}\}$  est supérieur à  $m$ . La valeur de RBP est influencée par la distance BP mais reste une approche différente. Lorsque  $t = 0$ , RBP = BP et lorsque  $t \rightarrow \infty$ , la valeur de RBP varie entre 0 (les structures sont identiques) et 1 (les structures sont opposées). Cette approche, contrairement à la distance BP, ne se base pas sur le maximum entre paires de bases grâce au paramètre de relaxation  $t$ . L'inconvénient majeur est que le paramètre de relaxation  $t$  est à définir pour chaque comparaison. Pour une même valeur  $t$  les résultats ne sont pas comparable d'une étude à l'autre avec des différences fortement dépendantes de la longueur des structures étudiées et des paires de bases impliquées. Le paramètre  $t$  doit être

décidé et ajusté selon le contexte de l'étude où plusieurs valeurs  $t$  doivent être étudiées pour des résultats complets, ce qui complexifie l'analyse des résultats.

#### 4.4. Distance d'édition des arbres

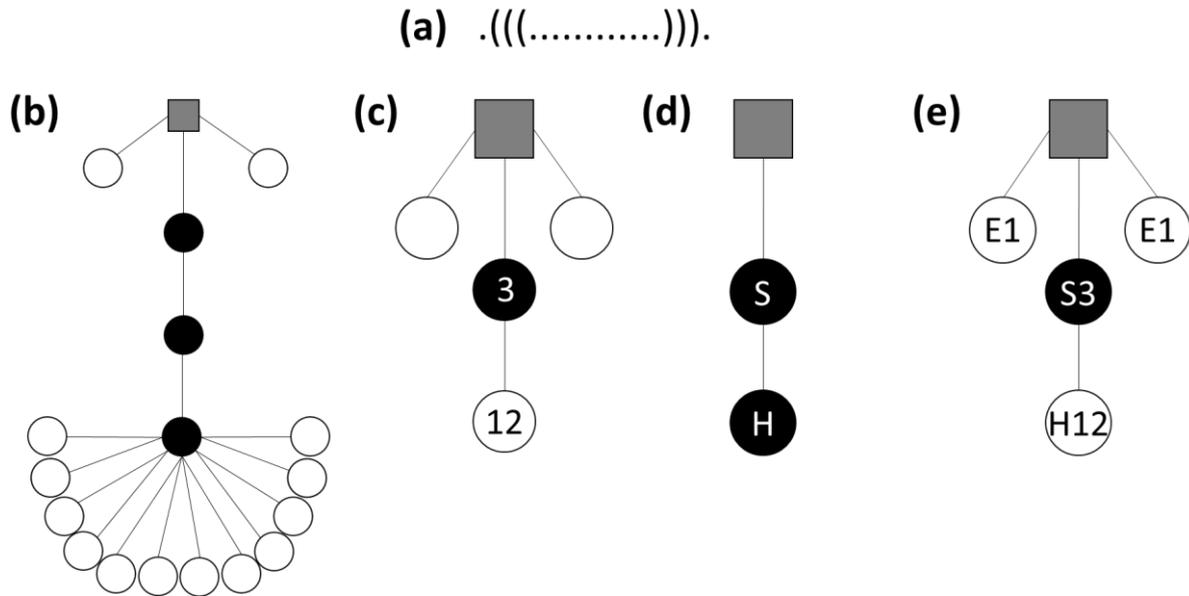


Figure 2-3. Représentation schématique des arbres associés à la structure secondaire (a) représentée en dot-bracket. 4 types d'arbres peuvent être étudiés avec RNAdistance soit (b) un arbre complet, (c) un arbre HIT, (d) un arbre gros grain et (e) un arbre gros grain pondéré.

RNAdistance est la méthode la plus versatile et la plus utilisée pour mesurer une distance entre deux structures secondaires car elle utilise la représentation en arbre de la structure secondaire. La comparaison de deux arbres se résume à calculer le nombre minimum d'opérations d'édition (insertions, suppressions ou renommages de nœuds) nécessaires pour passer de l'arbre représentatif d'une structure secondaire  $S_1$  à celui correspondant à une structure  $S_2$ . RNAdistance propose l'utilisation de 4 représentations d'arbres (Figure 2-3) qui sont adaptées à différents types d'analyse. La valeur de cette distance dépendra alors du type d'arbre utilisé en entrée. La représentation standard est la représentation en arbre complet introduit dans la Partie 2.2. La représentation HIT (Homeomorphically Irreducible Tree) groupe les bases en fonction de leur statut « appairé » ou « non appairé » et donne des arbres de taille réduite. Ensuite, les représentations « gros grains » permettent d'annoter les nœuds par type de motif rencontré. Le poids d'un changement de motif dans l'arbre varie en fonction de la matrice de coût utilisée pendant l'opération. L'inconvénient de cette représentation est la perte d'information occasionnée car il est difficile d'identifier les bases directement

impliquées. Pour pallier ce problème, chaque motif peut alors être accompagné d'un poids correspondant au nombre de bases qui forme le motif.

Ainsi RNAdistance propose plusieurs combinaisons de représentation et mode de calcul de distance détaillées dans le Tableau 4. L'utilisation du calcul par distance d'édition d'arbres représente l'intérêt premier de RNAdistance. Il est cependant possible de recourir à la comparaison de chaînes de caractères.

Tableau 4. Méthodes de calculs proposés par RNAdistance, utilisable via le paramètre -D et la lettre associée dans Méthode.

Méthode	Représentation	Calcul
f	Arbre complets	Distance d'édition d'arbre
h	HIT	Distance d'édition d'arbre
w	Gros grain pondérée	Distance d'édition d'arbre
c	Gros grain	Distance d'édition d'arbre
F	Arbre complets	Comparaison de chaînes de caractères
H	HIT	Comparaison de chaînes de caractères
W	Gros grain pondérée	Comparaison de chaînes de caractères
C	Gros grain	Comparaison de chaînes de caractères

#### 4.5. Autres approches

D'autres approches ont vu le jour avec pour objectif de pallier les problèmes de sensibilité. DoPloCompare se base sur le traitement d'images appliqué aux dotplot (Ivry et al., 2009). On définit donc deux structures  $A$  et  $B$  selon leur représentation dotplot et l'ensemble des points de  $A = (a_{ij})$  et  $B = (b_{ij})$ . La distance DoPloCompare entre  $A$  et  $B$ ,  $D(A, B)$  est dépendante de deux facteurs détaillés dans l'équation 7 :

$$D(A, B) = \frac{Dist(A, B)}{Corr(A, B)} \text{ (Équation 7)}$$

La composante  $Dist(A, B)$  désigne la distance quadratique moyenne entre les deux ensembles de points de  $A$  et  $B$  soit l'équation 8 :

$$RMSDist(A, B) = \sqrt{\frac{1}{n} \sum_{a \in A} \|a - N_B(a)\|^2} \text{ (Équation 8)}$$

avec  $n$  le nombre de points de  $A$  et  $N_B(a)$  le plus proche voisin de  $a$  dans  $B$ . En outre, la localisation des plus proches voisins utilise les diagrammes de Voronoï pour accélérer le temps de calcul.

La composante de corrélation  $Corr(A, B)$  est utilisée pour normaliser la distance. Cela correspond à la mesure de la corrélation entre les histogrammes représentatifs de la somme des points des matrices  $A$  et  $B$  sur les axes  $X, Y$  avant et après diagonalisation selon l'équation 9 :

$$Corr(A, B) = \sqrt{Xc(A, B) \times Yc(A, B) \times Dc(A, B) \times Ic(A, B)} \text{ (Équation 9)}$$

Avec  $Xc(A, B)$ ,  $Yc(A, B)$ ,  $Dc(A, B)$  et  $Ic(A, B)$  les valeurs de corrélation croisée pour les vecteurs somme des colonnes  $X$ , des lignes  $Y$ , des diagonales Sud-Ouest Nord-Est et diagonales Sud-Est Nord-Ouest. DoPloCompare se distingue des autres distances par sa normalisation qui facilite l'analyse des résultats. L'inconvénient de cette distance est son absence de symétrie car la mesure ne se fait que dans une direction et varie selon la référence choisie. Deux structures  $A, B$  dont la distance est calculée dans le sens  $A$  contre  $B$  ou  $B$  contre  $A$  peuvent ainsi générer des distances différentes. Cependant, cette caractéristique peut être utilisée dans certains contextes, notamment dans l'étude d'impact de mutations lorsque la structure de référence est clairement identifiée.

#### 4.6. AptaMat : un outil de comparaison de structures secondaires efficace

L'inconvénient de la plupart des méthodes de comparaison abordées provient de problèmes de sensibilité entraînant une incapacité à identifier les différences entre des structures secondaires très proches. L'exemple-jouet présenté en Figure 2-4, adapté de l'article de Ivry et al. 2009, montre les distances mesurées avec différentes méthodes pour les 3 structures (b), (c) et (d) face à la structure de référence (a). Ces trois structures alternatives présentent un bourgeon qui devient plus large entre (b) et (d) à cause de la variation de la troisième paire de bases qui se déplace progressivement vers l'extrémité 5'. Les distances proposées par la distance de Hamming, RNAdistance "f", distance BP, le score RBP, le score F1 ou CCM suggèrent que les trois structures (b), (c) et (d) présentent les mêmes différences face à la structure de référence malgré les variations identifiables. Pour pallier les défauts des autres algorithmes, nous avons développé notre propre algorithme de comparaison de structure

secondaire, AptaMat (Binet et al., 2023), qui se veut plus sensible aux changements de structures secondaires, versatile, et simple en termes d'algorithme et d'utilisation avec peu de paramètres à ajuster. Toutes les variations de paires de bases sont prises en considération dans le calcul, ce qui rend en principe possible la comparaison de structures contenant des pseudonœuds ou des G-quadruplexes. De plus, combiné à un algorithme d'alignement, AptaMat peut comparer des structures de longueurs différentes grâce à la pénalité de gap incluse dans l'algorithme. L'impact de ces changements de structures est estimé grâce à une simple mesure de la distance entre les points de deux matrices.

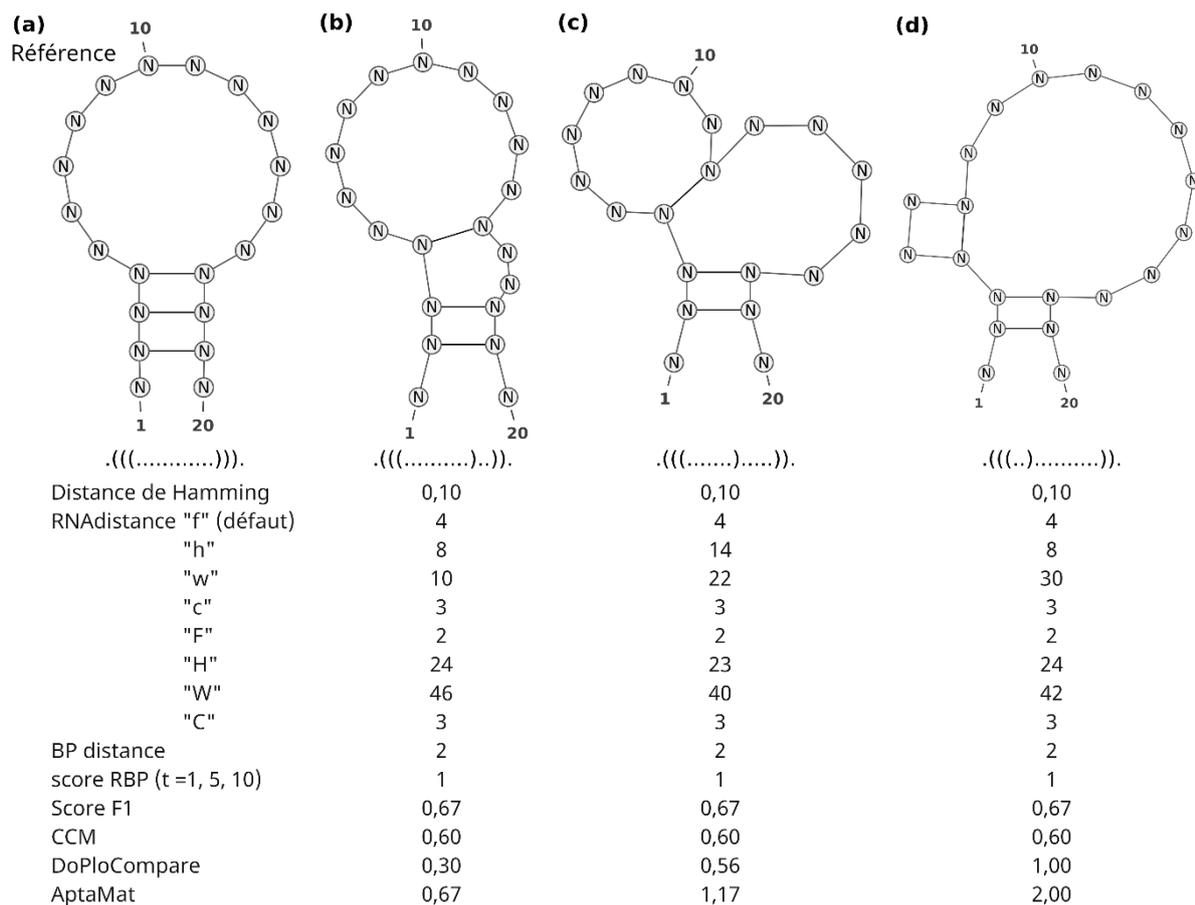


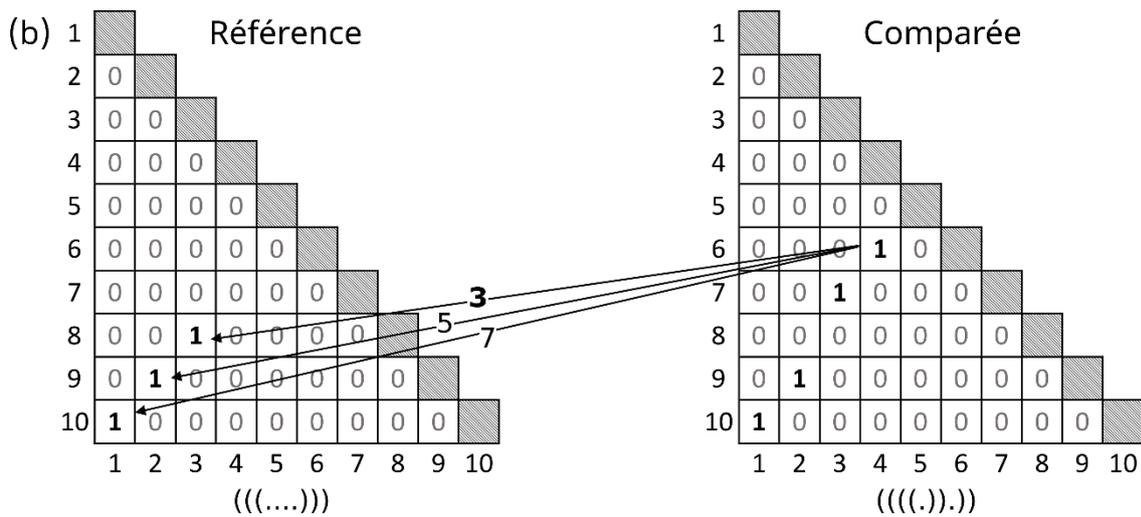
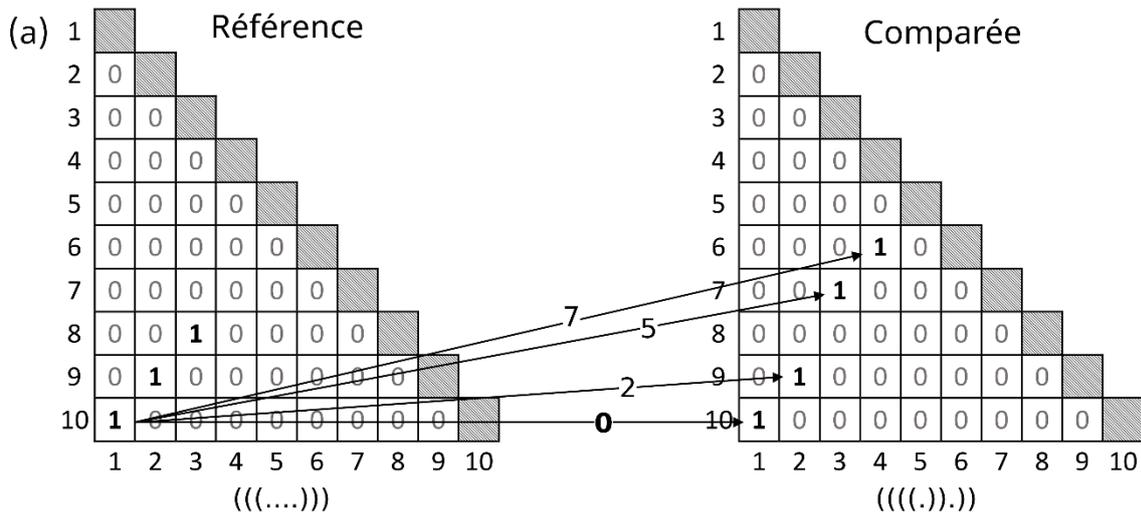
Figure 2-4. Comparaison d'une structure « épingle à cheveux » (a) et 3 structures alternatives (b, c et d) en utilisant plusieurs méthodes de calculs. La distance de Hamming, RNAdistance et ses différentes approches de calculs, BP, le score RBP (avec  $t = 1, 5$  et  $10$ ), le score F1, le Coefficient de corrélation de Mathews (CCM), DoPloCompare et AptaMat (structures dessinées avec VARNA).

#### 4.6.1 Présentation de l'algorithme

AptaMat se base sur une représentation matricielle de la structure secondaire similaire à DoPloCompare. AptaMat prend donc en entrée deux structures secondaires d'oligonucléotides alignées, l'alignement étant de longueur  $L$ , au format dot-bracket. L'alignement peut être effectué préalablement au programme en faisant appel à un outil d'alignement de structures, comme RNAAlign2D ou il peut être fourni directement par l'utilisateur. A partir des séquences  $S_A$  et  $S_B$  en notation dot-bracket, deux matrices,  $A = (a_{ij})$  et  $B = (b_{ij})$ , de taille  $L \times L$  sont créées, dans lesquelles chaque case  $(i,j)$  correspond à la position  $i$  d'un nucléotide de la séquence par rapport à la position  $j$  sur la même séquence. Chaque case  $(i,j)$  contient soit 1, si le nucléotide en position  $i$  est apparié au nucléotide en position  $j$ , soit 0 si les deux nucléotides ne sont pas appariés. Ensuite, pour pouvoir définir une distance entre la matrice  $A$  et la matrice  $B$  qui puisse décrire la proximité entre les cases contenant des 1 dans les deux matrices, chaque matrice est placée dans le plan de façon que chaque élément  $(i,j)^{\text{ème}}$  égale à 1 soit assimilé au point de coordonnées  $(j, L - i + 1)$ . Cela implique qu'à une matrice représentant une structure secondaire est associé un ensemble de points dans le plan, dont les coordonnées sont dans  $\{1, \dots, L\}^2$ . On note  $\mathcal{P}_A := \{(j, L - i + 1) \in \mathbb{N} : a_{ij} = 1, 1 \leq j < i \leq L\}$  et  $\mathcal{P}_B$ , défini de façon analogue, l'ensemble des points de la matrice  $A$  ou  $B$  dont les cases contenant 1 et qui donc indiquent la présence d'une paire de base entre les positions  $i$  et  $j$  dans la structure secondaire  $S_A$  ou  $S_B$ . AptaMat calcule la distance entre  $S_A$  et  $S_B$  en mesurant la distance entre les ensembles  $\mathcal{P}_A$  et  $\mathcal{P}_B$ . Pour ce faire, toute distance entre deux ensembles de points dans  $\mathbb{R}^2$  est adaptée, mais pour l'instant AptaMat se sert de la distance de Manhattan, définie en équation 10, où  $P_A$  et  $P_B$  sont deux points appartenant à  $\mathcal{P}_A$  et  $\mathcal{P}_B$ , respectivement, de coordonnées  $(i_{P_A}, j_{P_A})$  et  $(i_{P_B}, j_{P_B})$  :

$$d_{Man}(A, B) = |i_B - i_A| + |j_B - j_A| \text{ (Équation 10)}$$

Donc, pour chaque point  $P$  de  $\mathcal{P}_A$  on détermine la distance de Manhattan de son plus proche voisin dans  $\mathcal{P}_B$ , et vice-versa. Le calcul des distances est répété entre  $\mathcal{P}_B$  et  $\mathcal{P}_A$  pour obtenir toutes les distances dans les deux directions et identifier toutes les différences existantes entre les structures, car le nombre de paires de bases peut varier entre les deux structures.



(c)

		Référence			
		1-10	2-9	3-8	Minimum
Comparée	1-10	0	2	4	0
	2-9	2	0	2	0
	3-7	5	3	1	1
	4-6	7	5	3	3
	Minimum	0	0	1	

$$Apta_D = \frac{\sum_{P \in P_{référence}} d_{Man}(Référence, Comparée) + \sum_{P \in P_{comparée}} d_{Man}(Comparée, Référence)}{\#P_{référence} + \#P_{comparée}}$$

$$Apta_D = \frac{(0 + 0 + 1) + (0 + 0 + 1 + 3)}{7} = 0.714$$

Figure 2-5. Exemple appliqué d'AptaMat sur deux structures hypothétiques avec l'acquisition des distances de Manhattan dans les deux directions (a) et (b). L'ensemble des distances obtenues est détaillé en (c) et permet d'extraire les distances minimales pour chaque paire de points puis le calcul de  $Apta_D$  permet d'obtenir 0,714.

Par conséquent, les distances dans les deux directions peuvent différer et certaines paires de bases risquent d'être exclues de la comparaison.

Ensuite, les plus courtes distances entre les ensembles  $\mathcal{P}_A$  et  $\mathcal{P}_B$  sont additionnées. Dans le cas où l'on travaille à la comparaison d'oligonucléotides de longueur différente, AptaMat prend en compte la présence de gaps dans un alignement et leur attribue une pénalité de 1 pour chaque gap, car la présence de gaps augmente la distance entre deux points. Enfin, la distance obtenue est normalisée par le nombre de points des matrices  $A$  et  $B$ , égal au nombre de paires de bases des structures  $S_A$  et  $S_B$ , car certaines distances pourraient apparaître deux fois dans le calcul. Il est important de souligner que ce type de normalisation donne un poids plus important aux paires de bases en commun entre les deux structures comparées. On définit donc la distance AptaMat ( $Apta_D$ ) entre deux structures secondaires alignées  $S_A$  et  $S_B$  comme dans l'équation 11 :

$$Apta_D(S_A, S_B) = \frac{\sum_{P \in \mathcal{P}_A} d_{\text{Man}}(P, \mathcal{P}_B) + \sum_{P \in \mathcal{P}_B} d_{\text{Man}}(P, \mathcal{P}_A) + N_G}{\#\mathcal{P}_A + \#\mathcal{P}_B} \quad (\text{Équation 11})$$

Où pour chaque point  $P = (x, y) \in \mathbb{R}^2$  et pour tout sous-ensemble fini,  $\mathcal{C} \subset \mathbb{R}^2$ , on désigne par  $\#\mathcal{C}$  le cardinal de  $\mathcal{C}$ , et par  $d_{\text{Man}}(P, \mathcal{C})$  la distance de Manhattan d'un point  $P$  vers son plus proche voisin dans  $\mathcal{C}$ .  $N_G$  représente le nombre de gap dans l'alignement.

On peut ainsi voir que  $Apta_D$  est symétrique, qu'elle est égale à 0 si les deux structures comparées sont identiques et que, par conséquent, plus  $Apta_D$  est proche de 0, plus les structures comparées se ressemblent, indépendamment de leur longueur. Le schéma d'application d'AptaMat est représenté en Figure 2-5.

Enfin, les oligonucléotides sont des molécules très mobiles qui peuvent adopter une grande variété de conformations impliquant des variations de la structure secondaire, nécessitant parfois de travailler avec des ensembles de structures plutôt qu'une seule (Ganser et al., 2019). Pour cela, AptaMat est capable de travailler sur des ensembles de structures secondaires représentatives obtenues d'études expérimentales ou de données prédictives ayant des informations concernant la probabilité ou fréquence d'apparition de chaque structure secondaire de l'ensemble. Ainsi, grâce à son option "*-ensemble*", AptaMat prend en entrée un ensemble de  $n$  structures secondaires à comparer  $(B_i)_{i=1}^n$  et leur poids associé  $(w_i)_{i=1}^n$ ,

dérivés des données expérimentales ou prédites et la distance AptaMat pondérée est calculée comme l'équation 12 suivante :

$$D_{AM} \left( S_A, (S_{B_i})_{i=1}^n \right) = \sum_{i=1}^n w_i D_{AM}(S_A, S_{B_i}) \text{ (Équation 12)}$$

L'algorithme d'AptaMat a été développé sur Python 3.8 et le code est en libre accès sur Github<sup>4</sup>.

#### 4.6.2 Comparaison des d'AptaMat face aux autres distances

AptaMat offre une meilleure sensibilité par rapport aux autres méthodes utilisées communément dans la comparaison de structures secondaires d'acides nucléiques. Tout d'abord, AptaMat n'est pas basé sur une classification binaire des différences comme les comparaisons de chaînes de caractères, le score F1 ou le (CCM) qui peuvent entraîner une surestimation des différences. De plus, par rapport à RNAdistance, AptaMat mesure les changements en passant directement par la distance entre paires de bases, là où la distance d'édition d'arbres mesure indirectement les différences. La représentation matricielle permet de prendre en compte la structure secondaire dans sa globalité, puisqu'on compare la présence ou l'absence de paires de bases par la position des points dans la matrice. De plus, la distance de Manhattan utilisée dans AptaMat permet d'observer les variations des positions sur deux coordonnées là où la distance BP ne reporte la distance maximale que pour une seule coordonnée. En plus de ces observations, les performances d'AptaMat ont été comparées à celles des autres métriques pour la comparaison des structures secondaires mentionnées, notamment la distance de Hamming, RNAdistance, la distance BP, le score RBP, DoPloCompare, le score F1 et le CCM.

Tout d'abord, l'exemple-jouet repris de Ivry et al. (2009) a été utilisé (Figure 2-4). Les 3 structures alternatives (b), (c) et (d) proposées diffèrent de la référence (a) pour une unique paire de bases localisée à différentes positions, qui entraîne des changements de structure secondaire. En effet, la structure (a) de référence se compose d'une simple épingle à cheveux.

---

<sup>4</sup> <https://github.com/GEC-git/AptaMat>

Les structures alternatives (b), (c) et (d) présentent un bourgeon de plus en plus ample en passant de la structure (b) à la structure (d), car la troisième paire de bases se déplace progressivement vers l'extrémité 5', donnant pour la structure (d) un appariement improbable.

La comparaison des différents scores et distances montre une tendance à associer la même valeur aux 3 structures. Ainsi, l'absence de différence suggérée par la distance de Hamming, RNAdistance en paramètre « f », « c », « F » et « C », BP, le score RBP, le score F1 et le CCM semble indiquer que les appariements différents n'entraîneraient pas de changement notable de la structure secondaire, ce qui est incorrect. RNAdistance « h » et RNAdistance « H » ne semblent distinguer de différence que dans un cas sur les 3, mais l'ordre de classement ne coïncide pas. RNAdistance « h » classifie ainsi la structure (b) comme la plus proche à égalité avec la structure (d), pourtant très différente, reléguant la structure (c) comme la solution avec la plus forte distance. Pour RNAdistance « H » la structure (c) est identifiée comme la plus proche alors que (b) et (d) sont estimées à égale distance de la référence (a). Ces observations permettent ainsi de souligner la diversité de conclusions possibles en fonction du choix de paramètre pour RNAdistance notamment. Seuls RNAdistance « w », RNAdistance « W », DoPloCompare et AptaMat parviennent à différencier ces structures. Le classement (b) < (c) < (d) partagé entre DoPloCompare, AptaMat et RNAdistance « w » est cohérent avec les variations observées de la structure secondaire avec pour référence (a). En revanche, le classement proposé par RNAdistance « W » accorde une plus grande proximité de la structure (c), et classe incorrectement (b) comme la structure avec le plus de différences.

Le second exemple (Figure 2-6) est un aptamère d'ADN extrait de la structure PDB 5HRU qui se lie avec l'enzyme lactate déshydrogénase (LDH) de *Plasmodium vivax* par interaction des nucléotides T5 à T13 du bourgeon. Cet exemple témoigne de l'importance de ce motif pour la fonction souhaitée et pourtant les structures alternatives proposées (b), (d) et (e) ont tendance à perdre ce motif. Cet exemple montre que la plupart des outils de comparaison ne parviennent pas à différencier toutes les structures proposées bien qu'ils soient tous capables d'identifier la structure (c) comme la plus proche de la référence car c'est la seule qui ne présente qu'une unique perte d'appariement en T1-A32. Seuls RNAdistance « w » et AptaMat sont capables de différencier les 4 structures et partagent le même classement. La structure (e) s'approche de la référence en raison de la conservation d'une partie du bourgeon entre T10-T13. La structure (d) forme 2 appariements T11-G31 et G12-C30 non présents chez la

référence mais permet de maintenir l'existence d'appariements impliquant G31 et C30, non présents dans la structure (b) qui perd la totalité de l'hélice T1-A32, C2-G31, G3-C30, A4-T29.

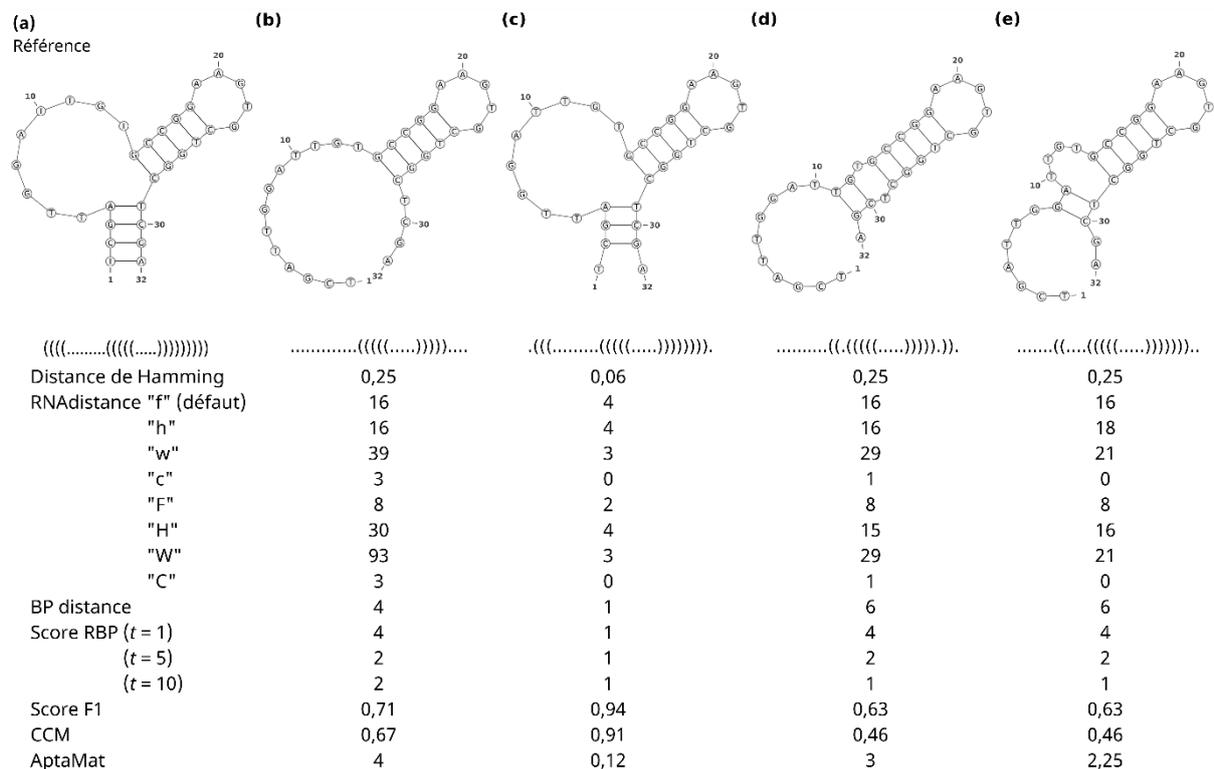


Figure 2-6. Comparaison d'une structure (a) et 4 structures alternatives (b, c, d et e) en utilisant plusieurs méthodes de calculs. La distance de Hamming, RNAdistance et ses différentes approches de calculs, BP, le score RBP (avec t = 1, 5 et 10), le score F1, le Coefficient de corrélation de Mathews (CCM) et AptaMat (structures dessinées avec VARNA).

La Figure 2-7 permet d'appuyer les différences de classement entre algorithmes. DoPloCompare, BP distance ou encore RNAdistance « w » et « c » catégorise la structure (c) comme plus proche de la référence (a) malgré des différences très marquées. A l'inverse les autres métriques comme la distance de Hamming, RNAdistance « f », « h », « F », « H », « W », « C », le score RBP et la distance AptaMat désignent la structure (b) comme la plus similaire à la référence. Il y a donc une variation de l'ordre de classement entre les métriques liée à la façon dont les changements d'appariements sont intégrés dans le mode de calcul. Les différences entre les structures (a) et (c) sont liées au déplacement de la jonction à 3 branches qui se séparent entre les nucléotides G4 et C53 dans la référence (a) alors que cette séparation intervient en G11 et U46 pour la structure (c). En cause, une hélice supplémentaire se forme

par 5 appariements de G7-U50 à G11-U46, ce qui induit de nombreux changements en aval dans la structure secondaire. La structure (b) ne présente qu'une jonction à 2 branches mais fait intervenir des appariements proches sur la portion de A3 à U30. Une seconde épingle à cheveux est également identifiable dans la région de A37 à U54 qui englobe les mêmes nucléotides que la référence, soit la région de G36 à C51. D'un point de vue conservation de structure secondaire, la structure (b) semble donc plus proche que la structure (c).

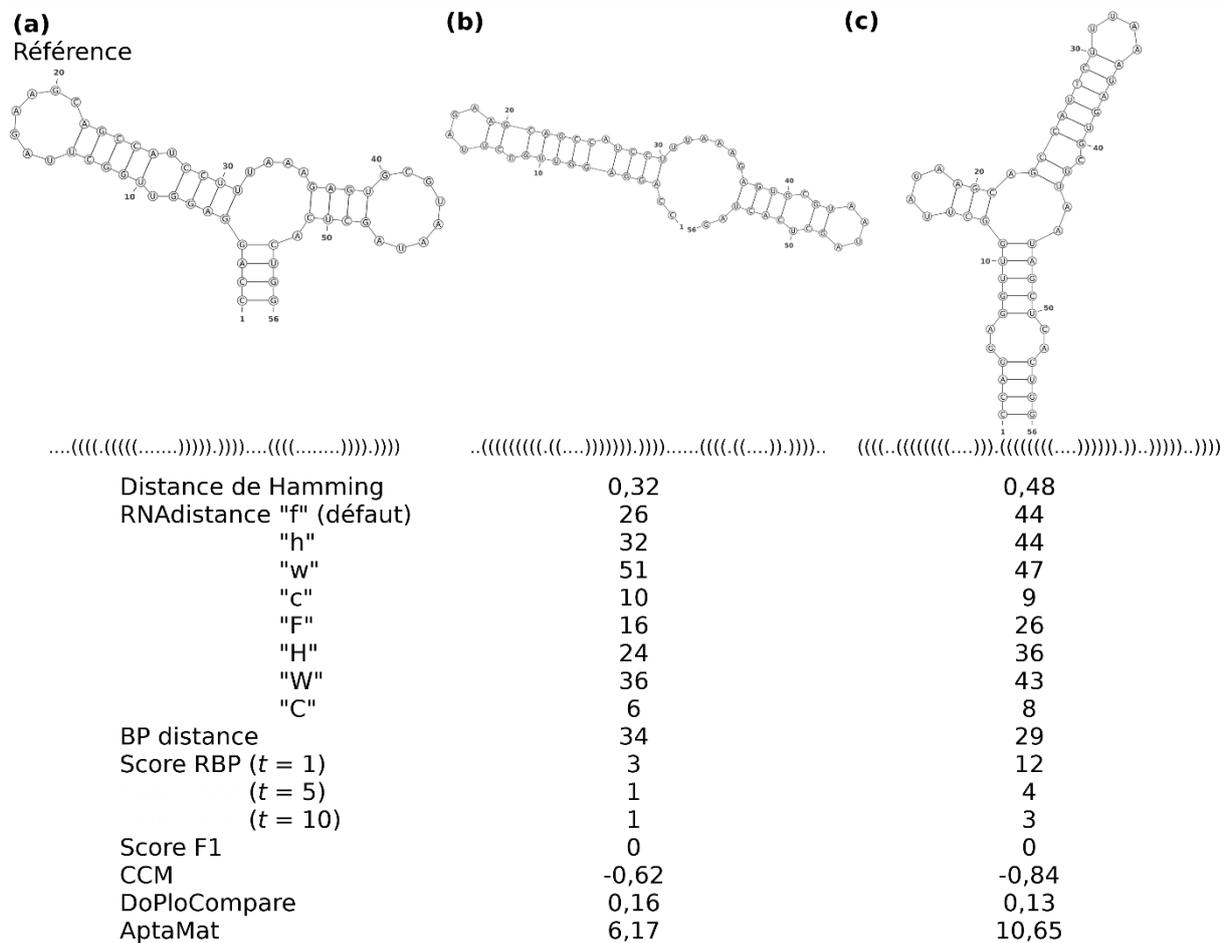


Figure 2-7. Comparaison d'une structure (a) et 2 structures alternatives (b, et c) en utilisant plusieurs méthodes de calculs. La distance de Hamming, RNAdistance et ses différentes approches de calculs, BP, le score RBP (avec t = 1, 5 et 10), le score F1, le Coefficient de corrélation de Mathews (CCM), DoPloCompare et AptaMat (structures dessinées avec VARNA).

#### 4.6.3 Calcul d'AptaMat appliqué aux ensembles de repliements

La flexibilité des oligonucléotides engendre fréquemment un ensemble de repliements possibles (Ganser et al., 2019; Herschlag et al., 2015) pour une séquence donnée, chacune existant dans des proportions différentes. Ceux-ci peuvent évoluer en fonction des conditions expérimentales (concentration en ions, pH, température, ...) ou en présence d'une molécule

reconnue par l'oligonucléotide (Haller et al., 2011). Par conséquent, il existe un intérêt à comparer indépendamment chaque structure alternative à une structure de référence, mais aussi à considérer simultanément l'ensemble conformationnel et sa distribution associée lors de la comparaison avec une structure de référence. L'option “-ensemble” de AptaMat a été développée pour répondre à ce besoin, qui permet le calcul de la distance d'AptaMat selon l'équation 12 après avoir fourni les structures alternatives et les poids associés. La structure de référence peut être une structure consensus connue ou la structure expérimentale à laquelle les structures co- ou sous-optimales prédites peuvent être comparées pour évaluer la fiabilité de la prédiction, comme dans l'exemple abordé en Figure 2-6. Il est aussi possible d'utiliser comme référence la structure la plus probable obtenue en changeant les conditions expérimentales, ce qui permettra d'étudier l'effet des changements conformationnels.

Tableau 5. Comparaison de l'ensemble conformationnel de l'ARN TAR à l'état libre face à la structure secondaire adoptée en forme lié au peptide ligand.

N°	Structure secondaire <sup>a</sup>	Effectifs <sup>b</sup>	Poids	AptaMat
1	(((((...(((.....))))).))))))	9	0,45	0,10
2	.(((...(((.....))))).)).	5	0,25	0,22
3	(((((...(((.....))))))))))	3	0,15	0,00
4	.(((.....(((.....)))..))))).	1	0,05	0,35
5	(((((.....(((...)).))..))))))	1	0,05	0,61
6	.(((...(((.....))))))))).	1	0,05	0,10
<b>AptaMat pondérée</b>				0,16

<sup>a</sup> ((((((...(((.....)))))))))), extrait de la structure pdb 2KDQ. <sup>b</sup> Effectif sur 20 conformations obtenues de la structure pdb 1ANR

Par exemple, les données expérimentales sur l'ARN de la réponse à la transactivation (TAR) de la protéine transactivatrice du VIH-1 TAT ont montré qu'elle peut adopter différentes conformations similaires en l'absence d'un ligand (PDB 1ANR, Tableau 5). À l'inverse, lorsque TAR se lie à un peptide mimant la protéine TAT (code PDB 2KDQ), une seule conformation est échantillonnée. Les résultats de l'algorithme par défaut et l'algorithme pondéré d'AptaMat ont été comparés en utilisant la structure provenant du complexe comme référence, quand une distance pondérée a été utilisée, les poids ont été indiqués comme les fréquences

relatives des différentes conformations. La distance AptaMat pondérée de l'ensemble conformationnel sans ligand par rapport à la structure liée aux peptides est de 0,16 (Tableau 5) ce qui indique globalement des changements conformationnels mineurs. Au regard des distances AptaMat des structures individuelles, la troisième structure alternative (poids = 0,15) est identique à la structure liée aux peptides suggérant que la présence du ligand stabilise une conformation mineure de l'ensemble structural sans ligand. De plus, la conformation sans ligand la plus probable est la plus proche de celle de référence, ce qui suggère que seul un petit changement de conformation se produit.

#### 4.6.4 Performances d'AptaMat dans le regroupement d'oligonucléotides en familles Rfam

Afin de défier ultérieurement AptaMat, ses capacités à classifier des structures secondaires d'ARN selon leur fonction dans une procédure de *clustering* ont été testées. Pour ce faire, un ensemble de structures secondaires connues issues de différentes familles d'ARN extraites de Rfam 14.9 (Kalvari et al., 2021) a été sélectionné. Seules les familles avec des structures expérimentales ont été sélectionnées, pour éviter l'introduction de biais liés à la prédiction des structures secondaires. Les familles intégrant plus de 10 structures obtenues expérimentalement ont été retenues. Parmi les familles contenant plus de 30 structures, une sélection aléatoire de 30 structures a été effectuée pour avoir un jeu de données équilibré. A partir de ces structures tridimensionnelles, les structures secondaires associées sous format dot-bracket ont été obtenues avec x3DNA (Lu, 2020). Le jeu de données testé contient ainsi 291 structures secondaires d'oligonucléotides issues de 14 familles différentes.

La technique de clustering retenue pour évaluer AptaMat est la propagation d'affinité qui permet de s'affranchir du choix du nombre de clusters à définir au préalable. Pour un ensemble  $\{S_1, \dots, S_N\}$  de N structures secondaires, le clustering par propagation d'affinité accepte en entrée une matrice d'affinité  $M_{Affinity} = (m_{ij})_{i,j=1}^N$  obtenue en utilisant l'équation 13 :

$$m_{ij} = \exp\left(-\frac{(Apta_D(S_i S_j))^2}{2\sigma^2}\right) \text{ (Équation 13)}$$

Pour chacune des structures secondaires de l'ensemble, la distance  $Apta_D$  qui la sépare de toutes les autres structures est calculée, nous donnant ainsi une matrice de taille  $N \times N$  structures. Le paramètre  $\sigma$  optimal a été déterminé en utilisant l'index de Caliński et Harabasz (Caliński & Harabasz, 1974). L'échange de messages à valeurs réelles entre les données est effectué jusqu'à ce qu'à l'obtention de groupes d'éléments représentatifs et ainsi, les clusters correspondants émergent. L'évaluation quantitative de la qualité du clustering est effectuée grâce à plusieurs mesures. Le score silhouette permet d'évaluer la différenciation des clusters en affectant une valeur de -1 lorsque les clusters sont difficilement dissociables ou jusqu'à 1 quand les clusters sont parfaitement distingués. Ensuite la précision de clustering a été déterminée comme la somme des éléments de la diagonale de la matrice de confusion divisée par le nombre total d'éléments utilisé pour le clustering. Une bonne précision de clustering est donc indiquée par des valeurs proches de 1, et à l'inverse un mauvais clustering donne des valeurs proches de 0. Enfin, le score aléatoire ajusté mesure les similitudes entre le clustering attendu et celui obtenu, avec des valeurs allant de 0, pour un clustering obtenu avec peu de similitudes avec celui attendu, à 1 pour un clustering identique à l'attendu. Ainsi, les scores silhouettes, la précision de clustering et le score aléatoire ajusté obtenus dans la procédure de clustering (0,55, 0,84 et 0,82, respectivement) sont indicatifs d'un clustering de bonne qualité.

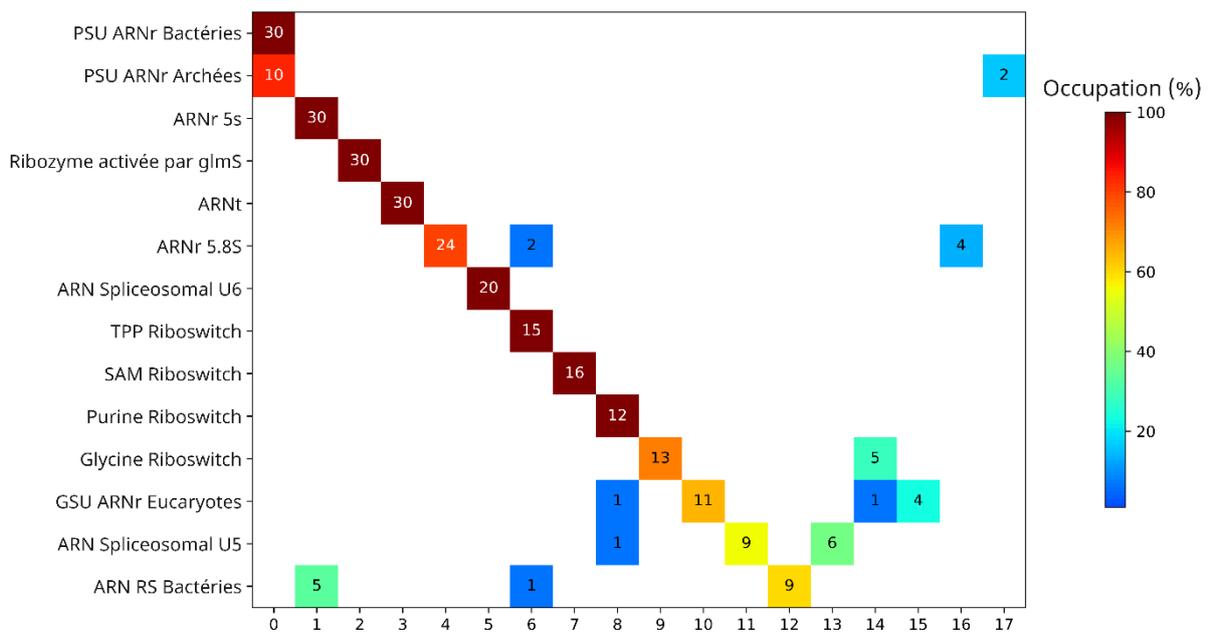


Figure 2-8. Représentation graphique du clustering effectué sur les 14 familles de structures de RFAM en utilisant la distance  $AptaMat$  et l'algorithme de propagation d'affinité. Le nombre de structures par cluster est indiqué dans chaque carré appartenant au cluster indiqué en abscisse. Le gradient de couleur indique l'occupation des structures pour leur cluster.

Les résultats du clustering montrent que *Apta<sub>D</sub>* est globalement capable de correctement classer les ARN du jeu de données en fonction de leur famille Rfam. En effet, la procédure de clustering a permis d'identifier 17 clusters, soit une surestimation de seulement 3 clusters par rapport au 14 familles attendues. De plus, la matrice de confusion (Figure 2-8) montre que la division en familles Rfam est plutôt correctement respectée. Notamment, 8 familles Rfam sur les 14 (i.e. l'ARN de la petite sous-unité ribosomique (PSU ARNr) bactérienne, l'ARN ribosomique (ARNr) 5S, l'ARN de transfert (ARNt), le ribozyme activé par la glucosamine-6-phosphate synthase (*glmS*), l'ARN splicéosomique U6, le riboswitch thiamine pyrophosphate (TPP), le riboswitch purine et le S-Adenosyl methionine (SAM) riboswitch) correspondent chacune à un cluster unique. De plus, 13 clusters sur les 17 identifiés ne contiennent pas des structures appartenant à des familles différentes.

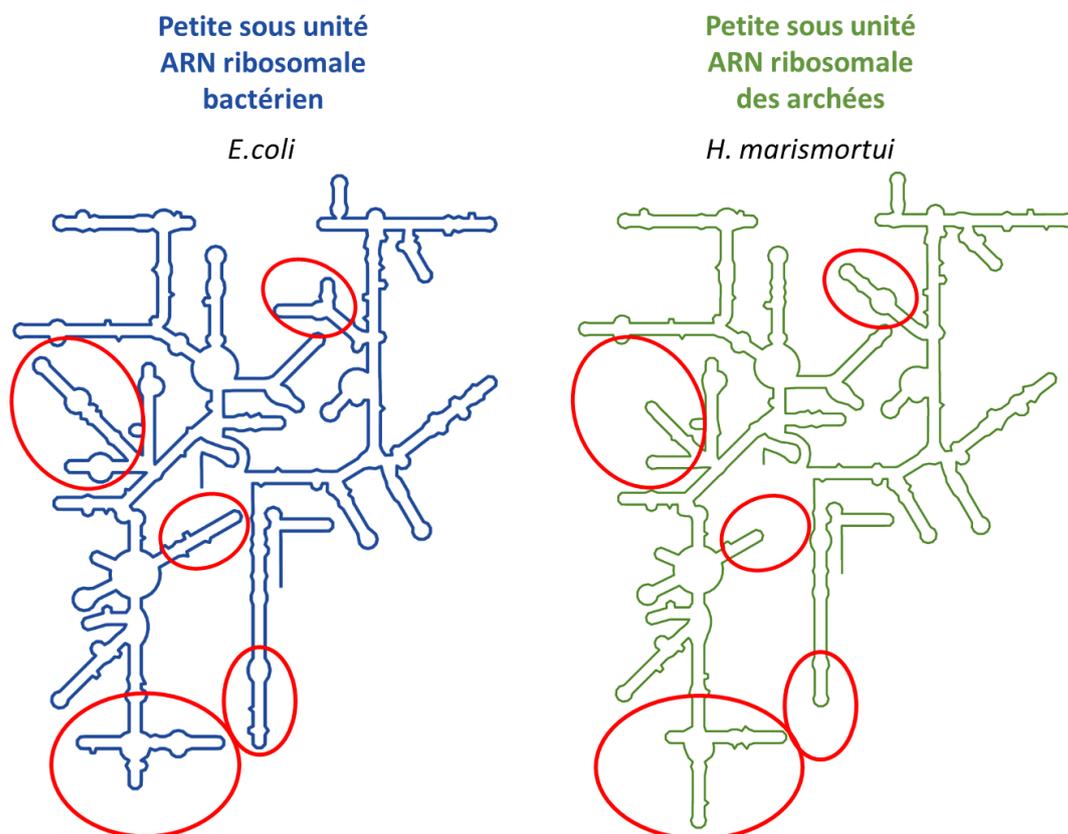


Figure 2-9. Représentation de la structure secondaire modèle des petites sous unité des ARN ribosomiaux des bactéries (en bleu) ou des archées (en vert). Les différences sont identifiées grâce aux cercles rouges. Modèles repris de RibosomeGallery<sup>5</sup>

Dans quelques cas, le clustering n'a pas permis de correctement catégoriser les structures mais ces variations sont explicables. Plus en détails, les deux familles de PSU d'ARNr issues des

<sup>5</sup> <http://apollo.chemistry.gatech.edu/RibosomeGallery/>

bactéries ou des archées partagent de grandes similarités de structures de leurs ARN ribosomaux 16S, avec uniquement quelques différences mineures (identifiées en Figure 2-9) au regard de la longueur des ARNr justifiant leur appartenance au même cluster (cluster 0).

Les deux structures de la PSU ARNr d'archées isolées dans le cluster 17 correspondent à deux structures entièrement caractérisées, alors que les autres ne le sont que partiellement. Dans les autres familles, nous pouvons identifier un cluster majoritaire et un ou plusieurs clusters minoritaires. La présence de ces derniers peut être facilement expliquée. Tout d'abord, la plupart des structures de la famille des ARNr 5.8S appartiennent au cluster 4. Les deux structures d'ARNr 5.8S incluses dans le cluster 6 (code PDB 2WWB, chaîne D, et 2WWA, chaîne D) sont beaucoup plus courtes (60 nucléotides) par rapport aux autres ARNr 5.8S, probablement à cause de problèmes de résolution de structure expérimentale, ce qui les rapproche de la famille des riboswitch TPP. Le cluster 16 contient 3 ARNr 5.8S de *Drosophila melanogaster*, qui a un ARNr 5.8S plus court (123 nucléotides) par rapport aux autres espèces, et un ARNr 5.8S humain qui présente plusieurs résidus manquants aux extrémités 5' et 3', ce qui résulte en une structure tronquée.

Pour les ARN de la famille des Glycine riboswitch, la division en 2 *clusters* dépend à nouveau de la longueur. Les structures du cluster 9 se limitent au domaine synthétique II du riboswitch tandis que le cluster 14 comprend une plus grande partie du riboswitch et provenant de différentes espèces.

La famille des ARN splicéosomiques U5 se divise en 3 *clusters*. Les *clusters* 11 et 13 sont les plus importants et sont représentatifs de portions manquantes aux extrémités 3' (cluster 11) ou 5' (cluster 13). La structure de l'ARN splicéosomique U5 de *Saccharomyces cerevisiae* est classée dans le cluster 8, partagé avec les Purine Riboswitch, en raison d'une insertion dans sa séquence qui n'est pas présente chez d'autres espèces (Mitrovich & Guthrie, 2007).

Les structures de la grande sous-unité ribosomique des ARN eucaryotes se divisent en plusieurs clusters, dépendant d'abord du type d'ARN ribosomique, soit 25S ou 26S, qui est intégré dans le cluster 10, ou 28S dans le cluster 15. Deux structures ont été distribuées dans d'autres clusters (8 et 14) car ils appartiennent à une espèce d'algue unicellulaire *Euglena gracilis* qui se démarque par sa chaîne d'ARN ribosomal atypique.

De la même façon, la famille des ARN de reconnaissance du signal (ARN RS) bactérien se divise en fonction du type d'ARN ribosomal. Le cluster 12 contient les ARN 4.5S et le cluster 1 contient le domaine S des ARN 7S, qui ne peut pas être différencié des ARNr 5S car les ARN 7S sont les précurseurs des ARNr 5S (Szeberenyi et al., 1984). Le cluster 6, contenant les riboswitch TPP, contient l'ARN 7S.S de l'espèce *Methanocaldococcus jannaschii*. L'ensemble des résultats et la réflexion associée ont été présentés dans l'article Binet et al. 2023.

## 5. Performances des outils de prédiction

Une fois la métrique de comparaison des structures secondaires des oligonucléotides sensible et fiable définie, il a été possible de comparer les performances des outils de prédiction de la structure secondaire sélectionnés en Partie 2.3, notamment mfold, RNAfold, Contrafold, CentroidFold, MC-fold, Linearfold, Ufold, SPOT-RNA et MXfold2. La capacité de retrouver la structure secondaire expérimentale des structures du jeu de données défini en Partie 2.1 de chacun de ces outils a été testée. Seules les structures contenant des G-quadruplexes ont été retirées car les limites des outils de prédiction vis-à-vis de ce motif particulier sont bien connues et la plupart des outils sélectionnés sont incapables d'en faire la prédiction (Afanasyeva et al., 2019; Luo et al., 2012). Seul RNAfold permet de gérer les G-quadruplexes et, pour une analyse exhaustive, les 27 structures contenant ce motif ont été étudiées avec cet outil et traitées séparément.

### 5.1. Génération et classification des structures

Tous les outils de prédiction sélectionnés prennent en entrée la séquence nucléotidique (structure primaire) des oligonucléotides et produisent une ou plusieurs structures secondaires. Avant les prédictions avec les outils non-compatibles avec les séquences ADN, les thymines ont été changées en uraciles. C'est le cas de MXfold2, Ufold, SPOT-RNA, CentroidFold, Linearfold, MC-Fold, mfold avec le modèle ARN Mathews (1999) et RNAfold avec le modèle ADN de Turner (2004).

Certains outils offrent la possibilité de préciser des paramètres impactant la prédiction, qui ont été donc testés. En particulier, mfold et RNAfold, basés sur l'estimation du minimum d'énergie libre (MEL), disposent de paramètres thermodynamiques adaptés à la prédiction d'ARN (paramètre par défaut) ou d'ADN. Mfold implémente les paramètres de Mathews

(1999) et SantaLucia (1998), pour l'ARN et l'ADN, respectivement, et RNAfold implémente ceux de Turner (2004) et Mathews (2004) pour l'ARN et l'ADN, respectivement. De plus, avec ces deux outils, il est possible de générer des solutions sous-optimales : mfold permet de définir un pourcentage de sous-optimalité ( $p$ , 5 % par défaut), qui correspond au pourcentage de la MEL qui sera considéré pour le calcul des solutions sous-optimales. RNAfold permet pas de choisir un seuil énergétique pour la génération des structures sous-optimales dans son implémentation dans RNAsubopt, faisant partie de la suite ViennaRNA. Les deux paramètres ont été modifiés pour obtenir entre 5 et 10 structures sous-optimales. Les autres paramètres modifiables dans mfold ont été gardés par défaut dans cette étude. Pour RNAfold, dans l'analyse des structures contenant des G-quadruplexes, le paramètre "gquad" pour induire la prédiction de ce type de motif a été utilisé.

Comme dit précédemment, LinearFold peut être couplé à l'approche *knowledge-based* (LinearFold-C) ou thermodynamique (LinearFold-V). Il a donc été testé avec les deux variantes. Par contre, CentroidFold peut être couplé à plusieurs distributions de probabilité, mais les développeurs ont montré que les meilleures performances sont obtenues avec le modèle Vienna RNAfold McCaskill et, donc, seule cette combinaison a été testée (Sato et al., 2009).

MC-Fold a été testé avec les paramètres par défaut, ainsi qu'avec le paramètre « *pseudoknotted* » (indiqué MC-Fold+P) afin de vérifier sa capacité à prédire les pseudonœuds.

Pour les trois outils de prédiction *deep learning* les paramètres ajustables sont associés aux bases de données d'apprentissage utilisées pour la prédiction. Les paramètres par défaut de ces 3 outils ont été conservés lors de ce travail.

Pour chaque oligonucléotide, la structure secondaire obtenue suite à la prédiction a ensuite été comparée à la structure secondaire expérimentale en utilisant AptaMat comme métrique. Pour chaque prédiction obtenue par outil, la distance  $Apta_D$  nous apporte une mesure de sa variabilité par rapport à la structure expérimentale. Pour faciliter l'interprétation des résultats, et bien qu'AptaMat soit une métrique continue, un seuil  $Apta_D$  a été fixé et optimisé grâce aux observations sur le clustering des familles d'ARN provenant de Rfam (0) pour classifier les prédictions. Parmi les familles qui ont été réparties dans les bon clusters, une faible  $Apta_D$  moyenne entre les membres a pu être observée, avec pour les familles des ARNt et les ribozymes activés par glmS des valeur  $Apta_D$  moyenne respectivement de 1,21 et 2,54.

Combinée aux observations faites précédemment sur quelques exemples de comparaison additionnels, le seuil a été fixé à  $Apta_D \leq 1,5$ . Trois rangs de qualité de prédiction sont évalués : i)  $Apta_D = 0$ , la prédiction est identique à la référence. ii)  $0 < Apta_D \leq 1,5$ , la prédiction est similaire à la référence. iii)  $Apta_D \geq 1,5$ , la prédiction est différente de la référence.

**ENTREES :**

SR = Dot-bracket de la structure référence

SC = Dot-bracket de la Structure comparée

L = Longueur des chaînes de caractères SR et SC

**INITIALISATION :** affecter la valeur  $N_{id} = 0$  pour le comptage des caractères communs

**TRAITEMENT :**

Lire SR et SC ;

**pour**  $u1 \in SR$  et  $u2 \in SC$  **faire**

**si**  $u1 = u2$  **alors :**

$N_{id} = N_{id} + 1$

**sinon si**  $u2 = "+"$  et **si**  $u2 \neq "."$  **alors :**

$N_{id} = N_{id} + 1$

**sinon :**

        Ne rien faire

**fin :**

        Affecter la nouvelle valeur de  $N_{id}$

**fin**

Calculer le tanimoto  $T = N_{id}/L$

**SORTIES :** Afficher la valeur de  $T$

Figure 2-10. Pseudocode du calcul du coefficient de Tanimoto modifié dans le but de calculer les similitudes entre les G-quadruplexes issus d'une structure de référence et les annotations de prédictions de RNAdistance.

En outre, RNAfold intègre un paramètre (*quadruplex*) pour retrouver les G-quadruplexes en identifiant les sites riches en guanines et répétés sur plusieurs positions dans la séquence. Les guanines candidates à la formation d'un G-quadret sont alors marquées avec un « + » en substitution des caractères se référant aux paires de bases. Cette annotation spécifique

obtenue n'est pas adaptée à l'utilisation d'AptaMat pour la comparaison de structure secondaire car les appariements de guanines ne sont pas clairement identifiés et seules les positions de site impliqué dans les G-quadruplexes sont marquées. Pour pouvoir procéder à la comparaison, une autre approche a été envisagée. Il est possible de compter le nombre de guanines marquées d'un « + » qui interagissent avec une autre guanine marquée et de quantifier les sites de formation des G-quadruplexes bien prédits. Pour cela, nous avons utilisé le calcul du coefficient de Tanimoto avec une modification permettant d'associer les « + » présents dans la structure prédite par RNAfold à n'importe quel caractère différent de « . » dans la structure de référence qui désignerait un appariement de guanine. Bien qu'approximative, cette modification permet de visualiser sur les données de structures le pourcentage de G-quadruplexes correctement positionnés par RNAfold avec le paramètre "gquad". Son fonctionnement est détaillé dans le pseudocode en Figure 2-10.

## 5.2. Performances des approches de Minimisation d'Énergie Libre (MEL)

Les approches MEL, premières approches développées pour la prédiction de structures secondaires des acides nucléiques simple brin, forment une classe de prédicteurs qui se base sur des règles thermodynamiques afin d'estimer l'énergie libre d'une ou plusieurs structures pour en retourner celle d'énergie minimale. Les MEL incluent donc mfold et RNAfold, qui forment une catégorie reconnue et très utilisée dans l'étude des structures secondaires simples, sans motifs type pseudonœuds ou G-quadruplexes (Hofacker, 2014).

### 5.2.1 Comparaison sur l'ensemble des données

Dans un premier temps, l'analyse de l'ensemble des structures a pu être effectuée en utilisant les paramètres d'estimation d'énergie libre par défaut, normalement appropriés pour l'ARN. Ainsi les paramètres pour l'ARN de Mathews (1999) et Turner (2004), respectivement associés à mfold et RNAfold, ont été utilisés pour la comparaison. Les prédictions résultantes avec mfold et RNAfold ont fourni des résultats comparables, avec ~45 % et 47 % de prédictions exactes et 82 % et 81 % de bonnes prédictions au vu des valeurs  $Apta_D \leq 1,5$  face à la structure secondaire expérimentale (Figure 2-11a).

Pour ce qui concerne le type d'oligonucléotide, les résultats montrent des différences mineures en termes de qualité de prédiction entre les séquences d'ADN simple brin et d'ARN pour mfold. En effet, mfold prédit correctement 48 % et 45 % des séquences d'ADN simple

brin et d'ARN, respectivement. Pour les prédictions avec des  $Apta_D \leq 1,5$  face à la structure secondaire de référence, les pourcentages augmentent et on obtient jusqu'à 75 % et 83 % de prédictions proches de la structure expérimentale. RNAfold propose une précision légèrement meilleure pour la prédiction des structures secondaires d'ARN par rapport aux structures d'ADN simple brin, avec 48 % et 40 % de prédictions exactes pour les séquences d'ARN et d'ADN simple brin, respectivement. Lorsque les prédictions avec un  $Apta_D \leq 1,5$  sont incluses dans l'analyse, RNAfold atteint un pourcentage de bonne prédiction de 84 % et 64 % pour les structures d'ARN et d'ADN simple brin, respectivement.

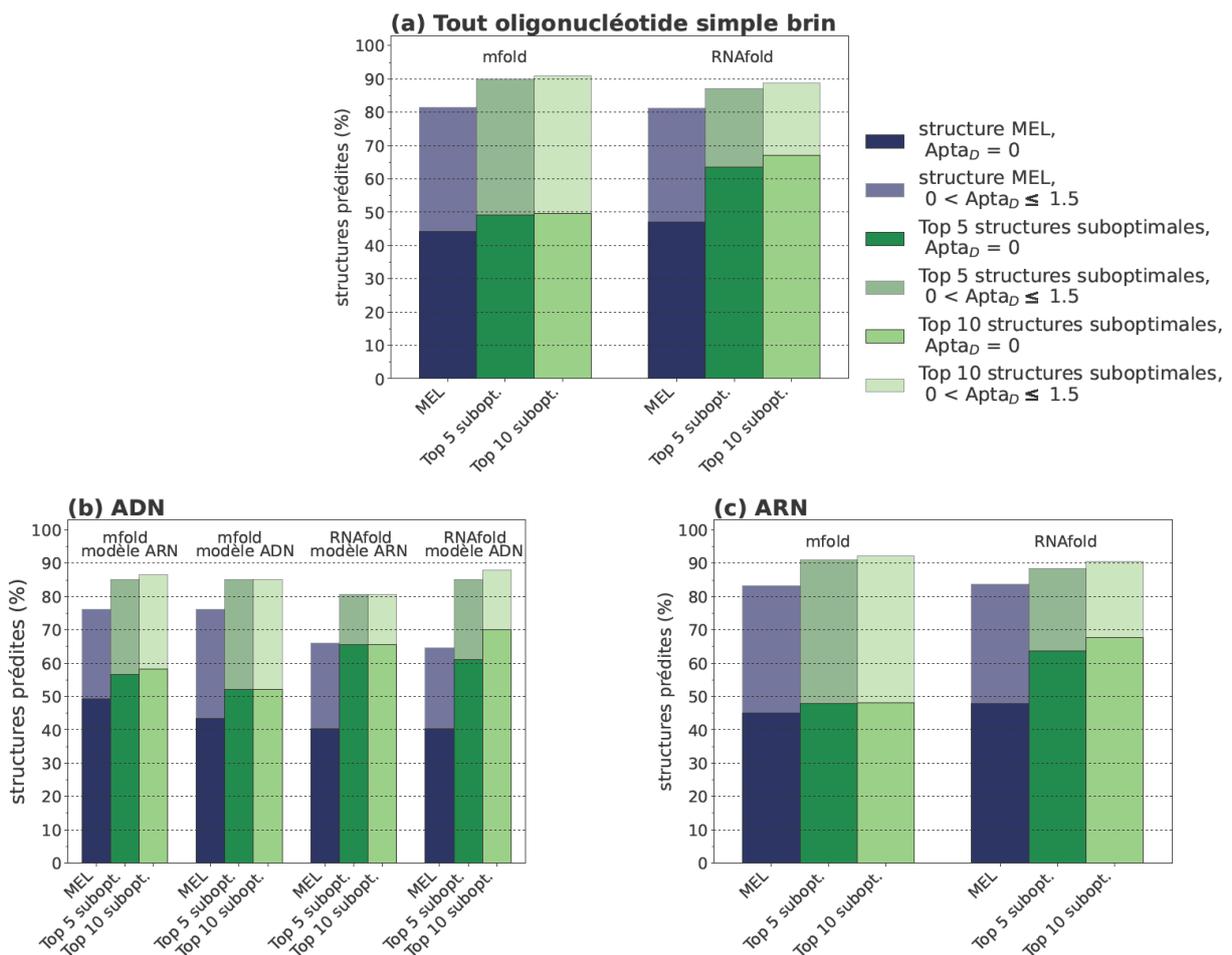


Figure 2-11. Proportion de prédictions exactes ( $Apta_D = 0$ , couleurs pleines), et prédictions similaires ( $Apta_D \leq 1,5$ , couleurs transparentes) suggérées par mfold et RNAfold par comparaison à la référence expérimentale. Les proportions sont mesurées sur a) l'ensemble du jeu de données (ADN + ARN), b) les ADN simple brin ou c) les ARN. Une triade de barre inclut les pourcentages pour les structures MEL, les solutions suboptimales du Top 5 ou du Top 10. Les résultats des modèles ARN Mathews 1999 (mfold) et Turner 2004 (RNAfold) sont inclus en plus des modèles ADN SantaLucia (mfold) et Mathews 2004 (RNAfold).

Il est important de souligner que les algorithmes mfold et RNAfold imposent une taille minimale d'une boucle d'une épingle à cheveux de 3 nucléotides. Or, dans l'ensemble de données considéré, la présence de boucles composées de 2 nucléotides a été observée (PDB ID : 1EZD, 1SNJ, 2N8A, 4ER8, 4F41, 4F43, 1RNG, 2L6I, 2B6G, 2JYM, 2ES5, 2PJP, 2UWM, 1EKZ, 2MTJ, 2M3Q, 6U8D, 4ZT0, 5VW1 et 5XBL). En conséquence, ces acides nucléiques monocaténares ne peuvent pas être prédits exactement à cause de cette contrainte de taille. De plus, les deux outils ont certaines limites dans la prédiction correcte des paires de bases impliquant les extrémités 5' et 3'. Ce problème pourrait être considéré comme négligeable puisque l'impact des paires de bases terminales sur le repliement global est limité mais peut expliquer  $\sim 10\%$  des structures secondaires affichant des valeurs  $0 < Apta_D \leq 1,5$ .

### 5.2.2 Effet du modèle thermodynamique sur la prédiction des structures d'ADN simple brin

Les approches MEL offrent la possibilité d'utiliser des modèles thermodynamiques différents en fonction du type d'oligonucléotide. Cette option a donc été explorée et les modèles de SantaLucia 1998 et de Mathews et al. 2004 pour l'ADN ont été testés avec mfold et RNAfold, respectivement. Pour ce qui concerne mfold, l'utilisation du modèle SantaLucia, adapté pour les ADN, produit une baisse dans le pourcentage des prédictions correctes, qui passe de 49 % avec le modèle ARN de Mathews à 43 % avec le modèle ADN de SantaLucia. Cette différence contre-intuitive peut s'expliquer par les différences de pénalités appliquées par les deux modèles, en effet le modèle SantaLucia pénalise les paires A-T positionnées aux extrémités d'une hélice, ce qui n'est pas fait par le modèle Mathews. Aussi, les cinq structures correctement prédites par le modèle ARN – et non par le modèle ADN – soit 6IY5, 6FKE, 1ECU, 2VIC et 2L5K (Figure 2-12) montrent pour la plupart une différence d'une paire de base (6IY5, 6FKE, 1ECU, 2VIC) et une perte d'hélice pour 2L5K. Pour ce qui concerne RNAfold, une différence de 1 % est identifiable dans la proportion de prédictions identiques entre les modèles ARN et ADN ( $Apta_D = 0$ ). Cette moindre différence peut être liée à la façon dont l'énergie des appariements terminaux est estimée ou aux difficultés à prédire la structure secondaire des oligonucléotides de taille inférieure à 15 nucléotides.

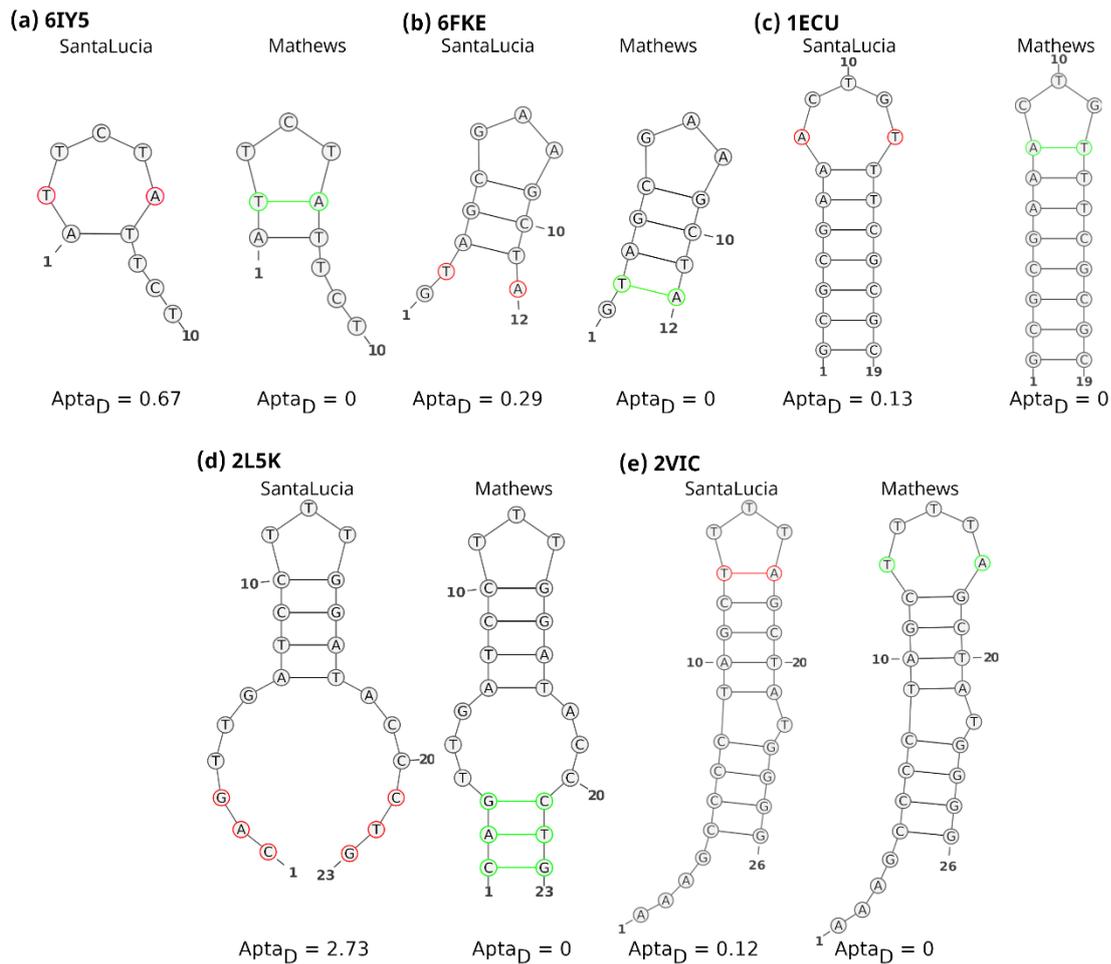


Figure 2-12. Prédiction des structures secondaires de (a) 6IY5, (b) 6FKE, (c) 1ECU, (d) 2L5K et (e) 2VIC en utilisant les modèles ADN (SantaLucia) et ARN (Mathews) de mfold. Les nucléotides dont l'appariement est mal prédit sont marqués en rouge, et à l'inverse sont coloré en vert lorsque l'implication du nucléotide dans la structure secondaire est bien prédite. La distance  $Apta_{Mat}$  par rapport à la référence expérimentale est reportée sous chaque structure (dessinée avec VARNA).

En extrapolant l'analyse des résultats sur les prédictions similaires ( $Apta_D \leq 1,5$ ), les modèles ARN et ADN démontrent les mêmes capacités prédictives pour les structures d'ADN simple brin, avec  $\sim 77\%$  et  $\sim 66\%$  de prédition similaires obtenues respectivement avec mfold et RNAfold.

### 5.2.3 Analyse des prédictions sous-optimales

L'étude des structures d'oligonucléotides requiert souvent de s'intéresser aux configurations alternatives. En effet, la mobilité de ces molécules fait qu'elles sont capables d'adopter plusieurs repliements stables énergétiquement et équiprobables en solution (Ganser et al., 2019; Havrila et al., 2018). En biologie structurale, l'obtention de ces repliements alternatifs est possible grâce à la RMN. Dans notre jeu de données de référence, plusieurs structures

obtenues via cette technique (ID PDB : 1M82, 1SCL, 1MFY, 1JO7, 5UZZ, 2FEY, 2N6W) montrent des repliements alternatifs, bien que limités à quelques paires de bases. Ceci questionne donc l'intérêt de se concentrer sur la seule prédiction de la structure à énergie libre minimale lors de l'obtention des prédictions. Les approches MEL incluent la possibilité de générer des structures sous-optimales dont l'énergie estimée reste proche de l'énergie libre minimale estimée. L'analyse des prédictions a donc été étendue sur les 5 et 10 premières structures sous-optimales afin de mesurer le potentiel des algorithmes MEL à retrouver la structure expérimentale de référence. Les résultats de prédictions sous-optimales ont donc été comparés à la référence avec AptaMat et les performances ont été comparées aux performances des prédictions limitées à la structure MEL en Figure 2-11b.

Si on considère le jeu de données entier, l'inclusion des solutions sous-optimales augmente le pourcentage de structures identiques ou similaires à la référence, indépendamment du modèle utilisé et de l'outil. Néanmoins, les prédictions de RNAfold sont plus affectées par la présence des solutions sous-optimales par rapport à mfold, car le nombre de prédictions exactes augmente de ~20 % pour RNAfold et seulement de ~7 % pour mfold. L'amélioration est identique pour les deux outils si on inclut les résultats de prédictions considérées comme proches ( $Apta_D \leq 1,5$ ), avec 90 % de bonnes prédictions (Figure 2-11a).

En outre, les séquences d'ADN simple brin bénéficient plus de l'inclusion des structures sous-optimales que les séquences ARN (Figure 2-11b). En incluant 5 structures sous-optimales, on observe une augmentation du pourcentage des prédictions exactes de structures d'ADN de ~10 % et ~25 % avec mfold et RNAfold, respectivement. L'inclusion de 10 structures sous-optimales n'apporte que peu d'amélioration : seul RNAfold avec le modèle de Mathews pour l'ADN retrouve ~11 % de structures exactement prédites en plus.

Il est intéressant de souligner que, avec la prédiction de la seule structure à énergie minimale, les performances de mfold et RNAfold sont équivalentes, mais, quand les solutions sous-optimales sont incluses dans l'analyse, RNAfold est capable de prédire correctement un pourcentage plus élevé de structures (72 %) par rapport à mfold (55 %) (Figure 2-11a). De plus, si les prédictions similaires ( $Apta_D \leq 1,5$ ) sont aussi prises en compte, mfold montre une augmentation moindre (< 10 %) par rapport aux données obtenues en analysant seulement la structure MEL. Au contraire, RNAfold montre un gain ultérieur de 21 %, atteignant un excellent

pourcentage de 90 % de structures proches de l'expérimentale en utilisant le modèle Mathews pour l'ADN.

L'analyse effectuée sur les structures ARN montre un impact moins important des prédictions sous-optimales obtenues par mfold, avec moins de 5 % de prédiction exactes supplémentaires par rapport à la prédiction à énergie minimale, et moins de 10 % pour les prédictions proches ( $Apta_D \leq 1,5$ ) (Figure 2-11c). En revanche, avec RNAfold le nombre de prédictions correctes augmente de ~20 %, indépendamment du nombre de structures sous-optimales incluses dans l'analyse, ce qui permet d'obtenir un pourcentage de succès > 67 %. L'inclusion des prédictions proches de la référence résulte en une augmentation moindre (<10 %), et les performances de RNAfold et mfold sont équivalentes.

### 5.3. Comparaison globale des outils testés

Les autres outils considérés ici n'offrent pas la possibilité de faire la distinction entre ARN et ADN et produisent une seule prédiction. Néanmoins, ils ont souvent été démontrés comme plus performants que mfold et RNAfold (Fu et al., 2022; Sato et al., 2021; Singh et al., 2019). Pour cela, ils ont été utilisés pour prédire les structures du jeu de données précédemment décrit. La Figure 2-13 présente la distribution globale des prédictions sur l'ensemble des structures testés pour chaque outil en utilisant  $Apta_D$  comme métrique.

#### 5.3.1 Performances des approches Machine Learning et Deep Learning

Les premières méthodes alternatives à celles basées sur la MEL sont les approches de Machine Learning. Ainsi, CONTRAfold écarte l'aspect thermodynamique de l'algorithme de prédiction et se base sur l'apprentissage des motifs observés sur des bases de données de structures secondaires d'ARN. En général, CONTRAfold présente des performances similaires à RNAfold avec 43,5 % de prédictions identiques aux structures obtenues expérimentalement et 82 % de prédictions proches de l'expérimental ( $Apta_D \leq 1,5$ ). Comme pour les méthodes basées sur la MLE, aucune différence significative n'a été observée entre les prédictions des séquences d'ADN et d'ARN : respectivement, on obtient 40 % et 44 % de prédictions exactes et 72 % et 84 % de prédictions proches de la référence.

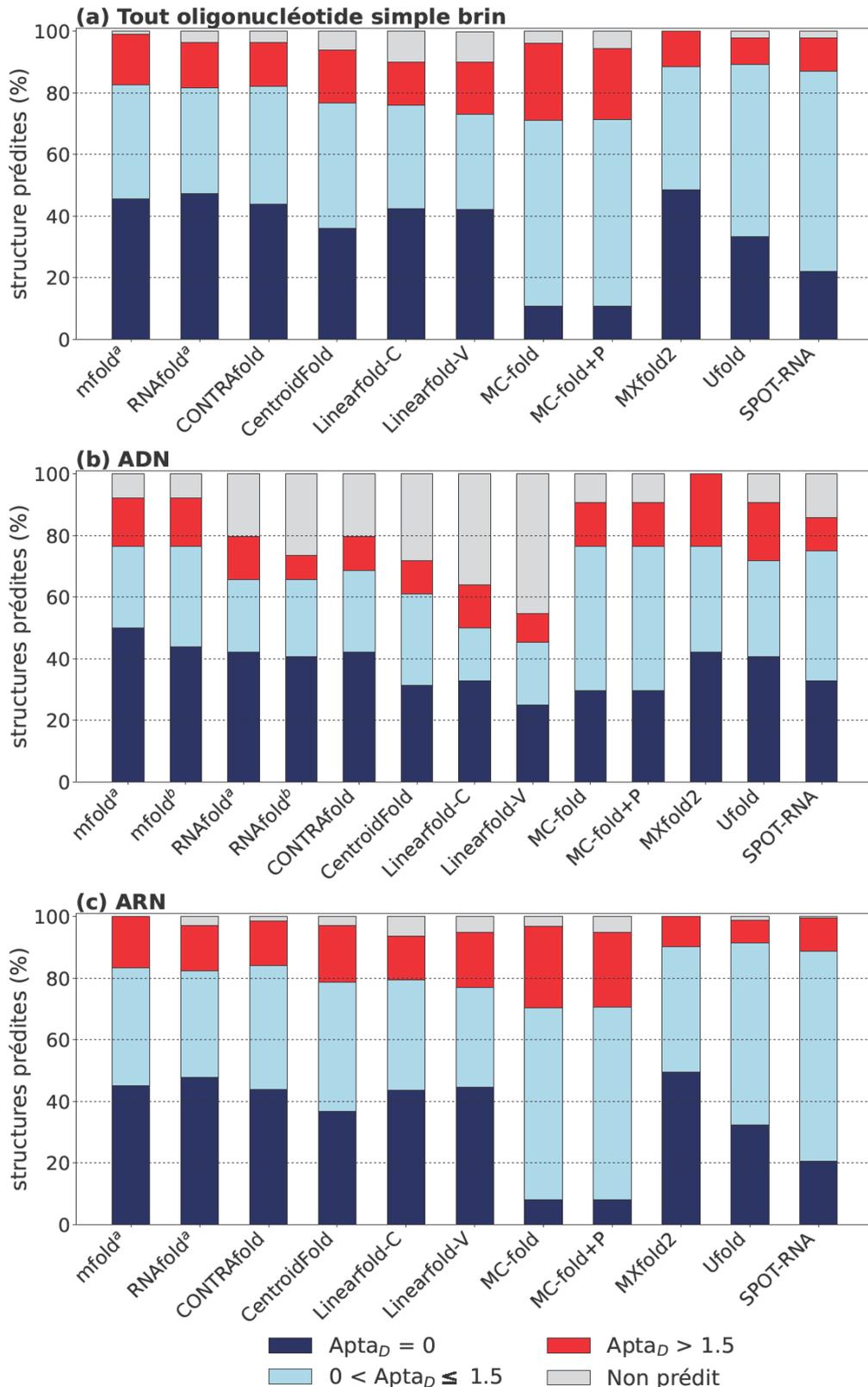


Figure 2-13. Pourcentage de structures prédites par différentes méthodes et la proportion de prédictions exactes ( $Apta_D = 0$ , barres bleues), prédictions similaires ( $Apta_D \leq 1,5$ , barres bleu ciel) et mauvaises prédictions ( $Apta_D > 1,5$ , barres rouges). Les pourcentages sont affichés pour (a) l'ensemble du jeu de données (ADN + ARN), (b) les ADN simple brin ou (c) les ARN. Les structures prédites sans repliement sont représentées dans les barres grises. Sont différenciés mfold et RNAfold selon <sup>a</sup> le modèle ARN, ou <sup>b</sup> le modèle ADN associé.

MC-fold se base également sur des données structurales et implémente une fonction permettant la prédiction des pseudonœuds de type H. La proportion de prédictions identiques à l'expérimentale montre une différence importante entre structures d'ADN (~30 % de prédictions identiques) et structures d'ARN (<10 % de prédictions identiques). Les structures d'ADN sont globalement mieux prédites avec ~23 % de mauvaises prédictions ( $Apta_D > 1,5$ ) contre près de 30 % des structures d'ARN, ce qui le classe parmi les plus faibles des outils testés.

Récemment, les approches d'intelligence artificielle se sont développées pour répondre à différents problèmes biologiques, notamment dans l'étude de la structure secondaire des molécules de nature protéique (Jumper et al., 2021) ou acides nucléiques. Ainsi, 3 outils très récents qui implémentent des algorithmes de *deep learning* (SPOT-RNA (Singh et al., 2019), UFold (Fu et al., 2022), et MXfold2 (Sato et al., 2021)), ont été testés. Ces 3 outils se basent sur des réseaux de neurones qui varient par leur construction, la quantité ainsi que la disposition des données d'entraînement et de test, les fonctionnalités ou le temps de calcul reporté dans le Tableau 6.

Tableau 6. Tableau résumé des différences majeures entre les 3 approches *deep learning* étudiées.

	<b>MXfold2</b>	<b>UFold</b>	<b>SPOT-RNA</b>
<b>ALGORITHME</b>	Réseaux de neurones		Réseaux de neurones et apprentissage par transfert
<b>DONNEES D'ENTRAINEMENT</b>	~ 48 000	~ 140 000	~ 13 000
<b>FONCTIONNALITEES</b>	Prédiction basée sur des critères thermodynamiques	Prédiction des pseudonœuds	
<b>LIMITE DE TAILLE<sup>a</sup></b>	500	1500	500
<b>TEMPS DE CALCUL<sup>b</sup></b>	0,31 s/ séquence	0,16 s/ séquence	78 s/ séquence

<sup>a</sup> La limite de taille est définie selon la taille maximale des structures du jeu de données d'entraînement.

<sup>b</sup> Le temps de calcul par séquence a été mesuré et reporté dans l'article de UFold (Fu et al., 2022).

SPOT-RNA et UFold ont des performances plutôt moyennes, avec ~33 % et ~41 % de prédictions identiques aux données expérimentales obtenues par SPOT-RNA et UFold, respectivement. Cependant, l'inclusion des prédictions proches de l'expérimentale leur fait atteindre un pourcentage de succès de 87 % et 89 %, respectivement. Si on divise le jeu de données selon le type d'oligonucléotide, le pourcentage des prédictions exactes est étonnamment supérieur pour les structures d'ADN pour les deux outils (32 % et 41 % pour SPOT-RNA et UFold, respectivement), contre ~21 % et ~32 % pour les structures d'ARN, respectivement. En revanche, en incluant les prédictions similaires de distance  $Apta_D \leq 1,5$ , la situation s'inverse : UFold et SPOT-RNA prédisent correctement jusqu'à ~93 % et ~89 % des structures d'ARN, contre ~72 % et ~75 % de structures d'ADN (Figure 2-13b-c).

MXfold2 (Sato et al., 2021) implémente un réseau neuronal profond couplé au modèle thermodynamique de Turner et il se présente comme l'outil avec les meilleures performances, avec ~48 % identique à l'expérimentale et seulement ~11 % de prédictions ayant un  $Apta_D > 1,5$ . En regardant les performances en fonction du type d'oligonucléotide, comme attendu, les prédictions de MXfold2 sur les structures d'ARN sont meilleures que celles sur les structures d'ADN : pour les premières, MXfold2 prédit ~49 % de structures identiques et jusqu'à ~90 % de structures similaires ( $Apta_D \leq 1,5$ ); pour les deuxièmes ~42 % des prédictions sont identiques à la référence et ~77 % des prédictions sont similaires ( $Apta_D \leq 1,5$ ).

La différence majeure entre SPOT-RNA, UFold et MXfold2 réside dans l'approche de prédiction qui intègre les paramètres thermodynamiques dans MXfold2. L'importance de la thermodynamique dans la prédiction de la structure secondaire semble donc avoir un impact sur les résultats obtenus. Cependant, les modèles basés sur des approches d'intelligence artificielle dépendent fortement du jeu de données utilisé, de sa complexité et sa pertinence, ce qui permet d'envisager des améliorations de ces modèles dans le futur avec des jeux de données plus complets et incluant les ADN, par exemple. De plus, au vu des limites des paramètres thermodynamiques citées précédemment, il serait aussi possible de travailler sur la précision de ces derniers.

### 5.3.2 Performances des approches mixtes

Au-delà des approches *knowledge-based* et basées sur la MLE, des algorithmes qui modifient ces approches existent. Par exemple, Linearfold permet d'utiliser les modèles basés sur la MLE (Linearfold-V) et ceux *knowledge-based* (Linearfold-C) mais il implémente un algorithme avec une complexité temporelle linéaire, ce qui les rend exploitables pour la prédiction des oligonucléotides longs. En effet, contrairement aux autres outils, les deux versions de Linearfold sont capables d'obtenir une structure proche de la structure expérimentale pour ce type d'oligonucléotides (ID PDB 1GRZ, 247 nucléotides, ID PDB 2R8S, 159 nucléotides), mais elles présentent une difficulté notable à prédire les oligonucléotides courts ( $\leq 15$  nucléotides), qui sont prédits comme non-structurés. Globalement, les deux versions de Linearfold se comportent de la même façon (basée sur la MLE pour Linearfold-V et *knowledge-based* pour Linearfold-C), mais, il faut remarquer que cet outil n'est pas adapté à la prédiction des structures d'ADN, car seul  $\sim 33\%$  et  $25\%$  des prédictions sont identiques à la référence avec Linearfold-C et Linearfold-V, respectivement. Ces pourcentages augmentent à  $50\%$  et  $\sim 45\%$  si les prédictions similaires à l'expérimentale sont incluses. En ce qui concerne les données d'ARN, les performances sont similaires à celles des approches MEL, avec  $\sim 44\%$  et  $\sim 45\%$  de prédictions identiques avec respectivement Linearfold-C et Linearfold-V. En incluant les structures similaires à la référence ( $Apta_D \leq 1,5$ ) Linearfold-C et -V parviennent à prédire respectivement  $\sim 79\%$  et  $77\%$  de structures secondaires similaires.

CentroidFold est un autre outil qui peut être associé aux modèles basés sur la MLE ou les *knowledge-based*, mais les auteurs ont montré que CentroidFold, sous le modèle implémenté par RNAfold, a des meilleures performances. Néanmoins, il est capable de prédire correctement  $36\%$  des structures oligonucléotidiques et ce pourcentage augmente à  $\sim 77\%$  si les prédictions proches de la référence sont incluses. Ses performances sont légèrement dépendantes du type d'oligonucléotide : pour les structures d'ADN on obtient  $31\%$  de prédictions identiques à l'expérimental et jusqu'à  $61\%$  de prédictions similaires; en revanche, pour les structures d'ARN on arrive à avec  $\sim 37\%$  de prédictions identiques ( $Apta_D = 0$ ) et  $\sim 79\%$  de prédictions similaires à la référence ( $Apta_D \leq 1,5$ ) (Figure 2-13c).

## 5.4. Les difficultés de prédictions des motifs structuraux particuliers

Au-delà de la détermination des performances relatives des outils choisis, l'analyse des données a permis d'identifier plusieurs limites structurales à la prédiction des structures secondaires des oligonucléotides.

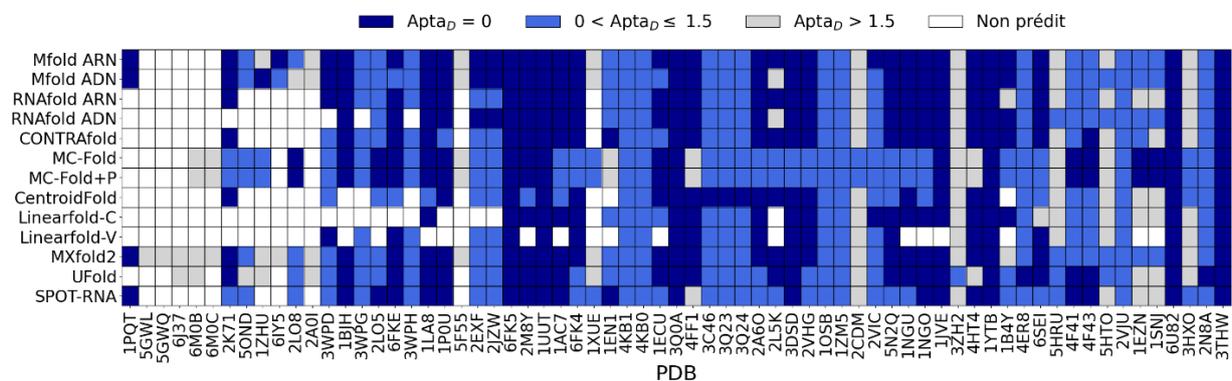


Figure 2-14. Qualité des prédictions par structure ADN et par outil, réparti en 4 catégories : prédictions exactes ( $Apta_D = 0$ , carrés bleus), prédictions similaires ( $Apta_D \leq 1,5$ , carrés bleu ciel), mauvaises prédictions ( $Apta_D > 1,5$ , carrés gris) et non prédit (carrés blanc)

### 5.4.1 Petits oligonucléotides et « Minidumbbell »

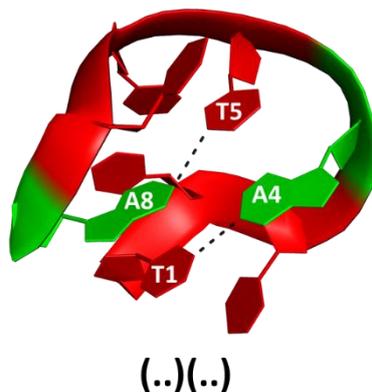


Figure 2-15. Représentation sur PyMOL de la structure tridimensionnelle d'un "minidumbbell" (ID PDB : 5GWQ). La structure secondaire associée est indiquée en bas de la figure. Les nucléotides sont colorés en rouge pour les thymines et en vert pour les adénines. Les appariements sont indiqués en pointillé noir.

Sur la totalité du jeu de données, les structures de petite taille ( $\leq 15$  nucléotides) sont celles qui sont fréquemment mal ou non prédites par les outils. 11 structures sur les 13 ayant une taille  $\leq 10$  nucléotides sont mal prédites par 6 des méthodes utilisées (RNAfold ARN et ADN, Linearfold-C et -V, CONTRAfold et CentroidFold). Parmi ces structures de très petite taille, aucune méthode n'est parvenue à retrouver le repliement de 5GWL, 5GWQ, 6J37, 6MOB et 6MOC, présenté en Figure 2-14 sur l'extrémité gauche de la matrice et intégralement

représentées par blocs coloré gris ou blanc indiquant une mauvaise prédiction. Ces 5 structures ont la particularité d’être un type de motif récemment caractérisé appelé “*minidumbbell*” (Figure 2-15) (Guo & Lam, 2016). Deux problèmes majeurs se posent à cause de cette récente caractérisation : i) les données structurales récentes et peu nombreuses ne sont pas incluses dans les outils de prédiction dont l’algorithme a été entraîné sur des données structurales connues (ex : CONTRAfold, MXfold2, Ufold, SPOT-RNA) ; ii) les paramètres thermodynamiques intégrés aux outils de prédiction MEL ont été définis dans les années ’90-2000 et, comme mentionné précédemment, n’incluaient pas la possibilité de former des boucles composées de moins de 3 nucléotides, typiques des “*minidumbbell*”. Il est important de noter que ce type de motif a initialement été observé sur des brins d’ADN natifs et que l’acquisition de ces structures était, avant tout, motivée par un besoin de caractérisation de la structure de ces motifs seuls. Il s’agit donc d’un type particulier de motif existant au sein des acides nucléiques mais probablement difficile à rencontrer libre en solution.

#### 5.4.2 Pseudonœuds

Le jeu de données utilisé reporte 77 structures (75 ARN et 2 ADN) contenant des interactions à longue distance, dont les pseudonœuds. Seuls 3 outils sont capables de prédire ces motifs : MC-fold, Ufold et SPOT-RNA. Les résultats ont démontré des capacités de prédiction très variables des 3 outils. Notamment, MC-fold, avec le paramètre *pseudoknotted*, est en principe capable de prédire uniquement les pseudonœuds type-H, ce qui représente 17 des 77 structures qui en contiennent. Cependant, seulement 8 de ces 17 structures ont été prédites avec une valeur  $Apta_D \leq 1,5$ . De plus, 3 autres structures présentant des interactions à longue distance ont été correctement prédites, ce qui amène le total de pseudonœuds correctement prédits à 11 sur 77, faisant de MC-fold le moins performant des 3 outils pour prédire les pseudonœuds.

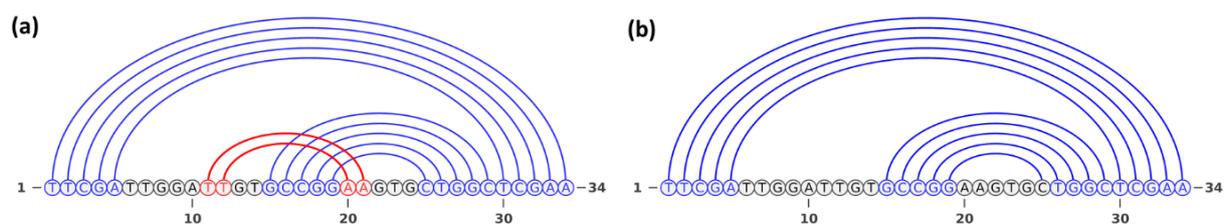


Figure 2-16. Représentation linéaire de la structure secondaire de 5HTO a) structure expérimentale b) résultat de prédiction de mfold ADN, RNAfold ADN, MXfold2 et UFold.

En effet, les performances de Ufold et SPOT-RNA dans la prédiction de structures contenant des pseudonœuds sont meilleures que celles de MC-fold. Plus en détail, Ufold montre les meilleurs résultats en prédisant 4 pseudonœuds sur 77 avec  $Apta_D = 0$ , et 61 sur 77 pseudonœuds montrant une proximité satisfaisante ( $Apta_D \leq 1,5$ ). SPOT-RNA n'a retrouvé aucune structure exacte, mais 59 pseudonœuds sur 77 ont été prédits comme étant proches de la référence ( $Apta_D \leq 1,5$ ).

Aucune des 2 structures d'ADN (5HTO et 5HRU), ayant de pseudonœud type H, n'a été exactement prédite et les meilleures distances  $Apta_D$  atteintes sont respectivement 1,091 et 1,2 avec MXfold2. Les mêmes résultats ont été obtenus par mfold, bien que ces deux outils ne soient pas aptes à prédire les pseudonœuds. L'analyse des appariements de ces meilleures prédictions a permis de constater que seuls deux appariements étaient manquants dans les structures prédites et concernaient les nucléotides impliqués dans la formation du pseudonœud. La distance relativement faible observée s'explique donc par la présence commune des 2 hélices entre la référence et la prédiction visible en Figure 2-16. La structure est donc partiellement bien prédite sur ces 2 oligonucléotides et seuls les appariements à longue distance n'ont pas été retrouvés. Cette observation peut également être faite avec plusieurs structures du jeu de données d'ARN, en fonction de l'outil de prédiction utilisé.

#### 5.4.3 Structures obtenues de complexes

Les oligonucléotides du jeu de données issus de complexes avec une protéine font partie des structures les plus difficiles à traiter en raison des interactions avec la protéine qui peuvent impacter la structure secondaire. Le jeu de données totalise 238 structures d'oligonucléotides qui sont issues de complexes avec une protéine. Les performances de prédictions sur ce type d'oligonucléotides sont variables selon l'outil sélectionné. En se focalisant sur les résultats des deux outils qui ont donné respectivement le plus haut et plus faible taux de prédictions correctes (MXfold2 et MC-Fold), 86 sur 238 prédictions exactes (~36 %) sont obtenues avec MXfold2 contre 20 sur 238 (~8 %) pour MC-Fold. Avec d'autres outils, le nombre de prédictions exactes est variable, avec 85 sur 238 pour avec RNAfold (modèle ARN par défaut), 79 sur 238 avec mfold (modèle ARN par défaut), 79 sur 238 avec Linearfold (-C et -V), 77 sur 238 avec CONTRAfold, 59 sur 238 avec CentroidFold, 54 sur 238 avec Ufold et 42 sur 238 avec SPOT-RNA.

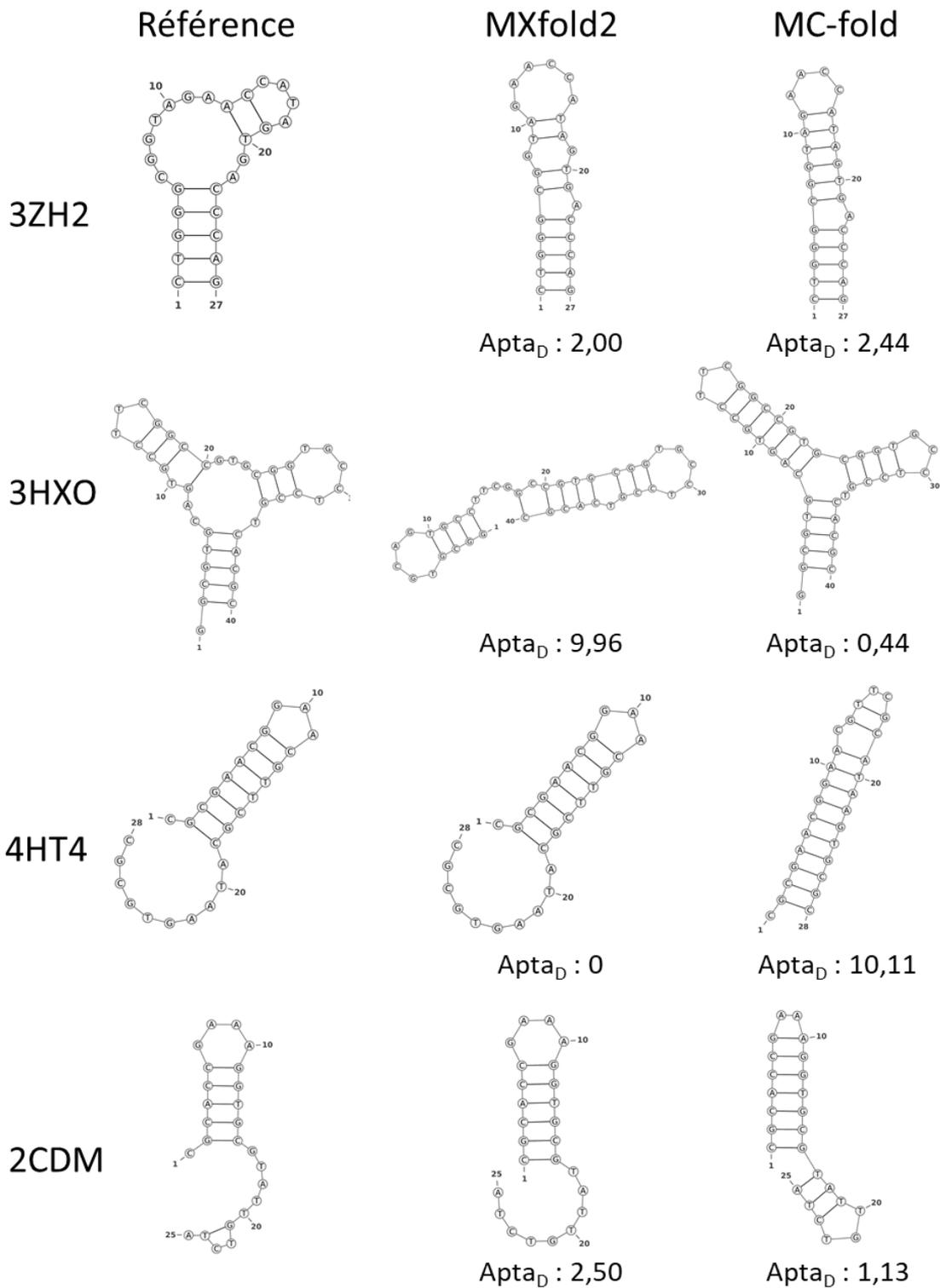


Figure 2-17. Exemples de 4 oligonucléotides en complexes avec la structure de référence et les différences de prédictions entre MXfold2 et MC-Fold ainsi que les valeurs *Apta<sub>D</sub>* associées.

En incluant les prédictions similaires ( $Apta_D \leq 1,5$ ) on retrouve la structure de 206 sur 238 (~86 %) oligonucléotides avec MXfold2, contre 122 sur 238 (~51 %) avec MC-Fold. UFold et SPOT-RNA parviennent à prédire respectivement 197 et 200 structures (~85 %) avec une



Ce problème est également transposable aux structures ARN. 3 exemples de longues structures issues de complexes ont été étudiés (Figure 2-18), soit 2XXA, 3WC1 et 5TF6 long de respectivement 102, 73 et 71 nucléotides. 2XXA est un oligonucléotide structuré en épingle à cheveux/tige-boucle avec 3 bourgeons et 3 boucles internes asymétriques. Pour cette structure, seuls MXfold2 et Ufold sont capables de retrouver le repliement exact. 3WC1 est un exemple de jonction à 3 branches qui n'est bien prédit que par Ufold. Tout comme 2XXA, RNAfold et mfold ne retrouvent pas la structure secondaire et les deux tendent à suggérer une structure en jonction à 2 branches, suggestion appuyée par d'autres outils (MXfold2, Linearfold-C et CONTRAfold). Enfin 5TF6 est mal prédit par tous les outils, car la plupart suggèrent la présence d'une jonction à 2 branches tandis que la structure expérimentale renvoie une épingle à cheveux/tige-boucle avec une absence de paire de bases entre les nucléotides 11 jusqu'à 29. Cependant, compte tenu de la longueur de ces 3 séquences la stabilité de la structure expérimentale hors complexe est discutable.

L'inclusion des structures sous-optimales dans l'analyse avec mfold et RNAfold a contribué à l'amélioration du nombre de prédictions identiques pour les oligonucléotides en complexe avec une protéine, surtout pour ce qui concerne les prédictions de RNAfold. L'analyse des 5 prédictions sous-optimales permet de retrouver la structure de 119 sur 238 oligonucléotides en complexe (modèle ARN par défaut) et jusqu'à 127 sur 238 oligonucléotides parmi les 10 prédictions sous-optimales.

Toutes les différences observées entre les prédictions de structure secondaire d'oligonucléotides d'ADN/ARN en complexe avec une protéine par rapport à la conformation expérimentale confirment le défi apporté par ce type de prédiction. La présence d'interactions entre l'oligonucléotide et la protéine favorise le maintien de conformations métastables à l'état libre qui ne sont pas nécessairement retrouvées par les outils de prédiction. Ainsi, les méthodes de prédictions incluant les solutions sous-optimales permettent de partiellement retrouver ces états métastables.

## 5.5. Prédictions de G-quadruplexes

Les structures présentant un G-quadruplexe sont mal prédites dans leur quasi-intégralité avec les approches sélectionnées car les appariements Hoogsteen ne sont pas des appariements canoniques. De plus, ceux-ci dépendent de la présence de cations stabilisateurs alors que la

plupart des outils n'intègrent pas de paramètres pour la concentration saline. Une unique structure, 5CMX composée d'une hélice et d'un G-quadruplexe a obtenu une prédiction acceptable avec SPOT-RNA et mfold présentant une distance  $Apta_D$  de 1,11 et 1,44 respectivement face à la référence. Cette distance correcte s'explique par une excellente prédiction des paires de bases impliquées dans l'hélice, alors que l'ensemble des appariements Hoogsteen impliqués dans le G-quadruplexe n'ont pas été identifiés par les algorithmes.

RNAfold, avec le paramètre pour la prédiction des G-quadruplexes, est cependant capable d'identifier la position de ce type de structure. Afin d'analyser la fiabilité de cette approche, le coefficient de Tanimoto modifié, présenté en Partie 2.5.1, a été utilisé pour la comparaison. Pour rappel, deux structures sont identiques si elles ont un score de Tanimoto de 1 et opposées si le score est de 0. Sur 27 structures disposant d'un repliement G-quadruplexe, 6 sont parfaitement identifiées avec un score de Tanimoto = 1. En étendant l'analyse aux structures bien prédites au seuil de Tanimoto 0,7, ce sont 18 structures (soit 62 %) dont les G-quadrets et les appariements canoniques ont été correctement prédits. En outre, parmi les 15 oligonucléotides composés de G-quadret et d'appariements Watson Crick, 11 affichent un coefficient de Tanimoto modifié supérieur à 0,7 (Annexe 7).

## 6. Discussion

AptaMat a permis d'obtenir les distances entre structures prédites et structures de référence avec une sensibilité accrue par rapport aux outils communément utilisés. Il est important de correctement estimer l'impact réel d'une variation d'appariement sur la structure secondaire qui peut être le résultat d'une ou plusieurs mutations ou simplement lié à la nature dynamique des oligonucléotides et l'existence de conformations métastables. Là où RNAdistance ou la distance BP peuvent parfois estimer certaines différences comme similaires vis-à-vis de la référence, AptaMat parvient à correctement classer les structures. De plus,  $Apta_D$  est une distance normalisée qui permet de fixer des seuils de proximité. Au-delà de la simple comparaison entre la référence et la prédiction, à partir d'un ensemble de prédictions correspondant aux solutions sous-optimales de mfold ou de l'association de plusieurs outils, la distance  $Apta_D$  peut être calculée pour en extraire les structures secondaires proches. Associé à un protocole de clustering, AptaMat est capable de différencier des familles d'ARN

issus de Rfam ainsi que nous l'avons démontré. Il est également possible d'obtenir une tendance de repliement qui se caractérise par plusieurs structures qui diffèrent peu (faibles valeurs  $Apta_D$ ). Dans le cas où une association d'outils de prédiction serait utilisée, la proximité des structures peut favoriser l'identification d'une structure consensus ou un ensemble de conformations possibles pour la séquence.

AptaMat a donc été utilisé pour tester les performances de 9 outils de prédiction de la structure secondaire nucléotidique (mfold, RNAfold, CentroidFold, CONTRAfold, Linearfold, MC-Fold, MXfold2, Ufold, SPOT-RNA) sur le jeu de données décrit en Partie 2.1. Les résultats de cette étude comparative ont montré que les outils basés sur la minimisation de l'énergie libre (i.e. mfold ou RNAfold) restent de bons choix. Ils présentent parmi les plus hautes proportions de prédictions identiques, de 43 à 49 % avec mfold en ne considérant que la structure de minimum d'énergie, et jusqu'à 88 % avec RNAfold en prenant en compte les structures sous-optimales, qui reflètent la variabilité de conformations accessibles autour de l'énergie libre minimale. D'autres modèles non-thermodynamiques sont prometteurs comme les approches basées sur du Machine Learning (CONTRAfold) ou celles implémentant des modèles de Deep Learning (MXfold2, Ufold). Ces approches sont basées sur les connaissances de la structure des ARN et restent cependant recommandées pour ce type de molécules.

Les difficultés communes à certains outils dans la prédiction des structures secondaires peuvent être justifiées par les cas particuliers qui composent le jeu de données. Les petits oligonucléotides ( $\leq 15$  nucléotides) possèdent des repliements parfois difficiles à prédire car résultant de conditions spécifiques pour l'obtention de la structure, comme les "*minidumbbell*".

Un autre type de motif difficile à prédire est le pseudonœud, car il implique des interactions à longue distance et seulement 3 parmi les 9 outils choisis sont explicitement capables de manipuler ce motif. Néanmoins, les approches basées sur des modèles de *Deep Learning*, notamment Ufold et SPOT-RNA, ont des bonnes performances vis-à-vis de la prédiction des pseudonœuds, avec 61 et 59 structures prédites avec une distance  $Apta_D \leq 1,5$ , respectivement.

Les G-quadruplexes forment un motif particulier difficile à prédire en raison de l'implication de 4 nucléotides interconnectés dans une tétrade. Ainsi, un G-quadruplexe peut être replié

selon différents schémas récemment classifiés par Popena et al. (2020) qui catégorise les G-quadruplexes en fonction de la topologie, le sens et les interactions entre guanines qui impactent le repliement de l'oligonucléotide. Le seul outil connu capable de localiser les sites de formation des G-quadrets est RNAfold et ses capacités se limitent à la prédiction de la position des sites riches en guanine potentiellement impliqués dans la formation d'un G-quadruplex sans être capables d'identifier les appariements. Il existe cependant d'autres outils plus récents capables de prédire la position et les interactions probablement formées dans un G-quadret (Rocher et al., 2021). L'intérêt récent porté à ce type de structure et leur comportement (Collie & Parkinson, 2011) encourage donc à envisager leur intégration dans de futurs travaux pour accroître la diversité de motifs disponibles pour la production d'oligonucléotides à des fins thérapeutiques ou diagnostiques. De la même façon, les pseudonœuds représentent le défi de la prédiction de structure secondaire et de nombreuses équipes travaillent sur des outils adaptés à la prédiction de ces motifs (Jabbari et al., 2018; Marchand et al., 2022; Reidys et al., 2011). Nos observations ont permis de conclure que pour l'instant les outils de prédiction conventionnels indiquent les nucléotides impliqués dans ces motifs sont prédits comme non-appariés.

L'existence de solutions et d'innovations ayant permis la prédiction de ces différents motifs rend envisageable la création d'un nouvel outil capable de prédire une combinaison d'appariements canoniques, de pseudonœuds, ou de G-quadruplexes. Avec les outils actuels, il est également possible de mettre en place un *workflow* qui compilerait les informations de repliements. Ce *workflow* permettrait de construire une structure probable en se basant sur i) les appariements partagés dans les structures sous-optimales ii) la probable formation de pseudonœuds sur certaines positions iii) l'identification de G-quadruplex pour les oligonucléotides riches en guanines. En marquant les guanines impliquées dans un G-quadret avec des contraintes bloquant toute interaction, il serait possible de retrouver les repliements les plus complexes.

Au-delà des motifs particuliers, la prédiction des structures secondaires obtenues pour les oligonucléotides en complexe avec une protéine est critique. En effet, la présence d'une protéine lors de l'obtention de la structure peut impacter la structure secondaire de l'oligonucléotide car la présence d'une interaction peut médier la formation de certains appariements (Davlieva et al., 2014). Ceci met également en lumière l'existence de différentes

conformations en raison de la mobilité naturelle des oligonucléotides prise en compte au travers des méthodes MEL, qui sont capables de retrouver jusqu'à 86 % des structures expérimentales en incluant entre 5 et 10 structures sous-optimales. Malgré les bonnes performances observées pour la plupart des oligonucléotides en complexe, plusieurs cas montrent de fortes différences ( $Apta_D \geq 1,5$ ) avec la structure de référence. Cela peut être dû au fait que la protéine peut stabiliser une conformation métastable de l'oligonucléotide (Hoetzel & Suess, 2022). Ces conformations sont potentiellement difficiles à prédire, ce qui explique les différences entre structure prédite et expérimentale, particulièrement dans les cas où un grand nombre d'outils ne parvient pas à retrouver la structure secondaire.

Enfin, il a été possible d'observer les limites des outils dans la prédiction des structures des oligonucléotides d'ADN. RNAfold et mfold sont les seuls outils qui offrent la possibilité de spécifier le type d'oligonucléotide et de choisir le modèle thermodynamique en fonction de cela. Néanmoins, les résultats obtenus en utilisant les modèles développés pour l'ARN sont meilleurs, ce qui suggère la nécessité d'optimiser les paramètres thermodynamiques propres à l'ADN. Les performances des autres outils CentroidFold et Linearfold vis-à-vis des ADN sont variables, mais les approches machine-learning et deep learning comme CONTRAfold, MC-Fold, MXfold2, Ufold et SPOT-RNA ont toutes montré des performances correctes en considérant qu'elles ont été créées pour la prédiction des structures d'ARN. Les bases de données de structures d'ADN simple brin étant limitées, ces approches ne peuvent bénéficier de données d'apprentissage exclusivement ADN. Néanmoins, il semble envisageable de pouvoir améliorer ces outils pour une application sur tout type d'acide nucléiques simple brin.

Les résultats obtenus ici ont fait l'objet d'une publication qui vient d'être acceptée dans BMC bioinformatics.



## Partie 3. Echantillonnage conformationnel des oligonucléotides d'ADN à simple brin

Au vu des limites identifiées dans la prédiction des structures secondaires des oligonucléotides et en raison de l'importance de tenir compte non seulement de l'appariement des bases, mais aussi de l'orientation spatiale des nucléotides pour une bonne compréhension de la fonction de ce type de molécules, une modélisation tridimensionnelle des oligonucléotides est fortement recommandée. Elle permet de mieux visualiser les interactions à longue distance et potentiellement de résoudre les erreurs de prédiction de la structure secondaire en évitant les appariements impossibles. En outre, si la modélisation 3D est effectuée en générant plusieurs conformations possibles, la flexibilité intrinsèque des oligonucléotides peut être examinée.

Pour ces raisons, la suite de ces travaux a consisté en l'optimisation d'un protocole basée sur la dynamique moléculaire pour la modélisation tridimensionnelle et l'échantillonnage conformationnel des oligonucléotides. Pour cette partie, seuls les oligonucléotides d'ADN ont été étudiés, et ce pour plusieurs raisons : tout d'abord, la prédiction de la structure secondaire de ce type d'oligonucléotide s'est révélée plus problématique. Il est donc intéressant de vérifier si l'utilisation des dynamiques moléculaires peut améliorer les prédictions obtenues. De plus, ces travaux s'insèrent dans un projet plus ample visant à développer une procédure pour la conception *de novo* d'oligonucléotides à simple brin capables de reconnaître des protéines d'intérêt. Pour ce projet, on s'intéresse à l'ADN à simple brin, car, comme décrit dans la Partie 1, il est plus stable par rapport à l'ARN et, de plus, le laboratoire d'accueil a une expertise sur la sélection d'aptamères à ADN reconnaissant des protéines (Avalle-Bihan et al., 2015; Loussouarn, 2014).

### 1. Méthodes de prédiction de structure tridimensionnelle

Les outils de prédiction de la structure tridimensionnelle des oligonucléotides à simple brin ont été développés pour la modélisation d'ARN et ils se basent essentiellement sur 2 approches complémentaires et utilisées simultanément avec différents degrés de prédominance dans chaque outil : les approches *knowledge-based* et l'échantillonnage stochastique. Les premières utilisent les connaissances sur la structure tridimensionnelle des

oligonucléotides acquises au travers des bases de données publiques et fonctionnent parfois par assemblage de sous-éléments structuraux. L'échantillonnage stochastique, souvent réalisé par l'application de la méthode Monte Carlo (Hénon, 1971), applique les lois de la biophysique pour explorer différentes conformations d'une même structure (repliée ou non) avec, pour objectif, la minimisation de son énergie libre.

Parmi les approches *knowledge-based*, on trouve la modélisation par homologie utilisée pour la prédiction de protéine et acides nucléiques. Cela consiste à aligner la séquence à modéliser avec une base de données de séquences dont la structure est connue, et à rechercher une ou plusieurs séquences qui puissent servir de modèle pour prédire le repliement de la séquence à modéliser. Plusieurs outils s'appuyant sur cette stratégie ont été développés, notamment ModeRNA (Rother et al., 2011) et RNAbuilder (Flores et al., 2010).

Selon le même principe, les approches par assemblage de fragments utilisent des fragments courts, extraits des données structurales connues, qui sont ensuite assemblés pour reconstruire la structure complète. Ces fragments peuvent être de toute taille en fonction de la base de données, en partant d'un simple enchaînement de 2 nucléotides jusqu'à des hélices entières. En cas d'absence de similarités entre une partie de la séquence à modéliser et les fragments reportés dans les bases de données, les atomes du fragment nucléotidique sont générés à partir de coordonnées cartésiennes compatibles avec les atomes et liaisons qui composent le fragment. La taille des fragments est définie par l'algorithme utilisé. Par exemple, l'approche adoptée par RNAComposer (Popena et al., 2012) consiste à traiter les structures secondaires comme des blocs indépendants extraits d'une base de données de fragments, la RNA FRABASE (Popena et al., 2010). Dans sa version plus récente, cette base de données reporte les informations de 2753 structures expérimentales et peut inclure des fragments de toute taille. Ainsi, RNAComposer se base sur le constat que ces éléments se répètent parmi les molécules d'ARN, et que deux fragments de même taille et présentant les mêmes appariements disposeront très probablement de la même topologie. Par conséquent, RNAComposer prend, en entrée, la structure primaire et secondaire d'un oligonucléotide d'ARN et assemble les éléments structuraux tertiaires dans la base de données qui coïncident avec les structures primaires et secondaires de l'ARN à modéliser pour reconstruire sa structure tertiaire. La structure tridimensionnelle résultante est ensuite minimisée au cours

de deux étapes de minimisation par gradient conjugué, appliquées en utilisant CHARMM (Popenda et al., 2012).

Un autre outil basé sur l'assemblage de fragments suivi d'une minimisation d'énergie est RNAAdenovo, qui est l'outil de modélisation des ARN dans la suite Rosetta (Das & Baker, 2007). RNAAdenovo intègre l'algorithme d'assemblage de fragments FARFAR (pour *Fragment Assembly of RNA with Full Atom Refinement*) (Cheng, Chou, and Das 2015), ou FARFAR2 (Watkins, Rangan, and Das 2020) selon la version. A partir des structures primaire et secondaire de l'ARN à modéliser, cet algorithme cherche dans une base de données de structures expérimentales des courts fragments de 1 à 3 nucléotides qui correspondent à ceux de la séquence à modéliser. Il les assemble ensuite pour former la chaîne de l'oligonucléotide en prenant en compte les contraintes imposées par la structure secondaire de l'ARN. L'assemblage de ces courts fragments résulte en une multitude de solutions possibles, permettant alors l'obtention de plusieurs conformations. Enfin, des simulations Monte Carlo de courte durée sont effectuées afin de minimiser l'énergie des modèles produits par l'approche FARFAR.

Une approche similaire à celle de RNAAdenovo est proposée au sein de MC-Sym (Parisien & Major, 2008), dont le fonctionnement est dépendant de MC-Fold. MC-Sym prend en entrée la séquence de l'ARN à modéliser. La prédiction de sa structure secondaire est effectuée par MC-Fold et, ensuite, il effectue la recherche et l'assemblage de fragments issus de la PDB et de la NDB d'un maximum de 6 nucléotides disposant d'informations sur leur structure secondaire et tertiaire. Ainsi MC-Sym est dépendant du résultat de MC-Fold et de sa fonction de score pour déterminer le repliement tridimensionnel optimal.

SimRNA (Boniecki et al., 2015) est un représentant des approches incluant en majorité la prédiction par échantillonnage aléatoire Monte Carlo. Comme la plupart des outils, il prend en entrée la séquence nucléotidique et éventuellement la structure secondaire de l'ARN à modéliser. Ensuite, la structure de l'ARN est décrite en utilisant un modèle de type gros grain, selon lequel chaque nucléotide est représenté par 5 grains correspondant aux atomes P, C4', C1, C3 et N4 pour les pyrimidines ou N9 pour les purines. La structure est alors construite selon un simple assemblage de grains centrés autour de C4' permettant la reconstruction du squelette-phosphate en respectant les contraintes d'angles du squelette et celles induites par

la structure secondaire (si existante). Pour l'obtention de la structure finale, SimRNA utilise l'approche Monte Carlo en simulation simple ou en échange de répliques. Cette dernière permet de lancer plusieurs simulations et d'échanger des états conformationnels entre ces répliques exécutées en parallèle avec des variations de température pour favoriser les changements d'état et permettre un échantillonnage plus efficace par rapport à la simulation Monte Carlo simple.

Parmi les outils disponibles et abordés ci-dessus, trois ont été sélectionnés afin d'être comparés comme représentant d'une catégorie d'outil. Ainsi, RNAComposer et RNAdenovo ont été testés comme outils basés sur des approches knowledge-based et SimRNA a été testé en tant qu'approche basée sur l'échantillonnage aléatoire.

## 2. Prédiction d'une structure tridimensionnelle

### 2.1. Sélection des données

Afin d'effectuer la comparaison des approches de prédiction sélectionnées, le jeu de données créé pour la comparaison des structures secondaires est également exploitable. Comme expliqué dans l'introduction de la Partie 2, les 67 oligonucléotides à ADN sans G-quadruplexes ont été sélectionnés pour cette étude (Tableau 2). Il est intéressant de noter que les 3 outils choisis ont été développés pour la prédiction de la structure tridimensionnelle des ARN et de nombreux travaux vantent les qualités de prédiction de RNAdenovo (Cheng, Chou, and Das 2015), SimRNA (Boniecki et al., 2015) et RNAComposer (Popenda et al., 2012) appliquées sur ce type d'oligonucléotide. Par conséquent, le choix de se focaliser uniquement sur les ADN à simple brin permettra également de vérifier s'il est possible d'étendre le domaine d'applicabilité de ces outils à ce type de molécule.

### 2.2. Outils et protocole de prédiction

Les trois outils sélectionnés, prennent en entrée un fichier en format fasta qui peut inclure la structure secondaire au format dot-bracket si disponible afin de guider la prédiction des repliements.

Pour RNAComposer, les calculs sont exécutables sur un serveur en libre accès<sup>6</sup> qui ne dispose d'aucun paramètre ajustable. RNAdenovo utilise la prédiction de la structure tridimensionnelle par assemblage de fragments FARFAR inclus dans la suite ROSETTA 3.6 (Das & Baker, 2007). Lors de son utilisation, les paramètres par défaut ont été gardés. Cela implique que 800 structures sont générées afin de disposer d'un nombre conséquent de conformations différentes, et elles sont minimisées au cours de 20 000 cycles Monte Carlo. Les structures sont ensuite classées selon l'énergie estimée et les 10 meilleures structures sont conservées.

Pour l'utilisation de SimRNA, les paramètres recommandés dans le manuel ont été fixés. Ainsi les poids des liaisons, des angles et des torsions qui impactent l'énergie du système sont ceux indiqués dans le Tableau 7. La diminution du facteur relatif à la température de 1,35 à 0,9 permet de favoriser l'exploration de différentes conformations. Le long de la simulation Monte Carlo, l'acquisition de la structure est effectuée toutes les 16 000 étapes pour un total de 16 000 000 itérations et permet d'acquérir un total de 1 000 conformations, soit un nombre proche de celui de RNAdenovo pour des temps de calculs inférieurs.

Pour SimRNA seule la structure à énergie minimale a été conservée. Dans le cas de RNAdenovo, en plus de la structure d'énergie minimale estimée, les 9 prédictions additionnelles ont été analysées. Enfin pour RNAComposer l'unique structure générée est conservée.

Pour les trois outils, une structure secondaire a été fournie en entrée avec la séquence nucléotidique. Ce choix permet de rendre la procédure homogène, car RNAComposer nécessite forcément une structure secondaire en entrée. Par conséquent, pour chaque outil et chaque ADN étudié, deux prédictions ont été effectuées. Une première a été guidée par la structure secondaire expérimentale afin d'évaluer la qualité de prédiction en supposant une prédiction exacte de la structure secondaire. Elle sert donc d'étalonnage de la qualité de prédiction. Puis une deuxième prédiction a été effectuée en fournissant comme structure secondaire en entrée une prédiction obtenue de l'un des outils de prédiction testés dans la Partie 2. Le choix s'est porté sur mfold, en raison de ses performances satisfaisantes dans la prédiction de structures secondaires. Cela permet d'évaluer la pertinence d'un protocole construit sur l'enchaînement de la prédiction de structure secondaire puis tertiaire en évaluant l'impact des différences de la structure secondaire suivie par la prédiction de

---

<sup>6</sup> [rnacomposer.cs.put.poznan.pl](http://rnacomposer.cs.put.poznan.pl)

structure tertiaire. De plus, il est possible de vérifier si des problèmes de prédiction de la structure secondaire sont résolus en passant à la structure tridimensionnelle. Dans le cas où la structure secondaire prédite par mfold est identique à la structure secondaire, une seule prédiction a été effectuée.

Tableau 7. Paramètres appliqués aux outils de prédiction RNAComposer, RNAdenovo et SimRNA

<b>RNAComposer</b>	<b>RNAdenovo</b>		<b>SimRNA</b>	
	Cycle Monte Carlo	20 000	Cycle Metropolis-Hastings	16 000 000
	Structures produites	800	Acquisition par n itérations	16 000
	ARN minimisé	Oui	Facteur température Initiale	1,35
	Hélices fixes	Oui	Facteur température Finale	0,9
	Structures générées	10	Poids de liaison	1
			Poids des angles	1
			Poids des torsions	0
			Poids d'angle $\eta + \theta$	0.4

### 2.3. Optimisation des modèles tridimensionnels obtenus

RNAdenovo, SimRNA et RNacomposer produisent une ou plusieurs structures de type ARN. Ainsi, si les oligonucléotides à modéliser sont des ADN à simple brin, comme c'est le cas dans cette étude, les prédictions obtenues doivent être modifiées pour avoir le type d'oligonucléotide voulu. Cela a été effectué manuellement avec PyMol (Schrödinger, 2015). En particulier, pour transformer le ribose en désoxyribose, le groupement hydroxyle en C2',

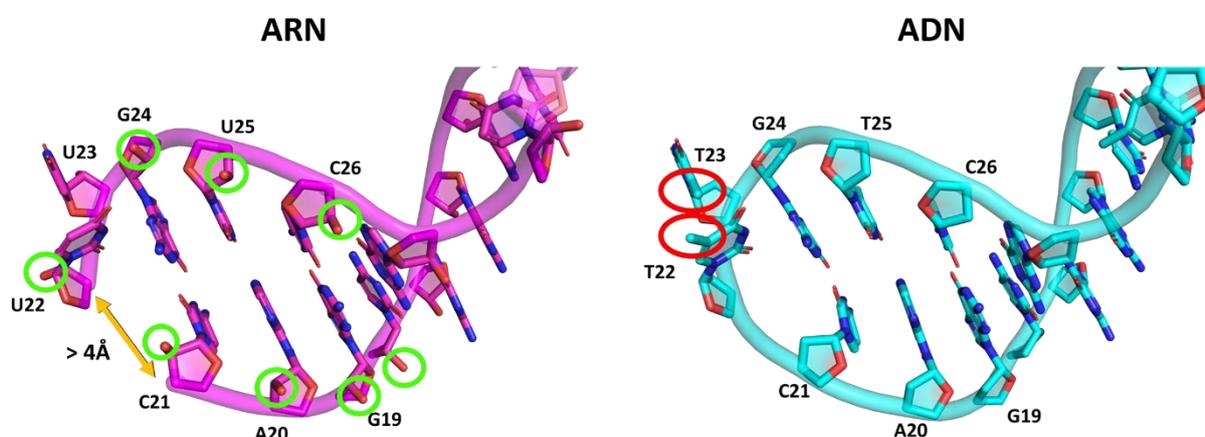


Figure 3-1. Comparaison de l'épingle à cheveux G19-C26 de 1EZN ARN généré par RNAdenovo (gauche), et ADN après transcription inverse minimisé avec Amber (droite) représentés sur PyMOL. Les groupements 2'OH présent sont entouré en vert. Les groupement 5-méthyl sont entouré en rouge. La distance séparant C21 et U22 est indiquée par la flèche orange.

entouré en vert dans la Figure 3-1, a été remplacé par un hydrogène. Les bases azotés uracile ont toutes été modifiées en thymine grâce à l’outil « mutagénèse » intégré dans PyMOL, ce qui n’impacte pas la position des atomes mais entraîne l’apparition d’un groupement 5-méthyl entouré en rouge dans la Figure 3-1.

De plus, les structures issues de la prédiction avec RNAAdenovo présentent parfois des césures du squelette sucre-phosphate comme montré sur l’exemple en Figure 3-1 avec un écart de 4 Å entre les nucléotides C21 et U22. Il a donc été nécessaire recréer manuellement la liaison manquante avec PyMol et d’effectuer une minimisation supplémentaire pour assembler correctement les deux portions séparées. En outre, en général, le changement des thymines en uraciles induit l’apparition d’un groupement plus encombrant pouvant engendrer des clashes stériques. Une étape de minimisation additionnelle a donc été effectuée. Cela a été réalisé *in vacuum*, en utilisant Amber20 (D.A. Case, T. Kur et al., 2020) pour toutes les structures tertiaires de RNAAdenovo, SimRNA ou RNAComposer. Une première minimisation des hydrogènes est appliquée avec 4000 cycles de descente de gradient suivis jusque 1000 cycles de gradient conjugué, puis la structure totale est minimisée sur 16 000 cycles de descente de gradient et jusqu’à 4000 cycles de gradient conjugué.

## 2.4. Mesure de la déviation Quadratique Moyenne des prédictions face à l’expérimental

Pour évaluer les performances des outils de prédiction de la structure tridimensionnelle, il est nécessaire de comparer la structure prédite à celle obtenue expérimentalement, qui constituera donc la référence.

La méthode choisie pour évaluer la variabilité entre deux structures tridimensionnelles est la déviation quadratique moyenne (ou RMSD pour *Root Mean Square Deviation* en anglais). Elle est définie selon l’équation 14 :

$$RMSD = \sqrt{\sum_{i=1}^n (S_{ref} - S_{pred})^2} \text{ (Équation 14)}$$

$S_{ref}$  désigne un ensemble des coordonnées atomiques de la structure de référence, et  $S_{pred}$  le même ensemble dans la structure prédite. Elle donne une mesure de la distance en Å séparant les atomes de la structure référence et les atomes correspondant de la structure prédite. Cette mesure a été obtenue en utilisant le module *cpptraj* de Amber20 en comparant

exclusivement les positions des atomes composant le squelette phosphate, soit les atomes P, O5', C5', C4', C3' et O3'.

Afin d'accorder une certaine liberté dans le positionnement des cycles composant le nucléotide, la mesure de la RMSD a été effectuée en comparant exclusivement les positions des atomes composant le squelette phosphate, soit les atomes P, O5', C5', C4', C3' et O3' (Figure 3-2).

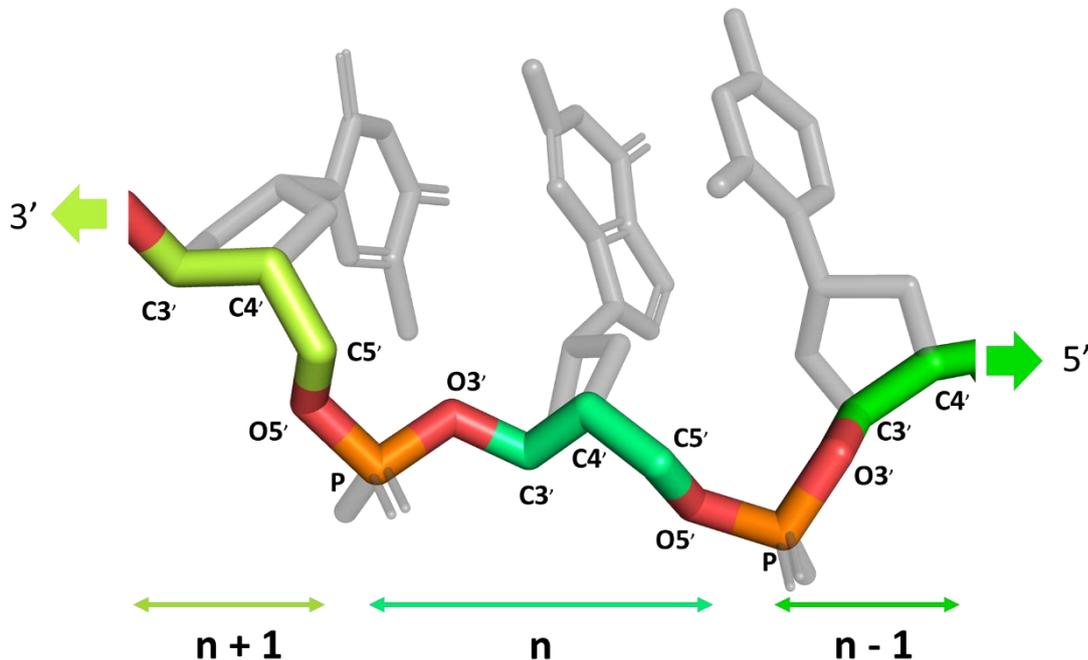


Figure 3-2. Sélection des atomes du squelette phosphate d'une chaîne nucléotidique représentée sur PyMOL. n désigne un nucléotide.

Bien qu'il n'existe pas de seuil idéal pour qualifier une RMSD satisfaisante, la valeur seuil a été fixée à 5 Å au regard de l'analyse globale des résultats obtenus ainsi que les résultats présentés dans d'autres études de modélisation des acides nucléiques simple brins, qui définissent des tendances médiocres d'alignement à partir de 4 Å, voire 7 Å selon les cas étudiés (De Beauchene et al., 2016; Ropii et al., 2023). Ce seuil, défini comme intermédiaire, permet de valider la stabilité d'une structure tout en identifiant les différences liées à la flexibilité des oligonucléotides.

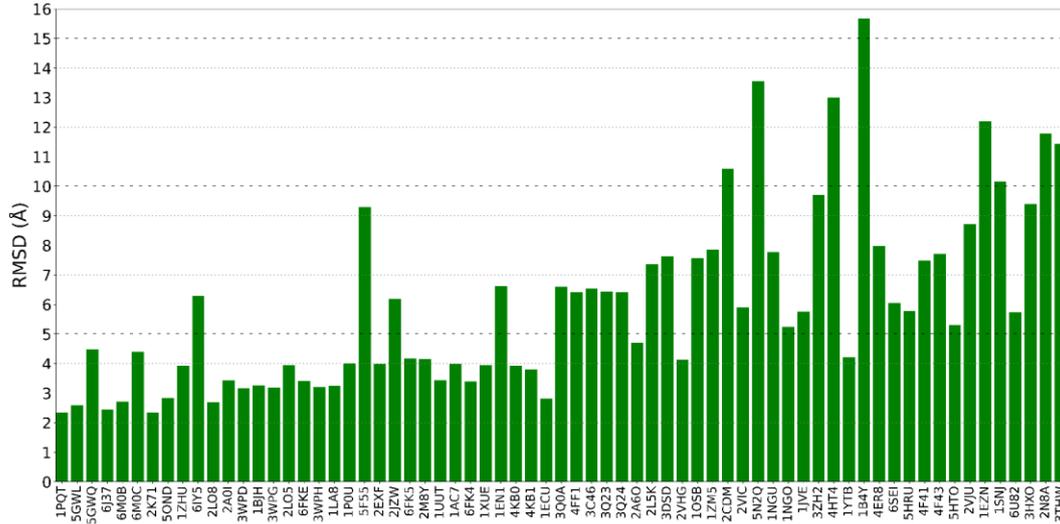
### 3. Comparaison des prédictions

#### 3.1. Prédiction des structures secondaires de référence

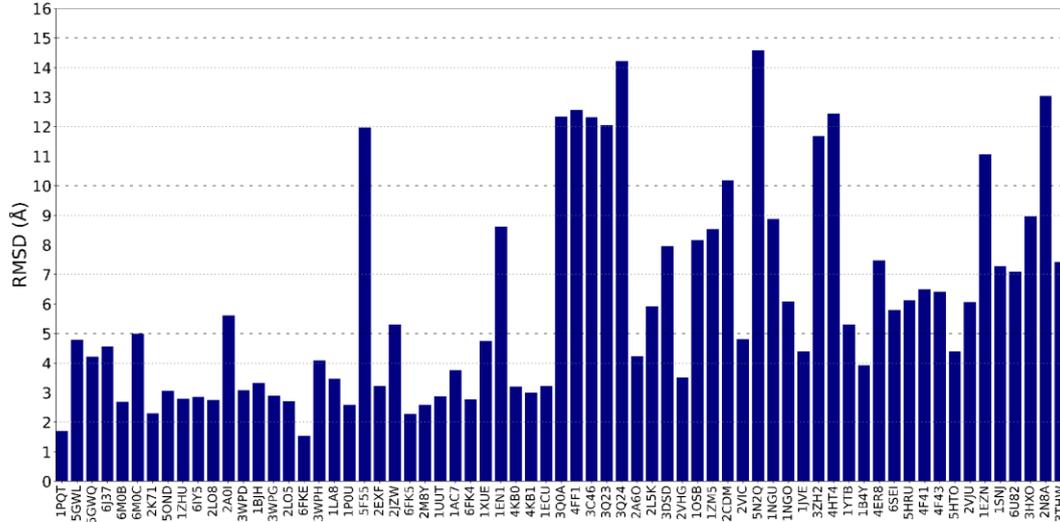
Dans un premier temps, les prédictions avec RNAComposer, SimRNA et RNAdenovo ont été générées en utilisant comme guide la structure secondaire de référence. La RMSD par rapport à la référence a été calculée pour les 67 prédictions des structures du jeu de données avec les trois outils. La Figure 3-3 présente l'ensemble de ces résultats sous forme de diagramme en barre, où chaque barre se rapporte à la RMSD calculée par rapport à la structure de référence pour chaque oligonucléotide. RNAdenovo produit 10 prédictions classées par énergie libre pour chaque oligonucléotide. Donc, dans ce cas, la RMSD moyenne des 10 prédictions et l'écart-type ont été indiqués dans la Figure 3-3.

RNAComposer, SimRNA et RNAdenovo prédisent 32, 35 et 33 structures (soit 48 %, 52 % et 49 %), respectivement, avec une RMSD inférieure au seuil de 5 Å par rapport à la référence expérimentale. Une majorité de ces bonnes prédictions (29, 28 et 29 pour RNAComposer, SimRNA et RNAdenovo, respectivement) correspond à des oligonucléotides de longueur < 20 nucléotides. Les performances sont donc relativement équilibrées entre les trois outils lorsqu'il s'agit de prédire la structure d'oligonucléotides relativement courts. Les différences se limitent à quelques structures : 5GWQ, 6MOC et 2A0I, bien que les valeurs de RMSD restent proches de la valeur seuil de 5 Å pour les trois outils, et 6IY5, dont la prédiction par RNAComposer montre une RMSD de 6,28 Å, quand SimRNA et RNAdenovo affichent respectivement 2,85 Å et 3,13 Å. De plus les prédictions de RNAdenovo permettent de retrouver la structure de 2L5K et 1YTB, avec une distance respective de 4,16 Å et 4,05 Å par rapport à la structure expérimentale, à l'opposé de SimRNA dont les prédictions dépassent le seuil de 5 Å (5,92 Å et 5,31 Å). Toujours sur ces deux structures, les prédictions de RNAComposer montrent des distances de 7,37 Å et 4,22 Å impliquant donc une mauvaise prédiction pour 2L5K et une bonne prédiction pour 1YTB. Enfin, dans le cas de 2VHG, RNAComposer et SimRNA retrouvent des repliements similaires à la structure expérimentale avec 4,12 Å et 3,51 Å respectivement, quand la prédiction de RNAdenovo montre une distance de 6,12 Å. Ainsi, dans plusieurs situations au moins un des trois outils est capable de retrouver un repliement tridimensionnel proche de la référence au seuil de 5 Å.

(a) RNAComposer



(b) SimRNA



(c) RNAdenovo

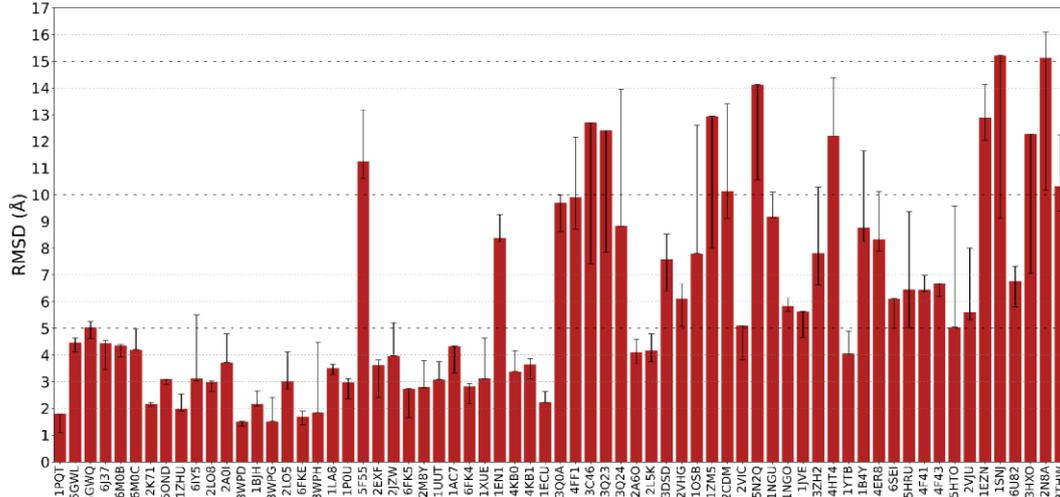


Figure 3-3. Diagrammes en barre de la valeur de RMSD pour les prédictions obtenues avec (a) RNAComposer, (b) SimRNA et (c) RNAdenovo. La longueur des séquences est croissante de gauche à droite. Les barres d'erreurs pour (c) RNAdenovo indiquent les valeurs minimales et maximales obtenues sur les 10 solutions, la barre colorée indiquant la valeur de RMSD pour la prédiction d'énergie minimale.

De façon similaire, les mêmes oligonucléotides (28 au total, 5F55, 1EN1, 3Q0A, 4FF1, 3C46, 3Q23, 3Q24, 3DSD, 1OSB, 1ZM5, 2CDM, 5N2Q, 1NGU, 1NGO, 3ZH2, 4HT4, 4ER8, 6SEI, 5HRU, 4F41, 4F43, 2VJU, 1EZN, 1SNJ, 6U82, 3HXO, 2N8A et 3THW) sont souvent mal prédits par les trois outils. Parmi eux, 26 sont des longs oligonucléotides ( $\geq 20$  nucléotides). Au-delà de la longueur, la présence de motifs complexes peut affecter la précision des prédictions. La présence de bourgeons et de boucles internes peut avoir le même effet car ces derniers entraînent une flexibilité augmentée dans le reste de la structure tridimensionnelle. La prédiction de la structure de 7 oligonucléotides (1NGU, 2VJU, 3THW, 3ZH2, 4ER8, 6SEI et 6U82) est ainsi négativement affectée par la présence de ces motifs.

En outre, les jonctions font partie des structures montrant de fortes valeurs de RMSD, avec 1EZN, 1SNJ, 3HXO et 2N8A qui dépassent tous les 7 Å de distance face à la structure de référence peu importe l'outil utilisé (RNAComposer, SimRNA, RNAdenovo). Enfin, la présence de portions de plus de 4 nucléotides sans appariements est responsable d'un haut degré de mobilité et donc de conformations possibles. C'est le cas par exemple de 1EN1, 3Q0A, 4FF1, 3C46, 3Q24, 3DSD, 1OSB, 1B4Y ou 4ER8 (également impacté par la présence d'une boucle interne), dont les valeurs de RMSD sont élevées en dépit de l'utilisation de la structure secondaire de référence. En procédant à l'alignement de ces structures avec la référence exclusivement sur la portion structurée, on constate au travers de la Figure 3-4 que le positionnement des nucléotides non structurés n'est pas aligné entre prédiction et référence. La RMSD calculée entre les portions impliquées dans des motifs impliquant des appariements de bases (Figure 3-4) est meilleure que l'alignement de la structure totale sur ces 9 structures (Annexe 8). La haute RMSD est donc le résultat de la forte déviation des portions libres. On remarque dans le cas de 1B4Y que l'absence de structure secondaire sur une portion n'exclue pas la formation d'interactions à longue distance stabilisant la structure. Ainsi, la structure de référence ne présente aucune portion libre car les nucléotides de T1 à T10 sont impliqués dans la formation d'une triple hélice (Van Dongen et al., 1999).

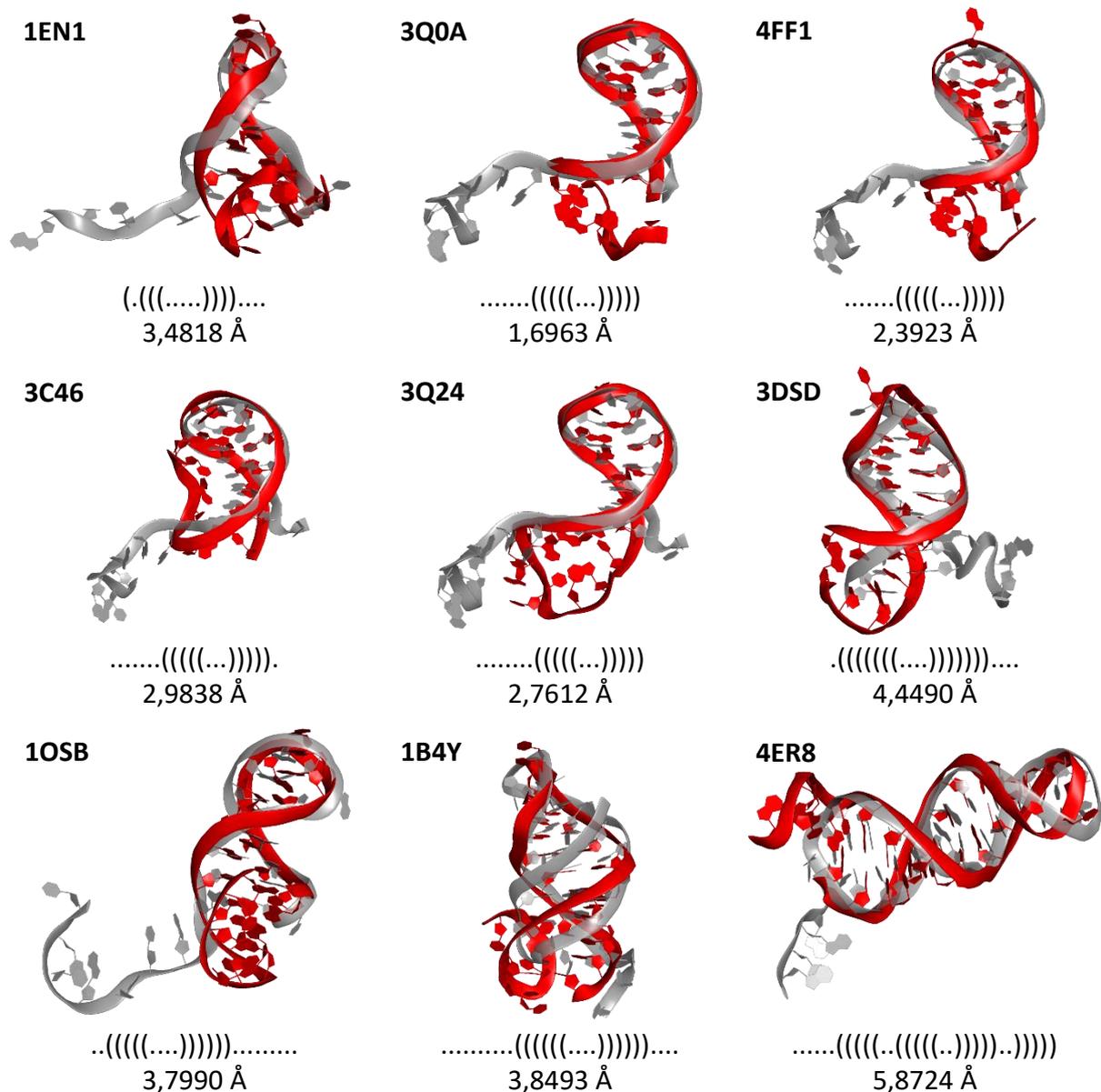
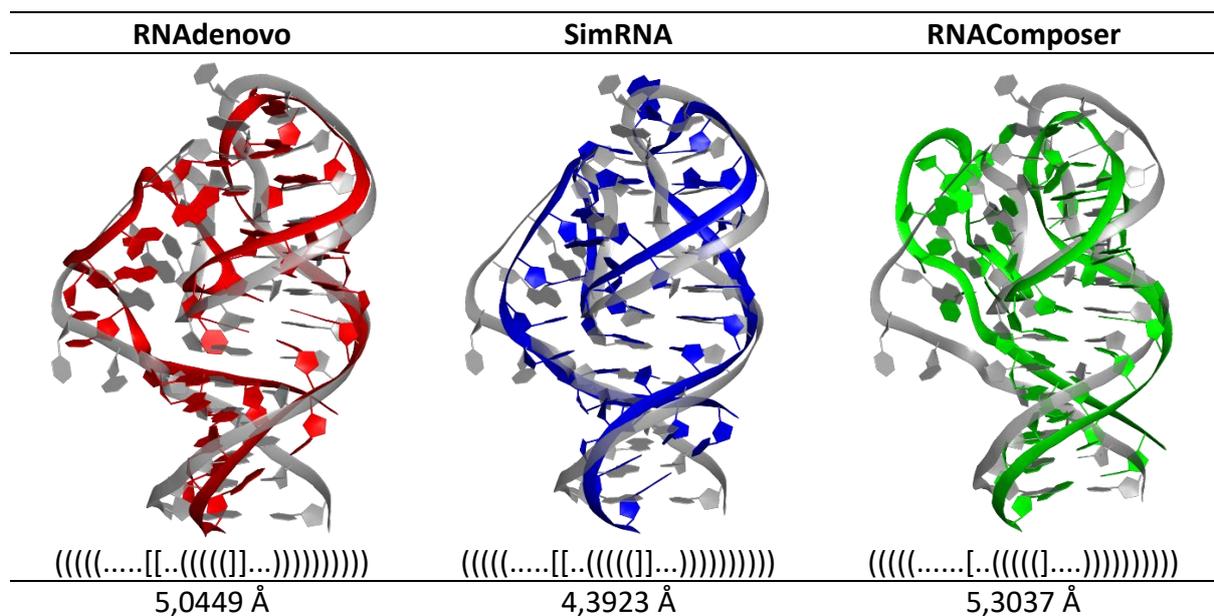


Figure 3-4. Alignement de la portion structurée (incluse dans une hélice ou une boucle) du squelette phosphate de 9 structures d'énergie minimale obtenues avec RNAdenovo (en rouge) avec la structure de référence (en gris) représenté sur PyMOL. La RMSD est mesurée pour chaque alignement. La structure secondaire dot-bracket est reportée à titre indicatif.

Pour ce qui concerne les pseudonœuds, on ne rencontre pas dans la prédiction 3D les mêmes difficultés eues pendant l'étape de prédiction de la structure secondaire. En effet, lorsque la structure secondaire exacte est utilisée pour guider la prédiction tridimensionnelle, les pseudonœuds peuvent être correctement prédits, car les trois outils sont compatibles avec l'incorporation de la notation dot-bracket étendue pour inclure les interactions à longue distance de ce type de motif. Ainsi, pour 5HTO on obtient une RMSD par rapport à l'expérimentale de 5,05 Å, 5,30 Å et 4,39 Å avec RNAdenovo, RNAComposer et SimRNA, respectivement, quand la structure secondaire expérimentale guide la prédiction (Figure 3-5a).

(a) Modèles obtenus à partir de la structure secondaire expérimentale

(((.....[.(((((]...)))))))))



(b) Modèles obtenus à partir de la structure secondaire expérimentale

(((.....[.(((((]...)))))))))

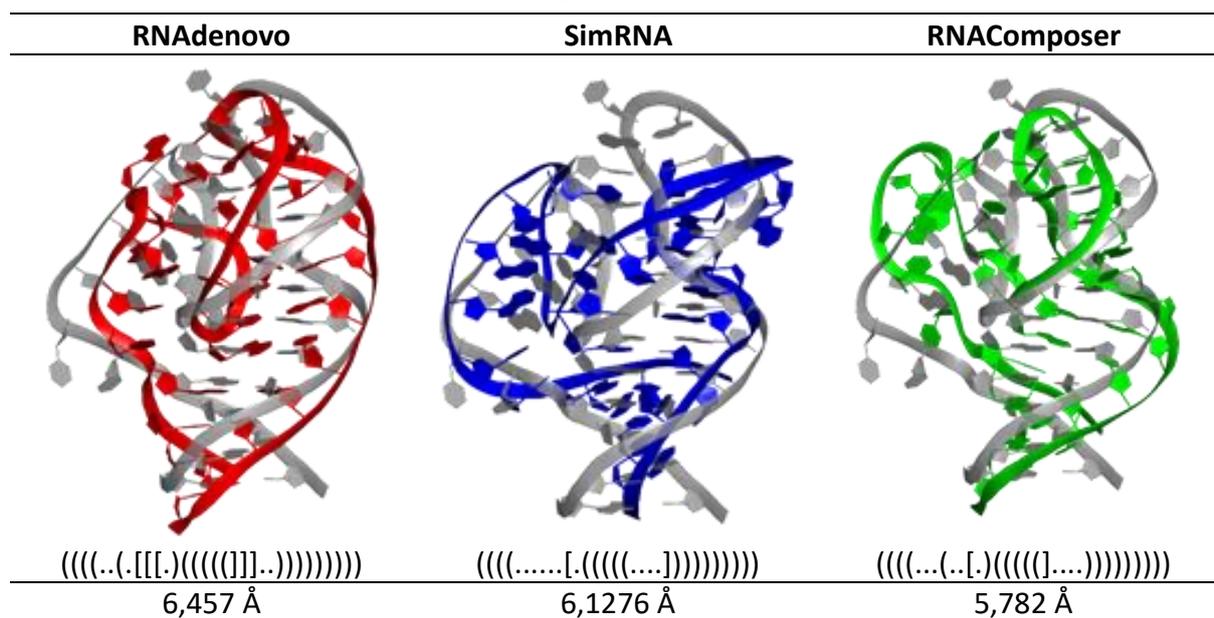


Figure 3-5. Alignement des prédictions de structure de (a) 5HTO et (b) 5HRU obtenues avec RNAdenovo (en rouge), SimRNA (en bleu) ou RNAComposer (en vert) avec la structure de référence (en gris) représentés sur PyMOL. La structure secondaire utilisée pour guider les prédictions est celle obtenue de la structure expérimentale. La RMSD est mesurée pour chaque alignement. La structure secondaire engendrée par les repliements est indiquée à titre indicatif.

Dans le cas de 5HRU, l'utilisation de la structure secondaire exacte résulte en des prédictions plus distantes de la structure expérimentale, avec des RMSD de 6,46 Å, 5,78 Å et 6,13 Å avec RNAAdenovo, RNAComposer et SimRNA respectivement (Figure 3-5b).

Pour RNAAdenovo, des observations complémentaires peuvent être obtenues en considérant les 10 prédictions sous-optimales obtenues pour chaque ADN. Globalement, la moyenne de RMSD des 10 prédictions de faible énergie est inférieure à la RMSD de la structure d'énergie minimale montrée en Annexe 8. Les écarts-type vont de 0,5 à 2,6 Å, mais dans la plupart des cas (81 %) ils sont inférieurs à 1 Å, ce qui indique la proximité des conformations prédites. En outre, la structure indiquée par RNAAdenovo comme celle à énergie minimale montre une RMSD inférieure au seuil de 5 Å ou, en tout cas, proche de la valeur moyenne. Cela suggère que, dans la plupart des cas, il est possible de se fier à la structure à énergie minimale pour les prédictions obtenues par RNAAdenovo. Néanmoins, dans certains cas (1PQT, 6J37, 2EXF, 6FK5, 1AC7, 3C46, 3Q23, 1ZM5, 2CDM, 1JVE, 6SEI, 1SNJ, 6U82 ou 2N8A) les valeurs des solutions alternatives proposées par RNAAdenovo montrent une proximité avec la structure de référence supérieure à celle de la structure d'énergie minimale comme les structures. Ainsi, dans 2 situations, les prédictions alternatives permettent de passer sous le seuil de 5 Å, comme avec les prédictions de 2VIC, avec la prédiction d'énergie minimale indique une RMSD de 5,088 Å face à la structure expérimentale contre 3,883 Å pour la prédiction la plus proche. De même, la prédiction RNAAdenovo d'énergie minimale de 1JVE montre une RMSD de 5,63 Å face à la structure expérimentale contre 4,67 Å pour la structure la plus proche.

En conclusion, RNAComposer, SimRNA et RNAAdenovo semblent être équivalents dans la prédiction de la structure tridimensionnelle des ADN à simple brin étudiés ici et, globalement, partagent les mêmes limites. Plus en détails, les modèles obtenus par les 3 outils montrent une augmentation de la RMSD des modèles face à la structure de référence qui semble corrélée à la longueur de séquence. Cela est attendu, car une séquence nucléotidique longue aura un nombre de degrés de liberté élevé qui impactera la phase de minimisation énergétique ou de recherche conformationnelle implémentée dans les 3 outils. De plus, il est attendu, pour de petites structures, de présenter une RMSD face à leur référence plus faible que les valeurs habituellement considérées comme satisfaisantes (Hajdin et al., 2010).

Si on regarde les performances des 3 outils, sur les 30 structures de longueurs > 20 nucléotides, on ne dénombre que 3 modèles pour RNAComposer et 5 modèles pour SimRNA qui montrent une faible RMSD (< 5Å). Avec RNAde novo, il est possible d'atteindre 6 modèles de grande taille proches de la structure résolue expérimentalement, si les 10 solutions proposées sont considérées, bien qu'il soit recommandé pour la prédiction de structures de fragments relativement courts d'ARN (Cheng, Chou, and Das 2015).

De plus, la présence de motifs (bourgeons, boucles internes et jonctions) ou de portions d'oligonucléotides flexibles augmente la difficulté de la prédiction.

### 3.2. Prédiction à partir des structures secondaires obtenues par mfold

En général, lorsque l'on cherche à modéliser la structure d'un oligonucléotide, sa structure secondaire n'est pas connue. L'utilisation de SimRNA ou RNAde novo sans fournir une structure secondaire guide est possible bien que les prédictions puissent résulter en une forte incertitude en raison de l'absence de contrainte. Pour cette raison, il est recommandé d'effectuer préalablement une prédiction de l'appariement des bases avec un outil de prédiction de structure secondaire. Il est néanmoins possible que la structure secondaire fournie pour la prédiction de la structure tridimensionnelle ne soit pas exacte. Dans cet esprit, une deuxième prédiction de la structure tridimensionnelle a été effectuée en utilisant les structures secondaires générées par mfold. Seuls les ADN simple brin pour lesquels mfold a donné une structure secondaire avec une distance AptaMat > 0 de la structure secondaire expérimentale (32) ont été inclus dans cette deuxième prédiction.

Sur les 33 structures dont la prédiction se base sur la structure secondaire prédite de mfold, RNAComposer, SimRNA et RNAde novo sont parvenues à prédire 10, 12 et 11 structures (soit 30 %, 36 % et 33 %), respectivement, avec une RMSD inférieure au seuil de 5 Å par rapport à la référence expérimentale. Seules 8 structures (2LO8, 3WPG, 6FKE, 3WPH, 1XUE, 4KB0, 4KB1 et 1ECU) sont correctement prédites par tous les outils. Ces 8 structures correspondent à des oligonucléotides courts (< 20 nucléotides). Il y a donc moins de fiabilité dans les prédictions à partir de la structure secondaire de mfold. Néanmoins, certains oligonucléotides (5F55, 1EN1, 3C46, 3Q23, 3Q24, 1OSB, 1ZM5, 2CDM, 3ZH2, 4ER8, 5HRU, 4F41, 4F43, 2VJU, 1EZN, 1SNJ, 3HXO et 2N8A) dont le modèle obtenu à partir de la structure secondaire produite par mfold est éloigné de la structure expérimentale étaient aussi mal prédits en partant de la structure

secondaire correcte. La forte RMSD ( $> 5 \text{ \AA}$ ) n'est donc pas une conséquence directe de la distance AptaMat observée pour ces structures.

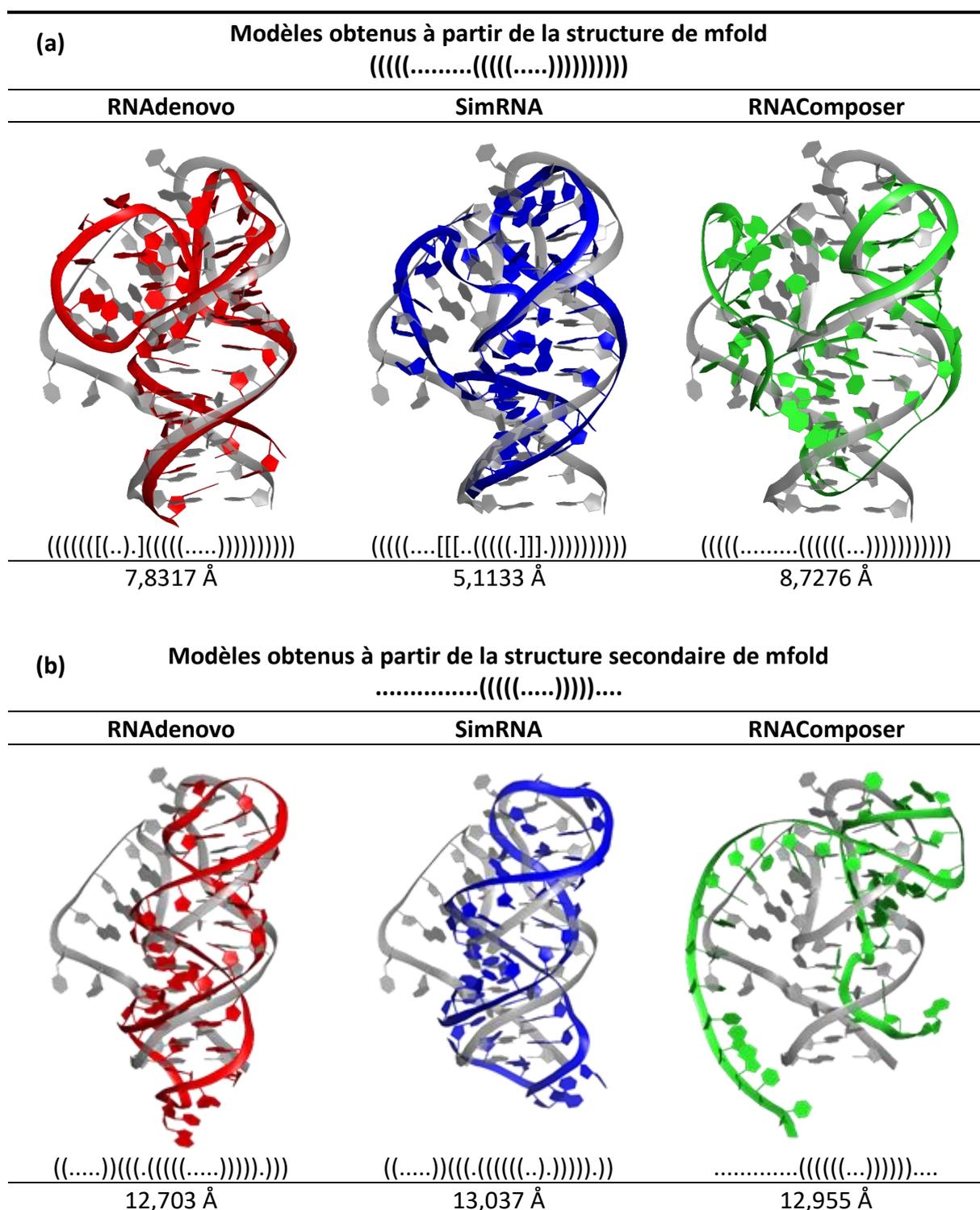


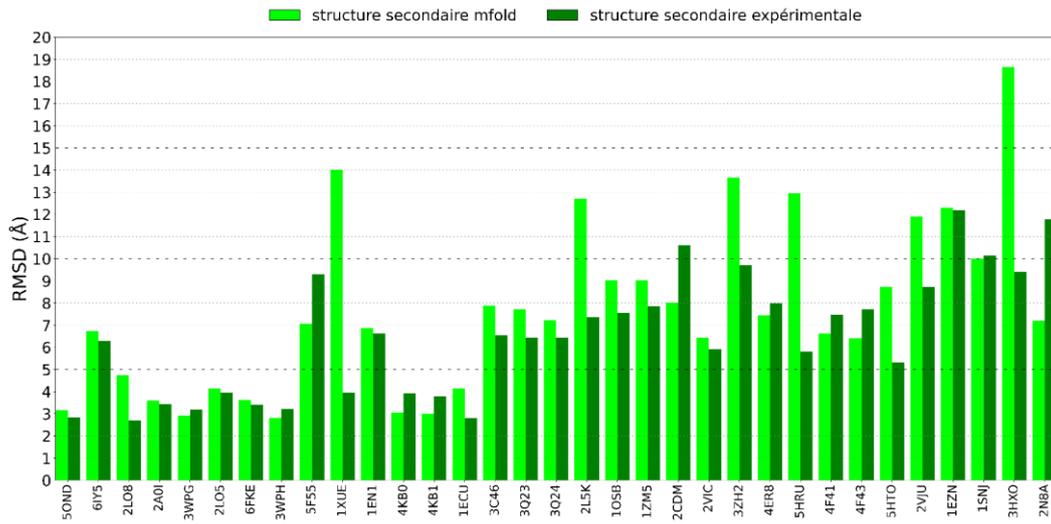
Figure 3-6. Alignement des prédictions de structure tertiaire de (a) 5HTO et (b) 5HRU obtenues avec RNAdenovo (en rouge), SimRNA (en bleu) ou RNAComposer (en vert) avec la structure de référence (en gris) représentés sur PyMOL. La structure secondaire utilisée pour guider les prédictions est celle obtenue avec mfold. La RMSD est mesurée pour chaque alignement. La structure secondaire engendrée par les repliements est indiquée à titre indicatif.

L'utilisation de la prédiction de mfold a donc permis de souligner l'impact d'une mauvaise prédiction de structure secondaire sur les résultats en prédiction tridimensionnel. 5HTO et 5HRU sont deux exemples qui illustrent ce problème car ils présentent des pseudonœuds non prédit par mfold, mais présentant des différences  $Apta_D$  avec leur structure de référence qui n'ont pas le même impact. Le prédiction mfold de 5HTO montre une  $Apta_D = 1,091$ , qui est considérée comme acceptable selon les critères définis en Partie 3.2.4. Ainsi les structures prédit par les montrent des valeurs de RMSD de 7,83 Å, 8,73 Å et 5,11 Å pour respectivement RNAdenovo, RNAComposer et SimRNA. Néanmoins, il est important de souligner que, même dans ces conditions, les trois outils sont parvenus à prédire correctement les repliements sur ~73 % de la structure. L'alignement de ces prédictions montré en Figure 3-6a permet d'identifier la partie de la chaîne plus fidèle à la structure de référence.

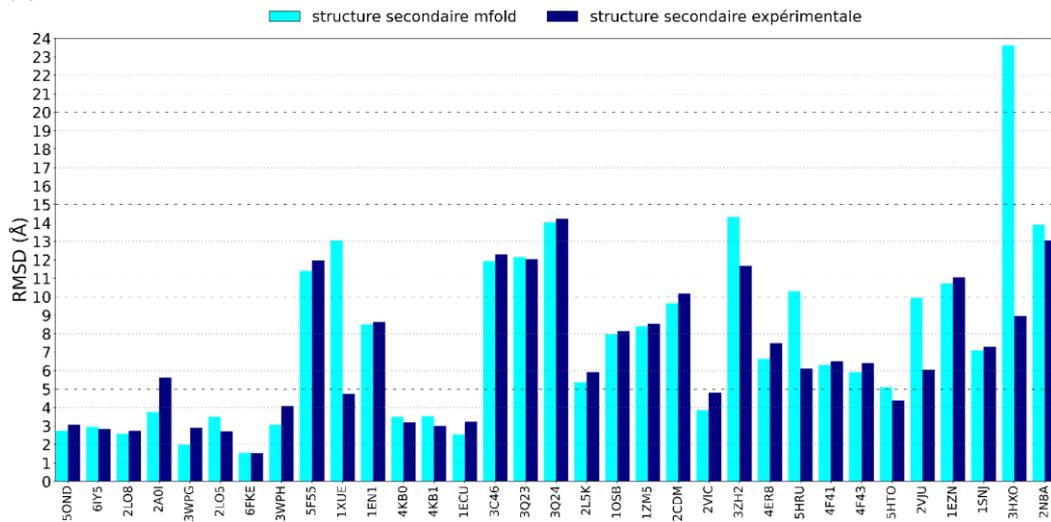
Dans le cas de 5HRU en raison de fortes différences entre la structure secondaire de référence et celle obtenue avec mfold, soit  $Apta_D = 5$ , il est cohérent d'observer de fortes variations de RMSD dans la prédiction de structure tertiaire. Les repliements obtenus atteignent ainsi des valeurs de RMSD dépassant le double par rapport aux prédictions avec structure secondaire de référence (Figure 3-5b), soit 6,46 Å, 5,78 Å et 6,13Å pour respectivement RNAdenovo, RNAComposer, et SimRNA (Figure 3-6b). Ainsi, les repliements obtenus avec RNAdenovo ou SimRNA intègrent les repliements contraints par la structure secondaire de mfold en entrée, auxquels sont ajoutés des appariements complémentaires stabilisant mais différents de la structure de référence. Pour la prédiction effectuée par RNAComposer en revanche, les repliements obtenus sont quasiment identiques à la structure secondaire prédite par mfold et, par conséquent, la structure tertiaire résultante ne montre aucune similitude avec la référence.

En outre, indépendamment de la valeur seuil de la RMSD par rapport à la structure expérimentale, ce n'est pas surprenant que la plupart des 33 structures prédites avec RNAdenovo et RNAComposer soit affectée négativement par le fait qu'une structure secondaire inexacte a été fournie : elles montrent une RMSD par rapport à la structure expérimentale plus élevée que celle obtenue par la prédiction faite à partir de la structure secondaire expérimentale (Figure 3-7, Figure 3-3).

(a) RNAComposer



(b) SimRNA



(c) RNAdenovo

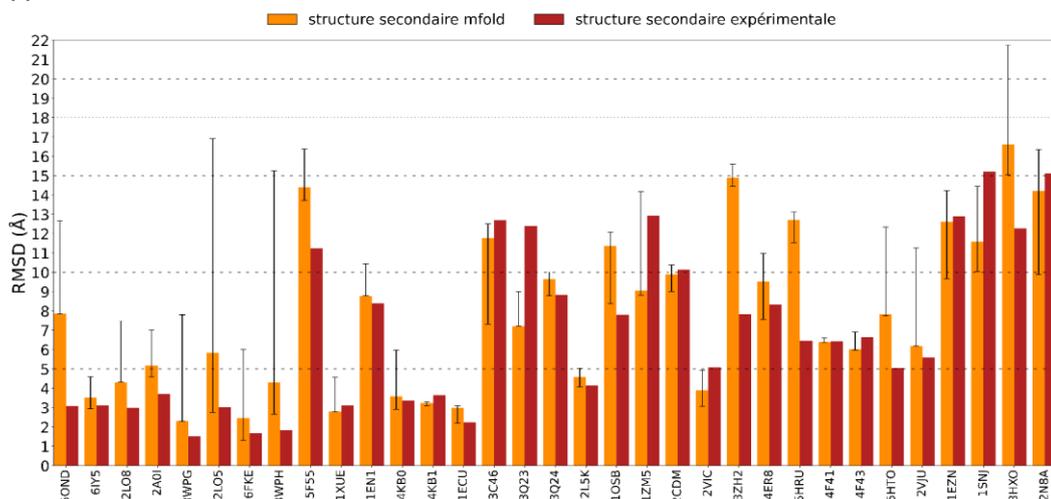


Figure 3-7 Diagrammes en barre de la valeur de RMSD pour les prédictions obtenues avec (a) RNAComposer, (b) SimRNA et (c) RNAdenovo en utilisant la structure secondaire prédit par mfold. La longueur des séquences est croissante de gauche à droite. Les barres d'erreurs pour (c) RNAdenovo indiquent les valeurs minimales et maximales obtenues sur les 10 solutions, la barre colorée indiquant la valeur de RMSD pour la prédiction d'énergie minimale.

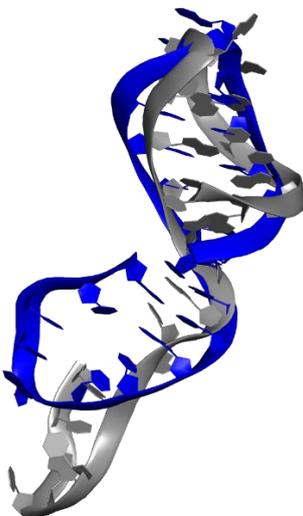
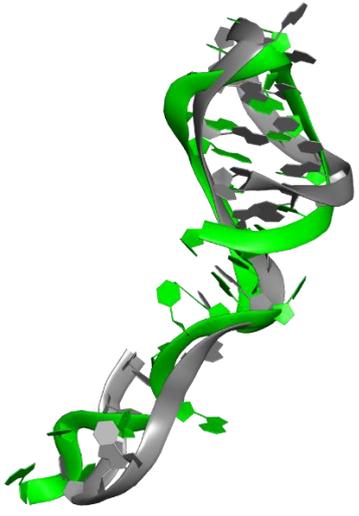
Structure secondaire en entrée ((((((.....)))))).....(..).		
RNAdenovo	SimRNA	RNAComposer
		
((((((.....))))))...(...)	((((((.....))))))((.....))	((((((.....)))))).....
9,889 Å	9,637 Å	8,0116 Å

Figure 3-8. Alignement des prédictions de structure de 2CDM obtenues avec RNAdenovo (en rouge), SimRNA (en bleu) et RNAComposer (en vert) avec la structure de référence (en gris) représentés sur PyMOL. La RMSD est mesurée pour chaque alignement. La structure secondaire engendrée par les repliements est indiquée à titre indicatif.

Plus en détails, RNAdenovo prédit 21 structures avec une RMSD supérieure à celle des structures obtenues en indiquant la structure secondaire exacte. Les 7 autres structures montrent des valeurs de RMSD légèrement inférieures malgré les variations de structure secondaire identifiées avec AptaMat. Pour 6 de ces structures (4KB0, 4KB1, 3Q24, 2VIC, 4F43 et 1EZN) la distance  $Apta_D$  est inférieure à 0,5, car leurs structures secondaires ne varient ici que d'un appariement par rapport à la référence. Si situés aux extrémités, ces appariements manquant ou additionnels peuvent offrir plus de flexibilité, ou à l'inverse des contraintes supplémentaires à la structure durant la prédiction et favoriser une conformation plus proche de l'expérimental. 2CDM est la seule exception, car sa structure secondaire montre une distance  $Apta_D$  de 2,5 par rapport à l'expérimentale, considéré comme élevée. Pour cette structure, la portion vers l'extrémité 3' présente un haut degré de flexibilité en raison de l'absence de bases appariées, résultant en un fragment de 9 nucléotides très mobile dans le cas de la prédiction de RNAComposer. Les prédictions de SimRNA et RNAdenovo en revanche semblent modéliser une structure plus compacte avec l'apparition d'appariements qui tendent à contraindre l'oligonucléotide au regard des structures tridimensionnelles obtenues.

L'explication de ce comportement peut provenir de la structure secondaire en entrée qui montre un appariement unique vers l'extrémité 3' et cette contrainte de repliement est prise en compte par RNAdenovo et SimRNA résultant en ces repliements après la minimisation par Monte Carlo (Figure 3-8).

Parmi les prédictions proposées par RNAComposer, 22 structures sont affectées négativement par les contraintes de structure secondaire inexacte, sans tenir compte de la valeur de RMSD par rapport à l'expérimentale. Parmi les 11 structures qui, au contraire, montrent une RMSD de l'expérimentale inférieure à celle obtenue pour les prédictions faites à partir de la structure secondaire de référence, 4 (4KB0, 4KB1, 4F43 et 2CDM) sont communes avec RNAdenovo. Les 7 restantes sont 3WPG, 3WPH, 5F55, 3Q24, 4ER8, 4F41 et 2N8A. Les 4 prédictions de la structure secondaire de 2N8A, 4F41, 4ER8 et 3Q24 ont une distance  $Apta_D < 0,5$ , ce qui suggère le même comportement que 4KB0, 4KB1, 3Q24, 2VIC, 4F43 et 1EZN. Pour 3WPH et 3WPG la structure secondaire prédite a une distance  $Apta_D = 1$  par rapport à l'expérimentale, qui est la conséquence d'une unique modification d'appariement, et peut être associée au même comportement que les précédents. La prédiction de la structure secondaire de 5F55 par mfold est très différente de celle expérimentale ( $Apta_D = 3$ ), et pourtant on observe une amélioration de la RMSD pour la structure obtenue à partir de la structure secondaire incorrecte (Figure 3-9), néanmoins cette RMSD reste supérieure à la valeur seuil de 5 Å. La structure tertiaire de 5F55 est probablement influencée par la présence d'une protéine, et la structure secondaire adoptée n'est pas celle attendue en forme libre. La forte déviation obtenue avec RNAdenovo et SimRNA peut s'expliquer par les interactions à longue distance favorisées lors de la prédiction tandis que la prédiction de RNAComposer résulte en un oligonucléotide disposant des mêmes contraintes comme le suggère la structure secondaire expérimentale.

Étonnamment, SimRNA est capable de trouver des structures plus proches de l'expérimentale que celles obtenues à partir d'une structure secondaire exacte. Ainsi, sans tenir compte de la valeur seuil de RMSD, seul 12 structures sont affectées négativement par les contraintes de structure secondaire inexactes (6IY5, 2LO5, 1XUE, 4KB0, 4KB1, 3Q23, 3ZH2, 5HRU, 5HTO, 2VJU, 3HXO et 2N8A). La différence est plus importante lorsque la distance  $Apta_D > 1$ . Les structures 1XUE, 3ZH2, 5HRU, 5HTO, 2VJU, 3HXO sont ainsi impactées avec une augmentation

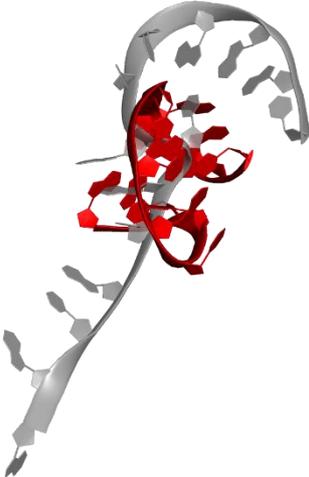
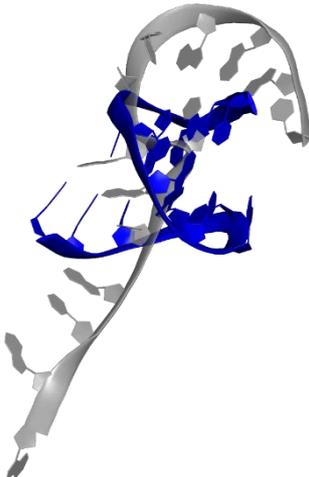
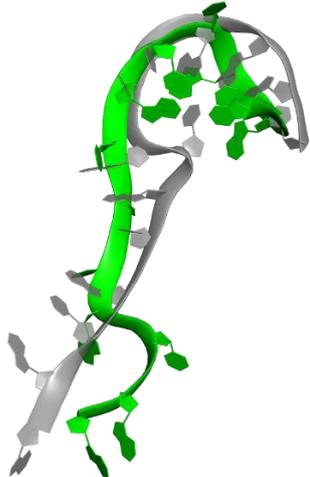
Structure secondaire en entrée .....((...))		
RNAdenovo	SimRNA	RNAComposer
		
.[.....(([.])	..[.....(([.])	.....((...))
14,4051 Å	11,4009 Å	7,0634 Å

Figure 3-9. Alignement des prédictions de structure de 5F55 obtenues avec RNAComposer (en vert), SimRNA (en bleu) et RNAdenovo (en rouge) avec la structure de référence (en gris) représentés sur PyMOL. La RMSD est mesurée pour chaque alignement. La structure secondaire engendrée par les repliements est indiquée à titre indicatif.

moyenne de la RMSD de 85 % par rapport à la prédiction SimRNA utilisant la structure secondaire expérimentale. 2CDM et 5F55 agissent comme des exceptions mais l'amélioration de la RMSD reste limitée, passant de ~12 Å à ~11 Å et ~10 Å à ~9,5 Å respectivement.

En outre, si on considère les oligonucléotides pour lesquels on observe des améliorations plus importantes de RMSD par rapport à l'expérimentale dans la prédiction à partir de la structure secondaire prédite par mfold, les oligonucléotides 2LO8, 2A0I, 5F55, 1XUE, 2L5K et 2CDM ont une structure secondaire prédite très différente de la référence ( $Apt_D \geq 1,5$ ), et montrent une meilleure RMSD qu'en partant de la structure de référence. Il est probable que ces améliorations par rapport aux modèles obtenus à partir de la structure secondaire exacte soient la conséquence de l'approche Monte Carlo adoptée par SimRNA pour l'échantillonnage des conformations. La minimisation de l'énergie recherchée implique que certaines contraintes de structures secondaires ne soient pas conservées. Cependant, ce comportement ne permet pas à ces structures d'atteindre une RMSD favorable ( $\leq 5$  Å), à l'exception de 2A0I, et les améliorations sont limitées avec une baisse moyenne de la RMSD de 8 % par rapport à la prédiction SimRNA utilisant la structure secondaire expérimentale. Au vu de ces résultats, il

est possible de conclure que, comme prévu, la qualité de la structure secondaire fournie en entrée de la prédiction de la structure tridimensionnelle des oligonucléotides affecte la qualité du modèle obtenu. Cet effet est moins évident pour les prédictions effectuées avec SimRNA, qui dépend de l'algorithme implémenté qui se base i) sur la construction de la structure avec un modèle gros grain, sans utiliser des fragments provenant des bases de données, et ii) sur des simulations Monte Carlo qui prévoient un nombre pas négligeable de cycles pour l'échantillonnage des conformations. Les outils de prédiction tridimensionnelle sont des méthodes qui ont l'avantage d'être faciles et rapides à utiliser. SimRNA calcul en ~1h les 16 000 000 itérations sur une structure de 20 nucléotides, RNAcomposer le fait en quelques secondes et RNAde novo, le moins rapide des trois, prédit 800 structures en ~1h en disposant de 40 processeurs. Les trois outils montrent des performances similaires en prédiction avec une RMSD moyenne des modèles oscillants entre 5,9 et 6 Å en utilisant la structure secondaire de référence. La moyenne augmente jusqu'à 6,5 Å si les structures secondaires de mfold sont utilisées. Au-delà des valeurs moyennes, on observe une tendance des trois outils à échouer à prédire les mêmes oligonucléotides, notamment les oligonucléotides de grande taille. Dans le choix du meilleur outil de prédiction de structure tridimensionnelle, il est attendu que la structure estimée comme celle de minimum d'énergie soit proche de celle attendue. Cette affirmation a été partiellement démentie par les résultats obtenus, notamment avec RNAde novo. En effet, en considérant les 10 structures de faible énergie, plusieurs prédictions estimées comme d'énergie sous-optimale sont plus proches de la référence que la première. C'est le cas dans 81 % des situations, incluant 3C46, 3HXO, 1SNJ et 2N8A dont les modèles ont une différence de RMSD de plus de 30 % la valeur de la structure classée comme celle à moindre énergie. La présence d'une forte variabilité de RMSD dans un ensemble de 10 prédictions d'énergie favorable, notamment pour les oligonucléotides les plus longs, suggère une forte flexibilité et donc une grande diversité de conformations. Ceci questionne le mode de sélection de la structure de référence. Néanmoins, il semble évident que tester l'intégralité ou même les 10 meilleures prédictions suggérées par RNAde novo serait très consommateur de temps. L'approche recommandée pour RNAde novo, abordée dans l'article de Cheng et al. (2015), est d'appliquer une étape de clustering sur l'ensemble des structures suggérées permettant de grouper les prédictions similaires en un unique cluster. L'intérêt est ainsi de limiter la redondance et la sur-représentation des solutions de plus faible énergie qui viendraient occulter les conformations alternatives existantes. Intégrer une étape

supplémentaire de clustering après la prédiction RNAde novo permettrait de mettre en valeur des conformations alternatives. Il serait envisageable de procéder à cette étape en utilisant l'approche proposée par TTclust qui s'est révélée efficace lors de l'analyse des dynamiques moléculaires dans l'étape qui suit. Rosetta RNAde novo continue d'implémenter des mises à jour qui pourraient impacter la précision de ses modèles dans le futur. Dans l'optique de maximiser la qualité des structures tertiaires modélisées avant dynamique moléculaire, SimRNA est une alternative démontrée comme viable au travers de nos résultats avec globalement un meilleur alignement des prédictions avec la structure résolue expérimentalement. Néanmoins, pour la suite de ces travaux, les modèles obtenus avec RNAde novo ont été utilisés. En se basant sur cela, il est possible d'étudier le comportement d'une hypothétique prédiction tridimensionnelle déviant de la structure expérimentale attendue pour appréhender les performances du protocole en situation non-optimale.

#### 4. Exploration de l'espace conformationnel des oligonucléotides par dynamique moléculaire à échantillonnage intensif

Une solution possible pour pallier les problèmes des outils de prédiction de la structure tridimensionnelle des oligonucléotides est d'utiliser des techniques de dynamique moléculaire à échantillonnage intensif qui impliquent l'échantillonnage de façon intensive de l'espace conformationnel des oligonucléotides, et qui peuvent donc générer un ensemble de conformations accessibles en solution par la molécule. Néanmoins, la dynamique moléculaire à partir d'une structure dépliée manque de rendement en raison des durées de simulations élevées (jusqu'à la ms) nécessaires pour atteindre la structuration et la convergence.

Les logiciels de prédiction tridimensionnelle testés précédemment ont contribué à l'obtention de structures tertiaires fiables pour les structures de petites tailles, mais sont majoritairement peu satisfaisants pour les structures les plus complexes. Ces méthodes permettent cependant de disposer d'une structure repliée et minimisée comme point de départ pour des méthodes d'échantillonnage comme la dynamique moléculaire. Une approche combinée, résumé en Figure 3-10 a donc été pensée pour bénéficier d'une structure d'énergie minimisée disposant de la structure secondaire attendue et qui servira de point de départ pour les simulations de dynamique moléculaire.

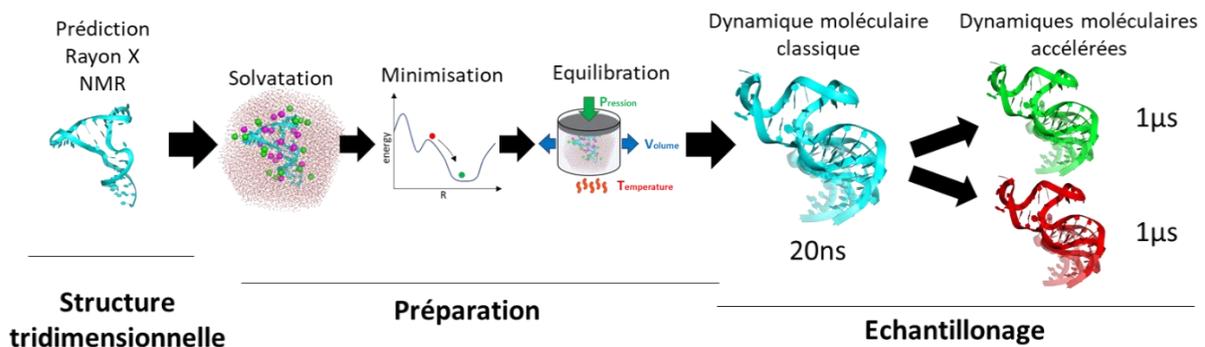


Figure 3-10. Etapes du protocole de dynamique moléculaire.

#### 4.1. Fondamentaux de la dynamique moléculaire

La dynamique moléculaire (MD) est une méthode computationnelle utilisée pour étudier l'évolution d'un système moléculaire en fonction du temps, en fournissant ainsi des informations, à une échelle atomique, sur les fluctuations et les changements conformationnels du système. Cette méthode est donc fréquemment appliquée pour l'étude de la structure, de la dynamique et de la thermodynamique des molécules biologiques et de leurs complexes.

La dynamique moléculaire se base sur la deuxième loi de Newton ou principe fondamental de la dynamique, qui s'exprime comme :

$$\vec{F}_i = m_i \vec{a}_i \text{ (Équation 15)}$$

Selon cette équation, la force  $\vec{F}_i$  exercée sur la particule  $i$  est calculée en fonction de sa masse  $m_i$  et l'accélération  $\vec{a}_i$ . Etant  $\vec{a}_i$  la dérivée seconde de la position des atomes  $\vec{r}_i$  en fonction du temps ( $t$ ), l'équation 16 peut s'écrire :

$$\vec{F}_i = m_i \frac{d^2 \vec{r}_i}{dt^2} \text{ (Équation 16)}$$

Par conséquent, à partir de la connaissance de la force exercée sur chaque atome, il est possible de déterminer l'accélération de chaque atome du système simulé. L'intégration de l'équation du mouvement sur des intervalles de temps courts (1 ou 2 fs) permet l'obtention d'une trajectoire qui décrit la variation en fonction du temps des positions des atomes, des vitesses et de l'accélération. La force  $\vec{F}_i$  peut être exprimée selon l'équation 17 comme un gradient du potentiel d'énergie  $V$  tel que

$$\Delta F_i = -\nabla_i V \text{ (Équation 17)}$$

Donc, en combinant les Equations 16 et 17, on obtient l'équation 18 :

$$\frac{-dV}{dr_i} = m_i \frac{d^2 r_i}{dt^2} \text{ (Équation 18)}$$

Les coordonnées initiales peuvent être obtenues par des méthodes d'acquisition de la structure tridimensionnelle, comme la cristallographie rayons X, la spectroscopie RMN, la cryo-électromicroscopie ou encore la modélisation 3D. L'évolution de ces coordonnées en fonction du temps définira la trajectoire des atomes après avoir défini des vitesses initiales et les forces appliquées.

En outre, la MD est une méthode basée sur la mécanique classique et, donc, elle repose sur l'approximation de Born-Oppenheimer, qui permet de considérer l'énergie potentielle exclusivement en fonction des coordonnées des noyaux atomiques. Cette approximation considère le noyau comme l'élément responsable des interactions dans un atome, car il représente 99,97 % de la masse d'un atome. Les électrons, beaucoup moins lourds, sont omis dans le calcul. Les atomes sont donc décrits comme une sphère avec type, un rayon, une masse et une charge positionnée au centre de la sphère, et les liaisons covalentes sont décrites comme des ressorts (Adcock & McCammon, 2006).

Il a été montré précédemment (Équation 17) que l'évolution du système dépendait du calcul de son énergie potentielle  $V(\vec{r})$ , qui peut tenir compte des interactions entre atomes covalentement liés et de celles entre atomes qui ne sont pas liés entre eux, car elle est fonction des positions atomiques. Ainsi pour un ensemble  $N$  particules dans leur position  $\vec{r}$  est connue, l'énergie de cet ensemble  $V(r^N)$  est communément calculée selon l'équation Hamiltonienne suivante :

$$V(r^N) = \sum_{\text{éléments liés}} K_l (l - l_0)^2 + \sum_{\text{angle}} K_\theta (\theta - \theta_0)^2 + \sum_{\text{dièdres}} K_\omega (1 + \cos(n\omega - \delta)) + \sum_{\text{torsions impropres}} K_\phi (\Phi - \Phi_0)^2 + \sum_{\text{éléments non liés}} \left\{ \epsilon_{ij} \left[ \left( \frac{R_{ij}^{\min}}{r_{ij}} \right)^{12} - \left( \frac{R_{ij}^{\min}}{r_{ij}} \right)^6 \right] + \frac{q_i q_j}{\epsilon_l r_{ij}} \right\} \text{ (Équation 19)}$$

Cette équation inclut 4 termes énergétiques qui décrivent les interactions interatomiques entre atomes liés, comprenant l'énergie associée à l'élongation/compression des liaisons et

celles associées à la torsion des angles, des dièdres propres et impropres, et des termes décrivant les interactions entre atomes pas directement liés, notamment les interactions électrostatiques et celles de van der Waals (Figure 3-11).

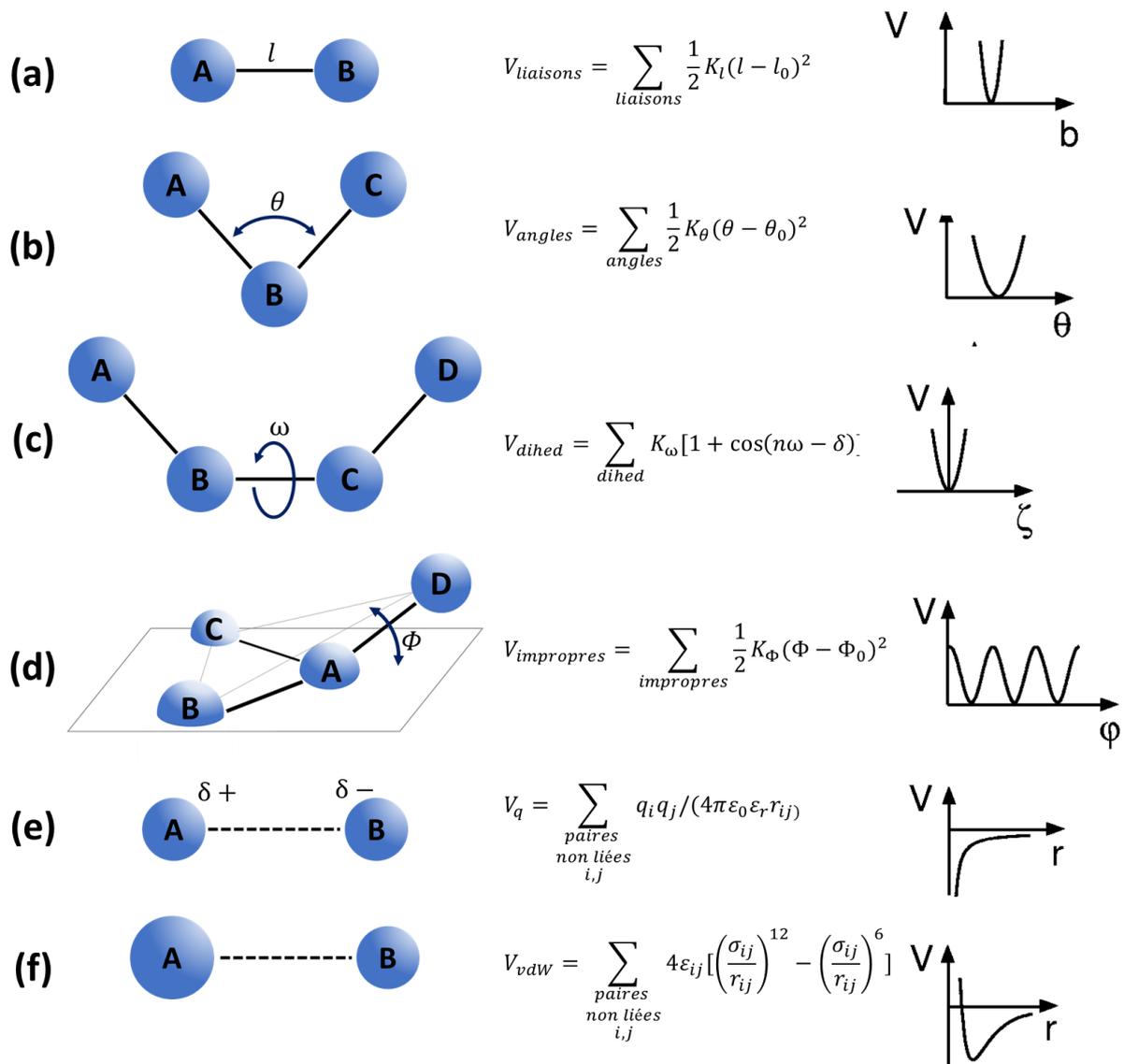


Figure 3-11. Type d'interactions entre atomes illustrant les termes de l'équation 19

Les liaisons covalentes étant décrites comme des ressorts, les termes énergétiques décrivant les interactions covalentes sont définis par la loi de Hooke pour les ressorts, qui établit que la force appliquée par un ressort est égale à l'étirement ou à la compression du ressort multiplié par la constante de rappel du ressort. Chaque terme inclut donc une constante, qui est spécifique à chaque type de liaison, angle et dièdre impropre ( $K_l$ ,  $K_\theta$  et  $K_\Phi$ , respectivement) et des valeurs d'équilibre et des valeurs d'équilibre ( $l_0$ ,  $\theta_0$ , et  $K_\Phi$ , respectivement). Le terme énergétique décrivant les dièdres est exprimé sous la forme d'une expansion en série de

cosinus, où  $\omega$  est le dièdre. Dans le calcul de l'énergie potentielle du système, ces termes sont calculés et sommés pour toutes les liaisons, les angles, les dièdres et les torsions impropres.

Le cinquième élément intègre la contribution des interactions non-covalentes à l'énergie potentielle. Le calcul concerne toutes les paires d'atomes non-covalents  $(i, j)$  provenant de la même molécule distante de trois liaisons, ou de molécules différentes. Les équations du potentiel sont généralement modélisées selon le potentiel de Coulomb pour les interactions électrostatiques et le potentiel de Lennard-Jones pour les interactions de van der Waals. Le potentiel de Coulomb est défini par les termes de charges  $q_i$  et  $q_j$ , la permittivité diélectrique dans le vide ( $\epsilon_0$ ), la constante diélectrique ( $\epsilon_r$ ) qui varie avec la distance qui sépare les deux particules chargées ( $r_{ij}$ ).  $r_{ij}$  est également utilisé dans le calcul du potentiel de Lennard-Jones, qui décrit les interactions de van der Waals, où on trouve aussi  $\epsilon_{ij}$ , qui correspond est la profondeur du puits de la fonction et il mesure la force d'attraction entre deux particules ( $i$  et  $j$ ), et  $\sigma_{ij}$  qui est la distance qui sépare les deux particules pour laquelle  $V_{vdw}$  est 0.

L'équation de l'énergie potentielle et ses paramètres définissent un champ de force qui est transposable à toute molécule de même type. Pour un même type de molécule biologique, plusieurs champs de force sont disponibles et ils diffèrent par la méthode d'obtention des paramètres ou par la forme des fonctions énergétiques. De plus, au sein d'une famille de champs de force (par exemple les champs de force Amber), des différences peuvent être observées, souvent liées aux paramètres des dièdres.

Les simulations MD ont pour objectif de reproduire ce qui a lieu dans un environnement biologique, donc le solvant, habituellement l'eau, qui doit être inclus dans la simulation. La façon plus réaliste de simuler les molécules d'eau consiste en l'utilisation d'un modèle d'eau explicite. C'est pour cela que cette approche a été choisie pour les simulations de cette étude.

En outre, toujours dans l'esprit de reproduire les conditions expérimentales, il est possible d'utiliser plusieurs ensembles thermodynamiques : l'ensemble microcanonique (NVE), où le nombre total de particules (N), le volume (V) et l'énergie totale (E) sont gardés constants, l'ensemble canonique (NVT), où le nombre total de particules (N), le volume (V) et la température (T) sont gardés constants et l'ensemble isothermique-isobarique (NPT), où le nombre total de particules (N), la pression (P) et la température (T) sont gardés constants.

## 4.2. Dynamiques moléculaires à échantillonnage intensif

L'espace conformationnel des biomolécules se caractérise par l'existence de plusieurs conformations métastables séparées par des barrières énergétiques. Explorer cet espace par l'approche de dynamique moléculaire classique requiert des simulations très longues et beaucoup de ressources. En fonction du système étudié, le temps de simulation en dynamique classique nécessaire pour atteindre la convergence se situe entre la microseconde et la milliseconde (Krüger et al., 2018). Sur ces échelles de temps, les temps de calculs peuvent prendre plusieurs jours et la quantité de données à stocker peut être très élevée (de l'ordre des To). De plus, il est fréquent, au cours des simulations de dynamiques moléculaires, que le système se maintienne dans un minimum d'énergie local, séparé d'une autre conformation stable par une haute barrière d'énergie. Cette approche n'est pas donc optimale lors de l'étude de macromolécules flexibles disposant de plusieurs états énergétiquement favorables, comme les oligonucléotides.

Afin de parvenir à échantillonner un maximum de conformations sur des échelles de temps relativement courtes, des méthodes de MD à échantillonnage intensif ont été développées.

Parmi ceux-ci, on identifie les approches de dynamique moléculaire accélérées (aMD) (Hamelberg et al., 2004), qui ont un coût computationnel équivalent à celui d'une MD classique, tout en explorant plus efficacement l'espace conformationnel, et ne nécessitent pas de connaissances *a priori* sur le système simulé. Pour améliorer l'échantillonnage conformationnel, pendant une simulation aMD, un biais énergétique ( $\Delta V(r)$ ) est intégré au calcul de l'énergie potentielle ( $V(r)$ ) afin de réduire la profondeur des minima locaux (Figure 3-12). La modification du potentiel se définit comme suit :

$$V(r) * = V(r) + \Delta V(r) \text{ (Équation 20)}$$

Dans la plupart de cas, le biais est appliqué sur l'énergie potentielle totale et sur l'énergie de torsion et il correspond à :

$$\Delta V(r) = \begin{cases} \frac{(E_p - V(r))^2}{\alpha_p + E_p - V(r)} + \frac{(E_d - V_d(r))^2}{\alpha_d + E_d - V_d(r)} & \text{lorsque } V(r) < E \\ 0 & \text{lorsque } V(r) \geq E \end{cases} \text{ (Équation 21)}$$

où  $V(r)$  est le potentiel standard,  $Vd(r)$  le potentiel de torsion standard, le biais d'énergie potentiel  $\alpha P$  et le seuil  $Ep$  ainsi que le biais d'énergie des dièdres  $\alpha D$  et le seuils  $Ed$ .

Comme décrit dans l'équation 21, le biais énergétique  $\Delta V(r)$  dépend de  $V(r)$ , de  $Vd(r)$  ainsi que du potentiel moyen  $Ep$  et des énergies des dièdres  $Ed$  obtenus lors d'une étape préliminaire de dynamique classique, et il est appliqué seulement si  $V(r)$  est inférieur à une valeur seuil. Les facteurs  $\alpha P$  et  $\alpha D$  (indiqués ensemble comme  $\alpha$ ) permettent d'ajuster l'intensité du biais, qui sera proportionnelle à la profondeur du minimum local. Plus la valeur  $\alpha$  fixée est faible, plus le profil énergétique sera modifié, s'approchant du seuil  $E$  (Figure 3-12).

Le biais d'énergie potentiel  $\alpha P$  et le biais d'énergie des dièdres  $\alpha D$  sont calculé selon les équations 22 et 23 :

$$\alpha P = \alpha_a * N_a \text{ (Équation 22)}$$

$$\alpha D = \frac{1}{5} * \alpha_n * N_{nuc} \text{ (Équation 23)}$$

Avec  $\alpha_a$  le biais appliqué en kcal·mol<sup>-1</sup>·atome<sup>-1</sup>,  $N_a$  le nombre d'atome du système,  $\alpha_n$  le biais appliqué en kcal·mol<sup>-1</sup>·nucleotide<sup>-1</sup> et  $N_{nuc}$  le nombre de nucléotides du système. Les valeurs seuil sont calculées pour  $Ep$  en fonction du facteur  $\alpha P$  et  $Ed$  en fonction du facteur  $\alpha D$  par les équations 24 et 25:

$$Ep = V(r) + \alpha P \text{ (Équation 24)}$$

$$Ed = Vd(r) + \alpha D \text{ (Équation 25)}$$

L'aMD permet ainsi d'explorer des transitions conformationnelles sur de courtes simulations entre 100 et 500 ns, non observées en dynamique moléculaire classique (Miao et al., 2013). De plus, en principe, le profil énergétique non-biaisé peut être récupéré.

Malgré cet avantage, la pondération appliquée au biais est fortement dépendante du profil d'énergie libre de la simulation. Celui-ci peut-être ponctué de cavités énergétiques surreprésentées dans le profil d'énergie corrigé par le biais, potentiellement source d'erreur ou de bruit dans la dynamique (Shen & Hamelberg, 2008).

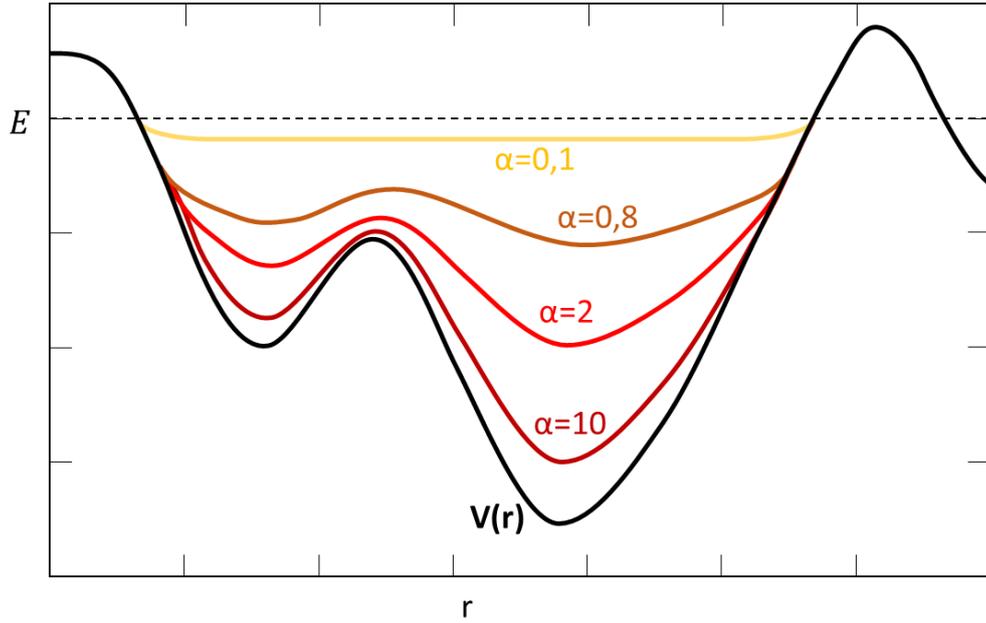


Figure 3-12. Exemple de profil énergétique d'un système simulé en dynamique moléculaire classique (noir) et l'effet de lissage induit par le facteur  $\alpha$  appliqué en aMD (soit  $\alpha = 10$  en rouge foncé,  $\alpha = 2$  en rouge,  $\alpha = 0,8$  en orange et  $\alpha = 0,1$  en jaune).

L'approche de dynamique moléculaire accélérée Gaussienne (GaMD) reprend le principe de biais appliqués sur les vallées d'énergie potentielle de l'aMD, mais son algorithme permet de pondérer la force du biais selon une gaussienne (J. Wang et al., 2021). Le biais  $\Delta V(r)$  est alors défini selon l'équation 26 :

$$\Delta V(r) = \begin{cases} \frac{1}{2} k (E - V(r))^2 & \text{lorsque } V(r) < E \\ 0 & \text{lorsque } V(r) \geq E \end{cases} \quad (\text{Équation 26})$$

où  $k$  est la constante de force harmonique et  $E$  le seuil d'énergie jusqu'à où le boost s'applique dans l'intervalle défini par l'équation 27 :

$$V_{max} \leq E \leq V_{min} - \frac{1}{k} \quad (\text{Équation 27})$$

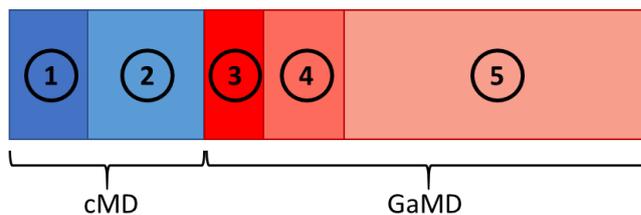
où  $k = k_0 \frac{1}{V_{max} - V_{min}}$ . Ici  $k_0$  est défini selon la limite supérieure ou inférieure  $E$  par l'équation 28 :

$$k_0 = \begin{cases} \min \left( 1, \frac{\sigma_0}{\sigma_v} \frac{V_{max} - V_{min}}{V_{max} - V_{moy}} \right) & \text{lorsque } E \text{ est la limite inférieure} \\ \left( 1 - \frac{\sigma_0}{\sigma_v} \right) \frac{V_{max} - V_{min}}{V_{moy} - V_{min}} & \text{lorsque } E \text{ est la limite supérieure} \end{cases} \quad (\text{Équation 28})$$

Un deuxième biais de la même forme est souvent appliqué sur l'énergie de torsion.

L'ensemble des équations appliquées permet la pondération selon une gaussienne lissant le profil d'énergie afin de réduire les barrières énergétique et l'impact des artefacts énergétiques sur le profil.

En GaMD, les paramètres de biais sont appliqués durant 4 étapes préparatoires (Figure 3-13). D'abord, 2 étapes de préparation en dynamique classique sont effectuées : une courte étape d'équilibration et une plus longue pour l'acquisition des statistiques concernant l'énergie potentielle  $V$  du système, soit l'énergie potentielle maximale ( $V_{max}$ ), minimale ( $V_{min}$ ), potentiel moyen ( $V_{moy}$ ), et son écart type ( $\sigma_V$ ). Deux étapes de préparation de GaMD sont effectuées pour ajuster les paramètres d'accélération et les statistiques d'énergie potentielle : une première pendant laquelle le biais est appliqué mais les paramètres ne sont pas mis à jour, et une deuxième pendant laquelle le biais est appliqué et les paramètres  $V_{max}$ ,  $V_{min}$ ,  $V_{moy}$ ,  $\sigma_V$  sont mis à jour. Le biais est, à chaque étape, dépendant de la valeur  $\sigma_0$ , qui fixe les limites de  $\sigma_V$  lors du recalcul de l'énergie potentielle. Si un double biais est utilisé (un sur l'énergie potentielle totale, l'autre sur l'énergie de torsion), 2 valeurs  $\sigma_0$  sont définies ( $\sigma_{0P}$  et  $\sigma_{0D}$ , respectivement).



- ① cMD préparatoire (équilibration du système)
- ② cMD avec collecte des données statistiques du potentiel
- ③ GaMD pre-équilibration: application du boost du potentiel sans mise à jour des paramètres
- ④ GaMD équilibration: application du boost du potentiel avec mise à jour des paramètres
- ⑤ GaMD production: application du boost du potentiel sans mise à jour des paramètres

Figure 3-13. Etapes de préparation de la dynamique accélérée Gaussienne.

Dans cette étude, les deux approches de dynamique moléculaire accélérée, standard (aMD) et Gaussienne (GaMD) avec un double biais, ont été testées.

### 4.3. Protocoles de simulation

Toutes les simulations ont été effectuées en utilisant le module *pmemd* de la suite Amber 20 (D.A. Case, T. Kur et al., 2020). Les simulations ont été effectuées en grande partie via l'Institut du développement et des ressources en informatique scientifique (IDRIS) et son supercalculateur Jean Zay HPE SGI 8600. D'autres simulations d'essais ou complémentaires ont été effectuées localement sur machine Dell Precision 7920 Tour équipée d'une carte graphique NVIDIA GeForce RTX 2080 Ti.

#### 4.3.1 Préparation du système

##### 4.3.1.1 Choix et préparation des structures des oligonucléotides

Au regard du nombre important de structures d'ADN contenues dans le jeu de données et des conditions à tester, une sélection de 5 structures a été faite, notamment 1NGO (Shiflett et al., 2003), 1EZN (Van Buuren et al., 2000), 3HXO (R.-H. Huang et al., 2009), 3THW (Gupta et al., 2012), 5HTO (Choi & Ban, 2016). Ces structures ont été sélectionnées afin d'étudier plusieurs niveaux de complexité structurale et vérifier les limites du protocole (Figure 3-14).

La structure de 1NGO correspond à une séquence transcrite et non traduite (UTR) en 3' du gène codant pour un antigène protecteur du *Bacillus anthracis*, jouant un rôle essentiel dans la pathogénèse des cellules de l'hôte. La structure secondaire de 1NGO, long de 27 nucléotides, est repliée en simple hélice-boucle, qui est correctement prédite par *mfold* ( $Apta_D = 0$ ). Cela représente donc un cas simple, qui permet d'analyser l'effet des différentes conditions de simulation testées.

1EZN présente une structure plus complexe : une jonction à 3 branches de 36 nucléotides. De plus, sa structure expérimentale a été résolue par RMN et présente 26 conformations relativement différentes les unes des autres. La prédiction de la structure secondaire de *mfold* montre une distance  $Apta_D$  de 0,143 par rapport à la référence. Néanmoins, les deux modèles obtenus par RNAde novo, un obtenu à partir de la structure secondaire expérimentale et l'autre obtenu à partir de la structure secondaire prédite par *mfold*, sont très éloignées de la structure de référence. Il est donc possible de vérifier si les dynamiques moléculaires sont capables d'échantillonner des structures plus proches à l'expérimentale que les simples méthodes de prédiction 3D.

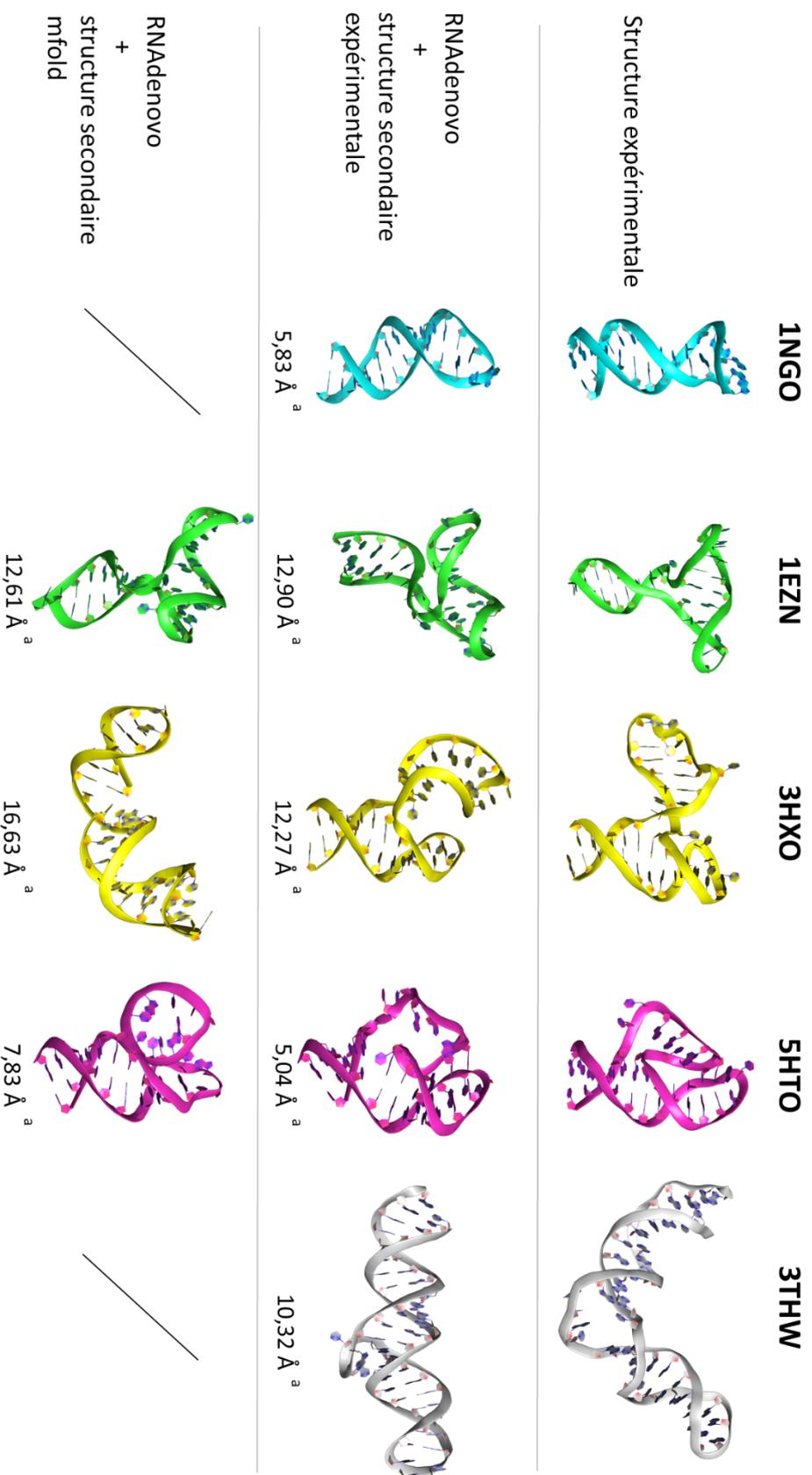


Figure 3-14. Représentation tridimensionnelle sur PYMOL des structures 1NGO, 1EZN, 3HXO, 5HTO et 3THW obtenues soit de la PDB, soit prédit par RNAdenovo en utilisant la structure secondaire expérimentale comme guide, ou en utilisant la structure secondaire prédit par mfold lorsque celle-ci est différente (soit pour 1EZN, 3HXO et 5HTO).<sup>a</sup> La RMSD face à la structure expérimentale a été mesurée pour tous les modèles RNAdenovo.

Les 3 autres structures sont issues de structures cristallographiques d'un complexe entre une protéine et l'oligonucléotide. 3HXO contient un oligonucléotide de 40 nucléotides replié en jonction à 3 branches comme 1EZN, mais la méthode d'acquisition de la structure ne permet de disposer que d'une unique conformation. Pour cet oligonucléotide, la prédiction de la structure secondaire par mfold a donné une distance  $Apta_D$  de 10,174 par rapport à l'expérimentale, indiquant une structure très éloignée de la référence. Par conséquent, cela a un fort impact sur la modélisation par RNAde novo (Figure 3-14). Ainsi les dynamiques moléculaires permettront d'observer l'effet de différences importantes dans les structures secondaires prédites et dans les modèles générés utilisés comme structure de départ.

3THW est la plus longue des 5 structures avec une chaîne de 53 nucléotides. Sa structure correspond à une longue hélice avec un bourgeon positionné en C36-A39 et a été correctement prédite par mfold. Le positionnement de ce bourgeon permet à l'oligonucléotide d'adopter une conformation courbée capable de s'insérer entre les deux sous unités MSH2 et MSH3 de MutatorS $\beta$  (Figure 3-15). Cette conformation n'a pas été obtenue par la modélisation avec RNAde novo, donc il est intéressant de mieux échantillonner les conformations accessibles par cet oligonucléotide.

Enfin, 5HTO est la seule structure intégrant un pseudonœud dans sa conformation expérimentale, qui implique les paires T11-A21 et T12-A20. A cause de son incapacité à prédire les pseudonœuds, mfold propose une structure secondaire pour 5HTO qui n'inclut pas le pseudonœud et impacte la prédiction de RNAde novo avec une rotation de la région entre T6 et T14 par rapport à l'expérimental (Figure 3-14). Cet oligonucléotide permet donc de vérifier l'efficacité du protocole basé sur la dynamique moléculaire sur la modélisation des pseudonœuds et des interactions à longue distance.

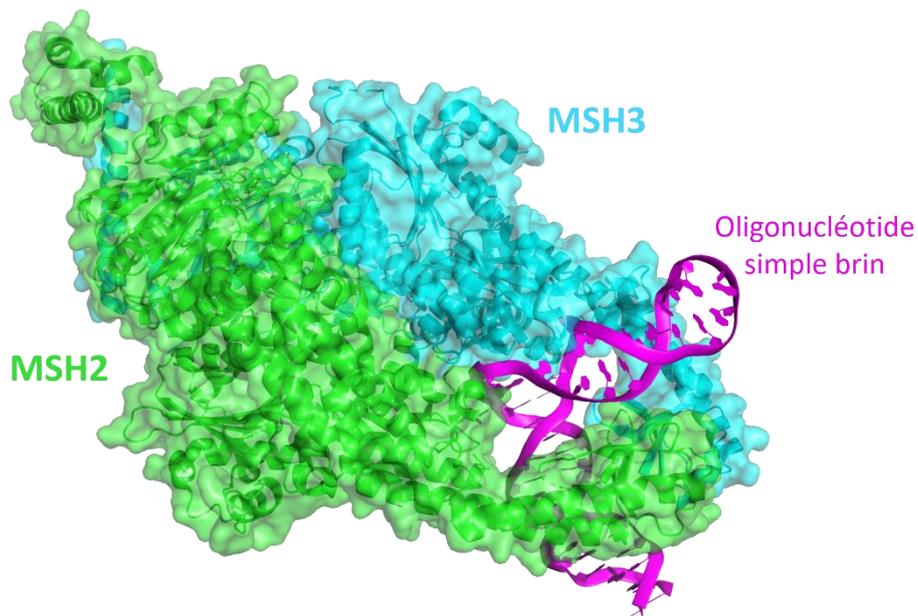


Figure 3-15. Représentation sur PyMOL de 3THW, intégrant la structure du complexe formé par MutS $\beta$  avec un fragment d'ADN simple brin présentant des lésions type insertion-délétion affectant la courbure de la boucle. En magenta est donc représenté l'oligonucléotide étudié dans ce manuscrit sous l'appellation 3THW.

Pour chacune des structures, il est attendu des dynamiques moléculaires de i) permettre l'exploration des conformations accessibles en utilisant comme structure de départ de la simulation l'oligonucléotide de la PDB, ii) retrouver une conformation proche de l'expérimental en partant des prédictions de RNAde novo, et iii) l'exploration d'autres conformations accessibles partagées entre les dynamiques moléculaires à partir de la structure expérimentale et de celle prédite.

Pour 1NGO et 3THW, dont la structure secondaire était correctement prédite par mfold, 2 structures de départ pour les dynamiques moléculaires ont été préparées : une prise de la structure expérimentale et l'autre correspondant au modèle obtenu par RNAde novo. Pour les oligonucléotides restants (1EZN, 3HXO et 5HTO), dont la structure secondaire prédite par mfold ne correspond pas exactement à celle expérimentale, 3 structures de départ pour les simulations de dynamique moléculaire ont été préparées : une première correspondant à la structure expérimentale, une seconde correspondant au modèle obtenu par RNAde novo à partir de la structure secondaire expérimentale et une troisième issue de la prédiction par RNAde novo à partir de la structure secondaire prédite par mfold.

Pour les structures expérimentales obtenues par RMN (1NGO et 1EZN), seule la première conformation, parmi celles disponibles, a été choisie. Pour les structures expérimentales obtenues par cristallographie aux rayons X en complexe avec une protéine (3HXO, 3THW, 5HTO), l'oligonucléotide est extrait de la structure et les atomes d'hydrogène ont été ajoutés avec le module *tleap* de la suite Amber20.

RNAAdenovo génère des modèles tridimensionnels d'oligonucléotides d'ARN, donc, pour tous les oligonucléotides de cette étude, les modifications décrites dans la Partie 3.2.3 ont été effectuées.

#### 4.3.1.2 Choix du champ de force

Une fois les coordonnées initiales des systèmes à modéliser obtenues, un choix du champ de force s'impose. Puisqu'il a été décidé de travailler avec la suite Amber20, ce choix s'est fait parmi les champs de force les plus récents et utilisés, disponibles sur Amber pour les ADN, notamment OL15 (Zgarbová et al., 2015) et bsc1 (Ivani et al., 2015).

OL15 correspond à l'optimisation des champs de force parm99 (J. Wang et al., 2000) et bsc0 (Pérez et al., 2007) avec un affinage des torsions  $\epsilon/\zeta$  du squelette phosphate, dont l'objectif est de favoriser la dynamique des ADN bicaténaires pour explorer la possible formation des ADN de forme Z, conformations difficiles à observer en dynamique moléculaire.

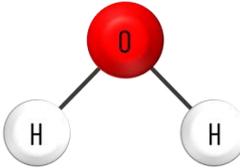
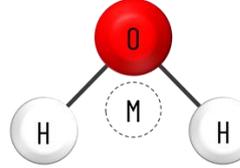
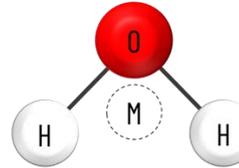
Bsc1 est également une optimisation du champ de force parm99 et bsc0 avec affinage des torsions  $\epsilon/\zeta$  du squelette phosphate, qui présente en plus une modification du plissement des cycles osidiques.

Les études comparatives disponibles dans la littérature ont montré une bonne capacité de ces deux champs de force à représenter la mobilité des ADN bicaténaires telle qu'observé sur des structures expérimentales RMN disposant de plusieurs conformations, le tout en solvant explicite (Galindo-Murillo et al., 2016). Le champ de force bsc1 semble cependant avoir montré de meilleures performances sur des simulations appliquées aux ADN simple brin, avec une meilleure tendance au repliement (formation de liaison hydrogènes, appariements) que OL15 (Oweida et al., 2021). Sur la base de ces résultats, le choix de bsc1 semble approprié.

#### 4.3.1.3 Choix du modèle d'eau

Comme indiqué dans la Partie 4.1, pour mieux reproduire les conditions expérimentales, les simulations ont été effectuées en utilisant un modèle d'eau explicite.

Tableau 8. Différences de paramètres entre les modèles d'eau en 3 points TIP3P et en 4 points TIP4P et TIP4P-Ew.

	<b>TIP3P</b>	<b>TIP4P</b>	<b>TIP4P-Ew</b>
			
<b>r(OH), Å</b>	0,9572	0,9572	0,9572
<b>r(OM), Å</b>	/	0,15	0,125
<b>HOH, degrés</b>	104,52°	104,52°	104,52°
<b>A x 10<sup>-3</sup>, kcal Å<sup>12</sup>/mol</b>	582	600	656,1
<b>B, kcal Å<sup>6</sup>/mol</b>	595	610	653,5
<b>q(O)</b>	-0,834	/	/
<b>q(H)</b>	+0,417	+0,52	+0,52422
<b>q(M)</b>	/	-1,04	-1,04844

Les modèles d'eau explicites recommandés en combinaison avec le champ de force bsc1 appartiennent à la classe TIP (Galindo-Murillo et al., 2016), notamment les modèles TIP3P et TIP4P. Le modèle TIP3P (Jorgensen et al., 1983) se base sur une molécule d'eau standard en 3 points pour les 3 atomes qui la composent (Tableau 8). Les distances séparant les atomes et l'angle HÔH se basent sur des données expérimentales obtenues en rayons X. Le modèle TIP4P ajoute un pseudo-atome de charge négative  $q = -1,04$  proche de l'oxygène, afin de reproduire le caractère dipolaire de l'eau, ainsi la charge partielle  $\delta^-$  portée par l'oxygène se retrouve déplacée sur ce point fictif. Les charges positives  $q$  appliquées aux hydrogènes sont également plus élevées pour le modèle TIP4P passant de  $q = +0,417$  à  $q = +0,52$ . En outre, le choix du modèle affecte le calcul des interactions électrostatiques, défini par l'équation 29 :

$$E_{ab} = \sum_i^a \sum_j^b \frac{q_i q_j e^2}{r_{ij}} + \frac{A}{r_{OO}^{12}} - \frac{B}{r_{OO}^6} \text{ (Équation 29)}$$

Avec  $q_i$  et  $q_j$  les charges partielles des atomes en interaction,  $e$  la constante de charge d'un électron et  $r_{OO}$  la distance séparant deux atomes d'oxygènes. Les termes  $A$  et  $B$  sont ensuite

modifiés selon le modèle sélectionné (Tableau 8) affectant les forces de dispersion et de répulsion.

Il existe plusieurs variantes du TIP4P appropriées pour différents types de simulations. TIP4P-Ew corrige les problèmes du modèle TIP4P standard dont les paramètres de densité, d'enthalpie ainsi que les propriétés cinétiques et thermodynamiques, montrent un certain écart avec les observations expérimentales. TIP4P-Ew ajuste certains paramètres d'angle et de liaison dans la molécule H<sub>2</sub>O (Tableau 8) et utilise la sommation d'Ewald pour le calcul des énergies d'interaction, faisant de TIP4P-Ew un modèle représentatif de l'eau plus adapté aux simulations de biomolécules (Horn et al., 2004). De plus, la plupart des paramètres influençant les interactions électrostatiques sont modifiés par rapport au modèle standard, avec notamment les paramètres de charges  $q$  ajustés et l'impact des forces de dispersion et répulsion revues à la hausse.

Les modèles TIP3P et TIP4P-Ew ont été testés durant ce travail dans le but d'identifier les éventuelles différences de comportements. TIP3P fait office de modèle d'eau standard, relativement rapide à calculer. En parallèle, le modèle TIP4P-Ew permet de modéliser avec une meilleure fiabilité le comportement des molécules d'eau en solution. En contrepartie le temps de calcul nécessaire à durée de simulation égale est allongé, car pour chaque molécule d'eau il y aura un atome en plus.

#### 4.3.1.4 Présence des ions

Les acides nucléiques sont des molécules chargées négativement en raison des groupements phosphate le long du squelette sucre-phosphate. Afin de travailler dans un système neutralisé, il est essentiel d'ajouter l'équivalent en contre-ions de la charge négative totale de la molécule. Ainsi le système est neutralisé au préalable avec des ions Na<sup>+</sup> afin d'équilibrer la charge négative des groupements phosphate des acides nucléiques.

En outre, en conditions expérimentales, une certaine concentration de sels de nature différente est souvent rajoutée à la solution. En dynamique moléculaire, l'ajout d'une concentration saline a démontré un effet sur le comportement en solution des acides nucléiques. Cet effet est dépendant de la concentration et tend à réduire la mobilité

conformationnelle du squelette phosphate des acides nucléiques mono- et bicaténaire mais également à impacter les appariements (Bell et al., 2020; Draper et al., 2005; MacKerell, 1997).

Par conséquent, pour chaque oligonucléotide simulé, deux conditions ont été testées : dans la première le système a été neutralisé par des ions  $\text{Na}^+$ . Dans la deuxième, la neutralisation de la charge a été suivie par l'ajout d'une concentration de 0,1 M de NaCl. La comparaison des différentes conditions permet de faire le lien entre les connaissances actuelles de la dynamique des oligonucléotides et de constater les éventuelles différences de comportement. Le choix des ions à inclure se base sur les conditions de solvant *in vitro* appliquées pour l'incubation d'oligonucléotides avec une protéine cible (Cho et al., 2006; Gong, 2023; Loussouarn, 2014). L'équivalent en ions des autres tampons détaillés dans ces travaux est négligeable dans un système de dynamique moléculaire. Les ions sélectionnés pour neutraliser le système sont des ions  $\text{Na}^+$  et  $\text{Cl}^-$  dans les concentrations détaillées dans les travaux mentionnés, soit l'équivalent de 0,1 M. Le nombre de molécules de NaCl à ajouter est déterminé selon le volume de la boîte et la concentration souhaitée, soit 0 pour les simulations effectuées en solvant neutralisé seul, ou l'équivalent de 0,1 M de NaCl soit le nombre d'atome calculé selon la formule :

$$N_{\text{atomes}} = V_{\text{Tot}} \times N_A \times C \text{ (Équation 30)}$$

Avec  $V_{\text{Tot}}$  le volume totale de la boîte en L qui varie en fonction du système,  $N_A$  le nombre d'Avogadro et  $C$  la concentration souhaitée en M, dans ce cas 0,1 M. Pour 0,1M de molécules de NaCl, on ajoute donc  $N_{\text{atomes}}$  d'ions  $\text{Na}^+$  et  $N_{\text{atomes}}$  d'ions  $\text{Cl}^-$ .

#### 4.3.2 Protocole de dynamique moléculaire appliqué

Une fois le système préparé, il peut être soumis à la dynamique moléculaire. Celle-ci prévoit plusieurs étapes détaillées dans les sections suivantes.

##### 4.3.2.1 Minimisation, chauffe et équilibration

Tout d'abord une ou plusieurs étapes de minimisation sont effectuées afin de rapprocher le système au minimum énergétique le plus proche en ajustant la position des atomes en fonction de l'énergie de liaison, si l'atome implique une liaison covalente, ou de répulsion si l'atome implique une certaine proximité avec un autre.

Dans le protocole utilisé, en première étape, les hydrogènes sont minimisés sur 1000 cycles de descente de gradient suivi jusqu'à 4000 cycles de gradient conjugué. Ensuite, les molécules d'eau et les ions sont minimisés à leur tour sur 2000 cycles de descente de gradient suivi jusqu'à 3000 cycles de gradient conjugué.

Ensuite, des étapes d'équilibration du potentiel total, de la pression, du volume et densité du système, sont menées à la température de simulation souhaitée. Dans le protocole utilisé, deux étapes de 100 ps d'équilibration ont été effectuées. La première équilibration a été effectuée à volume constant (NVT) et la seconde à pression constante (NPT) pour ajuster l'état du système en maintenant la pression à 1 atm. L'ensemble des atomes du système équilibré est minimisé sur 2500 cycles de descente de gradient puis 2500 cycles de gradient conjugué.

La température du système est ensuite incrémentée progressivement de 0 à 300 K avec 6 étapes de 5 ps avec une augmentation de 50 K par étape (à pression constante et volume constant). Enfin, le volume et la pression du système ont été rééquilibrés à 300 K sur 200 ps en NVT et 1 ns en NPT, respectivement.

#### 4.3.2.2 Phase de production

Après les étapes de minimisation et équilibration, une phase de production est effectuée, pendant laquelle on laisse le système évoluer en fonction du temps. Le protocole utilisé prévoit 20 ns de production dans un système à pression et température constante, soit une pression fixée à 1 atm et une température de 300 K, en conditions périodiques. L'approche thermostatique type Langevin est utilisée pour le maintien de la température, avec une fréquence de collision de  $2,0 \text{ ps}^{-1}$ , et la sommation d'Ewald à maillage de particules pour les interactions électrostatiques longues distance, avec un seuil limite des interactions électrostatique défini à 8 Å. Enfin, l'algorithme SHAKE pour le calcul des contraintes de liaisons a été utilisé dans toutes les simulations.

#### 4.3.2.3 Dynamiques moléculaires accélérées et dynamiques moléculaires accélérées gaussiennes

Comme déjà anticipé, les oligonucléotides, en tant que molécules très mobiles, nécessitent de longues simulations de dynamiques moléculaires pour parvenir à explorer l'espace conformationnel accessible pour ces molécules. La dynamique moléculaire classique peut

ainsi requérir des simulations de l'ordre de la micro/milliseconde pour explorer un grand nombre de conformations. Plusieurs essais de simulations de dynamique moléculaire classique ont été effectués sur structure expérimentale afin d'observer d'éventuels changements sur 500 ns. Aucune simulation n'a atteint la convergence, et l'exploration des conformations était limitée voire absente comme observé sur les structures 3THW et 3HXO en Figure 3-16, avec une stabilité de 3HXO sur toute la durée, et pour 3THW quelques pics de RMSD suggérant des changements de conformations qui ne se maintiennent pas. Ces observations ont donc motivé l'utilisation des méthodes d'échantillonnages intensifs, qui suivent la phase de production en dynamique classique.

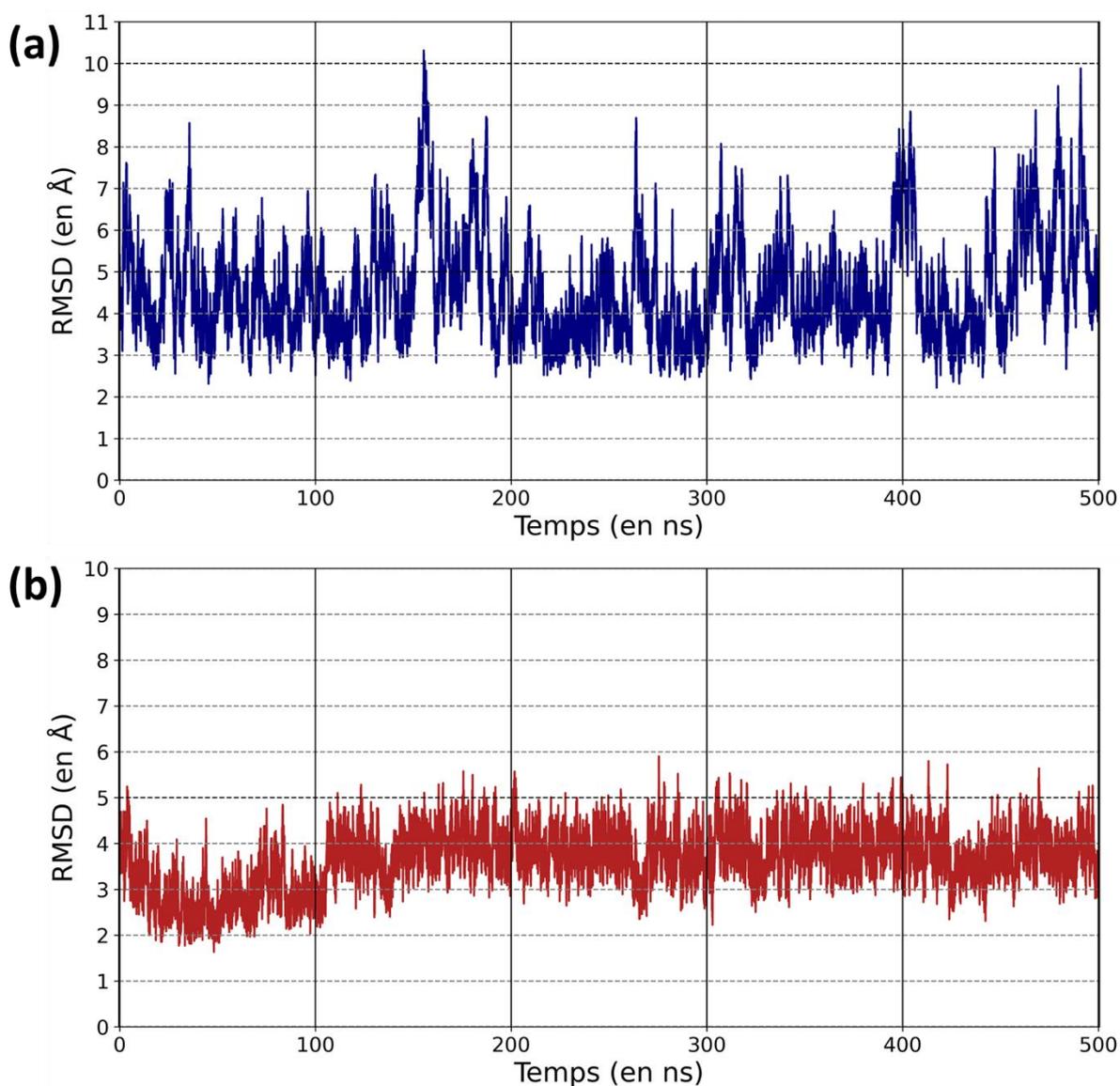


Figure 3-16. Courbes de RMSD des simulations en dynamique moléculaire classique de 500 ns effectuées sur les structures expérimentales de (a) 3THW et (b) 3HXO. Les deux structures ont été simulées en systèmes neutralisés et avec le modèle d'eau TIP3P.

Dans cette étude, les deux approches de dynamique moléculaire accélérée, standard (aMD) et Gaussienne (GaMD), ont été testées. Le passage en GaMD a été plus tardif et lié à la volonté de pouvoir ensuite récupérer un profil énergétique sans artefacts. Dans les deux cas le biais d'énergie est appliqué sur les dièdres et sur l'énergie potentielle totale.

Le seuil d'énergie  $E$  pour les simulations d'aMD et de GaMD a été sélectionné sur la borne inférieure, donc  $E = V_{max}$ .

Pour les simulations aMD, comme indiqué par le manuel d'Amber 20, les valeurs  $\alpha$  pour les facteurs de torsion et potentiel ont été fixées à 0,2 et 0,16 kcal mol<sup>-1</sup> atome<sup>-1</sup>, respectivement. Les autres paramètres ont été gardés comme dans la phase de production de la dynamique classique. Deux courses indépendantes, de 1  $\mu$ s chacune, ont été effectuées pour chaque système étudié, en sauvegardant la trajectoire toutes les 10 ps, pour avoir un nombre suffisant d'instantanées à analyser tout en limitant la quantité de données à stocker.

Pour ce qui concerne les simulations GaMD, le double biais, sur l'énergie potentielle totale et sur l'énergie des torsions, a été appliqué, en fixant les valeurs de  $\sigma_{OP}$  et  $\sigma_{OD}$  à 8 kcal/mol pour un recalcul de  $V$  sur des conformations d'énergie potentielle plus importante. La procédure décrite dans la Partie 3.4.2 a ensuite été effectuée. Notamment, après la dynamique classique de 20 ns, 2 courtes étapes de préparation en dynamique classique ont été effectuées : la première de 500 ps et la deuxième de 2 ns pour l'acquisition des statistiques concernant l'énergie potentielle  $V$  du système : l'énergie potentielle maximale ( $V_{max}$ ), minimale ( $V_{min}$ ), la moyenne ( $V_{moy}$ ), et l'écart type ( $\sigma_v$ ). Ensuite, deux étapes de préparation de dynamique accélérée Gaussienne ont été effectuées pour ajuster les paramètres d'accélération et les statistiques d'énergie potentielle, d'abord avec 500 ps de dynamique incluant les biais sans les mettre à jour, puis 10 ns où les données d'accélération sont calculées et les statistiques ( $V_{max}$ ,  $V_{min}$ ,  $V_{moy}$ ,  $\sigma_v$ ) réajustées. Les biais mis à jour dans cette dernière phase ont été appliqués dans la phase de production de 1  $\mu$ s. Dans ce cas aussi, les autres paramètres de simulation ont été gardés comme dans la phase de production de la MD classique et 2 courses indépendantes ont été effectuées pour chaque système.

Pour résumer, un total de 140 simulations a été effectué pour un total de 140  $\mu$ s, dont 80 simulations aMD et 60 GaMD. Deux modèles d'eau explicites ont été testés, ainsi que deux

conditions environnementales (neutralisation de la charge et neutralisation suivie par un ajout de 0,1 M de NaCl). De plus, pour chaque oligonucléotide, plusieurs coordonnées de départ ont été utilisées. Le Tableau 9 résume toutes les simulations effectuées.

Tableau 9. Résumé de l'ensemble des simulations effectuées sur les différents oligonucléotides soit 1EZN, 1NGO, 3HXO, 3THW ou 5HTO

		1EZN				1NGO				3HXO				3THW				5HTO			
		aMD		GaMD		aMD		GaMD		aMD		GaMD		aMD		GaMD		aMD		GaMD	
Concentration NaCl (M)		0	0,1	0	0,1	0	0,1	0	0,1	0	0,1	0	0,1	0	0,1	0	0,1	0	0,1	0	0,1
TIP3P	Expérimentale	2x1 µs	2x1 µs		2x1 µs																
	RNAde novo structure secondaire expérimentale	2x1 µs	2x1 µs		2x1 µs																
	RNAde novo structure secondaire mfold				2x1 µs				2x1 µs					2x1 µs				2x1 µs			
TIP4P-Ew	Expérimentale	2x1 µs	2x1 µs		2x1 µs																
	RNAde novo structure secondaire expérimentale	2x1 µs	2x1 µs		2x1 µs																
	RNAde novo structure secondaire mfold				2x1 µs																

Les modèles RNAde novo obtenus à partir de la structure secondaire de mfold ont été simulés exclusivement en GaMD pour une question de temps et avec l'objectif pour la suite de ce projet de pouvoir récupérer un profil énergétique plus précis.

#### 4.3.2.4 Obtention des conformations majoritaire par clustering

Une fois les trajectoires générées, il est nécessaire de les analyser pour étudier les conformations majoritaires échantillonnées et appréhender les similitudes et différences avec les structures expérimentales. Cela a été effectué en utilisant une procédure de clustering, qui permet de regrouper des conformations similaires en fonction d'un critère choisi au préalable.

L'approche choisie utilise un module Python, TTclust (Tubiana et al., 2018), optimisé pour l'analyse de longues dynamiques avec différentes approches de clustering hiérarchiques disponibles impliquant différents modes de répartition des conformations observées au cours d'une simulation : le clustering par la méthode de Ward, la méthode *simple-linkage*, la méthode *complete-linkage*, la méthode *average linkage*, la méthode médiane et la méthode des centroïdes. La structure représentative du cluster est celle qui présente le moins d'écart

avec les autres membres de son cluster. Pour la génération des clusters avec TTclust, le choix s'est porté sur l'alignement et la mesure de RMSD entre atomes du squelette phosphate soit pour chaque nucléotide composant le polymère, les atomes P, O5', C5', C4', C3' et O3'. La classification par la méthode de Ward a été utilisée pour le clustering, et le nombre de cluster à générer est déterminé automatiquement par TTclust. D'abord, l'approche des k-means regroupe les conformations selon leur proximité en incrémentant progressivement le nombre de clusters à créer puis identifier la variabilité moyenne dans les clusters par la somme des carrés. Ensuite, la méthode Elbow cherche à retrouver le nombre optimal en cherchant le point où la variabilité est la plus basse avant la convergence (Kaufman & Rousseeuw, 1990).

Le clustering avec TTclust est appliqué sur chaque réplique de dynamique moléculaire (aMD ou GaMD). Les clusters ainsi créés sont distincts entre les répliques pour un même oligonucléotide. Afin de faciliter la compréhension, les clusters seront nommés en fonction de la simulation à laquelle ils appartiennent selon la notation cluster X.Y avec X le numéro de réplique de simulation (aMD ou GaMD précisé) et Y le numéro du cluster.

#### 4.4. Etude conformationnelle des oligonucléotides par dynamique moléculaire

Pour tous les oligonucléotides sélectionnés, plusieurs simulations de dynamiques moléculaires ont donc été effectuées. Les dynamiques moléculaires accélérées standard (aMD) ont dans un premier temps été testé sur ces systèmes. Puis, au regard des résultats obtenus sur plusieurs systèmes, les dynamiques accélérées Gaussienne (GaMD) ont été testé afin d'évaluer les éventuelles variations de comportement dynamique par rapport à l'aMD qui présentent plusieurs défauts ayant un impact potentiel sur l'exploration de l'espace conformationnel.

Le changement de la concentration en NaCl dans le système de 0 à 0,1 M nous permet de d'observer l'effet des ions sur les dynamiques moléculaires. Ces deux conditions ont été testées en aMD, et en GaMD seuls les systèmes concentrés à 0,1 M de NaCl ont été testé pour se focaliser sur des systèmes potentiellement aux conditions similaires à celles appliquées expérimentalement. Enfin, les deux modèles d'eau TIP3P et TIP4P-Ew ont été utilisés dans le but d'évaluer l'effet du modèle d'eau sur la dynamique des systèmes. Pour les simulations effectuées à partir du modèle obtenu par RNAde novo depuis la structure secondaire prédite par mfold, seulement la méthode GaMD a été testée.

#### 4.4.1 1NGO

La séquence de 1NGO provient de la région 3'UTR du transcrit codant pour un antigène protecteur du *Bacillus anthracis*, essentielle donc pour la stabilité du transcrit et caractérisé par une structure tertiaire peu complexe constituée d'une unique hélice-boucle comprenant 27 nucléotides, dont la structure secondaire est correctement prédite par mfold.

Les simulations effectuées ont donc comme point de départ soit la structure expérimentale soit le modèle obtenu par RNAde novo à partir de la structure secondaire expérimentale. Ce dernier a une RMSD par rapport à la structure expérimentale de 5,83 Å.

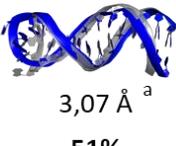
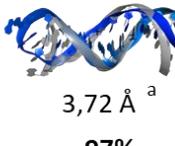
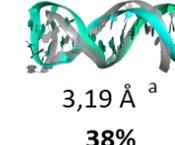
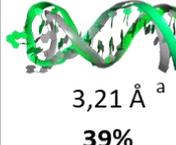
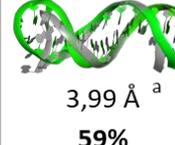
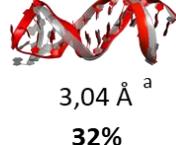
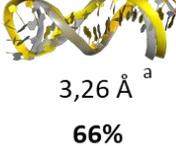
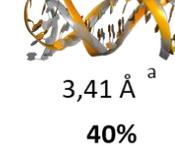
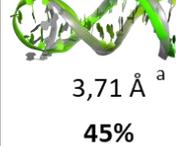
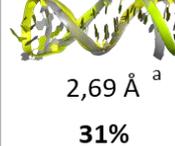
Cluster		aMD		GaMD	
		n°1	n°2	n°1	n°2
TIP3P	0 M NaCl	 3,07 Å <sup>a</sup> 51%	 3,72 Å <sup>a</sup> 97%		
	0,1 M NaCl	 3,82 Å <sup>a</sup> 65%	 3,19 Å <sup>a</sup> 38%	 3,21 Å <sup>a</sup> 39%	 3,99 Å <sup>a</sup> 59%
TIP4P-Ew	0 M NaCl	 3,04 Å <sup>a</sup> 32%	 2,89 Å <sup>a</sup> 45%		
	0,1 M NaCl	 3,26 Å <sup>a</sup> 66%	 3,41 Å <sup>a</sup> 40%	 3,71 Å <sup>a</sup> 45%	 2,69 Å <sup>a</sup> 31%

Figure 3-17. Alignement des clusters majoritaires des différentes simulations (cluster X.1 avec X le numéro de réplique de simulation) effectuées en partant de la structure résolue expérimentalement de 1NGO. Les conditions varient selon le modèle d'eau utilisé (TIP3P ou TIP4P-Ew), la concentration en NaCl dans le système (0 ou 0,1M) et le type de simulation (aMD ou GaMD). La RMSD obtenue face à la structure expérimentale est précisée pour chaque alignement.

L'ensemble des simulations de dynamique effectuées sur les structures de 1NGO ont montré la stabilité de la structure dans le temps indépendamment des conditions appliquées. En aMD dans un premier temps, cette stabilité a été vérifiée au travers des 2 modèles d'eau appliqués soit TIP3P et TIP4P-Ew, ainsi que les 2 niveaux de concentration en NaCl introduits dans les

systèmes (0 ou 0,1 M). Ainsi les clusters majoritaires (i.e. regroupant au minimum 10 % de la trajectoire) identifiés avec TTclust montrent tous une structure représentative avec une distance de la structure résolue expérimentalement  $\leq 5 \text{ \AA}$  (Annexes 9 à 16).

Par exemple la deuxième simulation aMD effectuée sur 1NGO en système neutralisé et modèle d'eau TIP3P dispose d'un cluster représentatif de  $\sim 96 \%$  de la dynamique avec une RMSD de  $3,7 \text{ \AA}$  (Figure 3-17). Cette stabilité est également observée dans les simulations réalisées à partir du modèle obtenu par RNAdenovo, déjà très similaire à la structure expérimentale, et aussi dans les simulations GaMD, où l'ensemble des dynamiques exécutées sont stables sur  $1 \mu\text{s}$  comme montré en Figure 3-18 pour les GaMD effectués avec le modèle d'eau TIP3P.

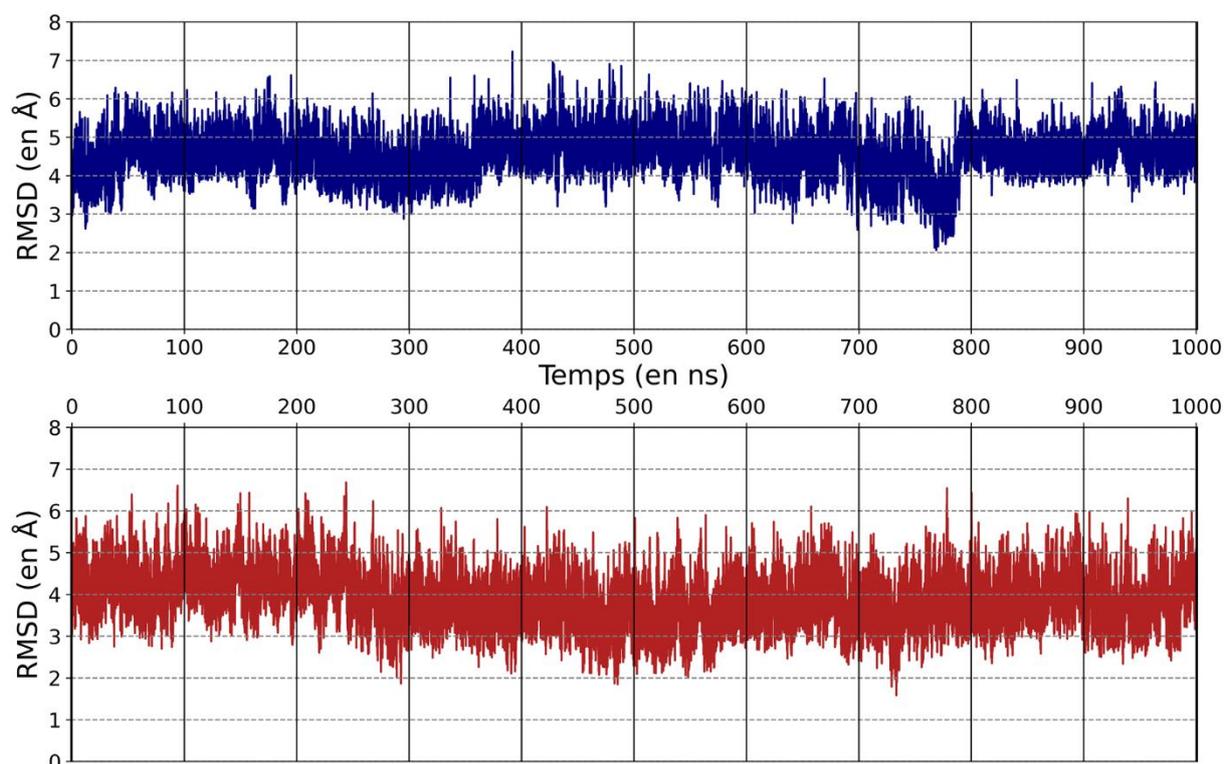


Figure 3-18. RMSD des simulations de GaMD sur 1NGO modélisé avec RNAdenovo basé sur la structure secondaire expérimentale en TIP3P. La RMSD face à la structure résolue expérimentalement est mesurée le long des dynamiques de GaMD n°1 (bleu) et GaMD n°2 (rouge).

Néanmoins, parmi les clusters minoritaires, des structures représentatives des clusters avec une RMSD par rapport à l'expérimental  $> 5 \text{ \AA}$  peuvent être observées. Cela est le cas par exemple des simulations sur la structure expérimentale de 1NGO i) en système neutralisé sans NaCl avec le modèle TIP3P en aMD, ii) en système concentré à 0,1 M de NaCl avec le modèle TIP4P-Ew en aMD, iii) en système concentré à 0,1 M de NaCl avec le modèle TIP3P en GaMD.

Cela a également été observé pour les simulations faites à partir de la structure modélisée avec RNAde novo, dans un système concentré à 0,1M de NaCl avec le modèle TIP3P en aMD. Ces structures maintiennent la structure secondaire expérimentale, mais elles montrent une flexibilité de l'hélice et de la boucle.

La stabilité de ces simulations dans les conditions testées n'est pas surprenante, car l'étude structurale qui a mené à la détermination de la structure expérimentale de 1NGO souligne la forte stabilité de cette boucle-hélice, qui est fondamentale pour sa fonction. (Shiflett et al., 2003)

#### 4.4.2 1EZN

1EZN contient une structure de type jonction à 3 branches de 36 nucléotides, déterminée par RMN pour étudier extensivement ce type d'organisation tertiaire dans les ADN à simple brin (Van Buuren et al., 2000). La structure de départ contient 26 conformères, on s'attend donc à pouvoir identifier plusieurs conformations. Dans ce cas, 3 structures de départ ont été utilisées pour les dynamiques moléculaires : la première conformation prise de la structure expérimentale, le modèle de RNAde novo obtenu à partir de la structure secondaire expérimentale et le modèle de RNAde novo obtenu à partir de la structure secondaire prédite par mfold, qui montrait une distance  $Apta_D$  de 0,143 par rapport à l'expérimental. Les valeurs de RMSD indiquées dans la discussion suivante ont été calculées par rapport au premier conformère de l'ensemble obtenu expérimentalement par RMN, sauf si indiqué autrement.

##### 4.4.2.1 Simulations effectuées à partir de la structure expérimentale

Les aMD effectuées en système neutralisé en partant de la structure résolue expérimentalement et indépendamment du modèle d'eau utilisé montrent, comme attendu, une forte mobilité de l'oligonucléotide avec une majorité de clusters ayant une structure représentative avec une RMSD élevée ( $> 5 \text{ \AA}$ ) face à la référence expérimentale (Figure 3-19). On identifie néanmoins un cluster dont la structure représentative a une faible RMSD ( $\leq 5 \text{ \AA}$ ) représentatif de  $\sim 27 \%$  de l'aMD n°2 en TIP3P (Annexe 9) et  $\sim 30 \%$  de l'aMD n°1 en TIP4P-Ew (Annexe 10). En parallèle, les aMD effectuées en système concentré à 0,1 M de NaCl n'ont pas permis d'identifier de structure proche de l'expérimental, aussi bien en TIP3P qu'en TIP4P-Ew, bien que quelques clusters semblent s'approcher du seuil de  $5 \text{ \AA}$  (cluster 2.2 TIP3P, cluster 2.1 TIP4P-Ew, Annexes 13 et 15). Cela peut être dû au fait que l'obtention de la structure

expérimentale a été faite dans une solution contenant 50 mM de NaCl. Par conséquent, la concentration en NaCl utilisée pour les simulations peut avoir eu un effet sur l'échantillonnage de la conformation correspondante à l'expérimentale. Cela justifierait un choix plus précis des concentrations salines en simulation.

### Structure expérimentale 1EZM

#### aMD n°1

#### aMD n°2

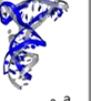
Cluster		n°1	n°2	n°3	n°4	n°5	n°1	n°2	n°3	n°4
NaCl 0M	TIP3P	 10,80 Å <sup>a</sup> 45%	 5,40 Å <sup>a</sup> 28%	 11,15 Å <sup>a</sup> 17%	 10,97 Å <sup>a</sup> 10%		 12,46 Å <sup>a</sup> 33%	 10,56 Å <sup>a</sup> 30%	 4,98 Å <sup>a</sup> 27%	 10,94 Å <sup>a</sup> 10%
	TIP4P-EW	 6,67 Å <sup>a</sup> 34%	 4,87 Å <sup>a</sup> 31%	 8,66 Å <sup>a</sup> 13%	 10,66 Å <sup>a</sup> 12%	 6,52 Å <sup>a</sup> 10%	 5,29 Å <sup>a</sup> 35%	 5,62 Å <sup>a</sup> 30%	 10,80 Å <sup>a</sup> 24%	 6,60 Å <sup>a</sup> 11%
NaCl 0,1 M	TIP3P	 12,54 Å <sup>a</sup> 45%	 10,98 Å <sup>a</sup> 38%	 9,85 Å <sup>a</sup> 17%			 7,36 Å <sup>a</sup> 46%	 5,56 Å <sup>a</sup> 39%	 6,91 Å <sup>a</sup> 16%	
	TIP4P-EW	 5,41 Å <sup>a</sup> 52%	 5,93 Å <sup>a</sup> 34%	 10,94 Å <sup>a</sup> 14%			 13,05 Å <sup>a</sup> 57%	 12,89 Å <sup>a</sup> 36%	 10,05 Å <sup>a</sup> 7%	

Figure 3-19. Alignement des clusters des différentes simulations effectuées en partant de la structure résolue expérimentalement de 1EZM en système neutralisé avec et sans concentration de 0,1 M de NaCl testé avec les modèles d'eau TIP3P ou TIP4P-Ew en aMD. <sup>a</sup> La RMSD obtenue face à la structure expérimentale est précisée pour chaque alignement.

En GaMD, le comportement observé est similaire. Les simulations effectuées avec le modèle TIP3P et 0,1 M NaCl ne montrent aucun cluster avec une structure représentative ayant une  $RMSD \leq 5 \text{ \AA}$  par rapport à la première conformation expérimentale (Figure 3-20, Annexe 15). Quelques conformations tendent à s'approcher du seuil du 5 Å, notamment le cluster 1.2 avec une RMSD de 5,36 Å. Néanmoins, en considérant l'ensemble des conformations obtenues en RMN (Figure 3-21), on observe une certaine proximité du cluster 1.2 avec 22 sur les 26 conformations obtenues expérimentalement. Au regard de cette observation, une certaine flexibilité du seuil de 5 Å peut être autorisée. Il est intéressant d'observer que, dans les deux

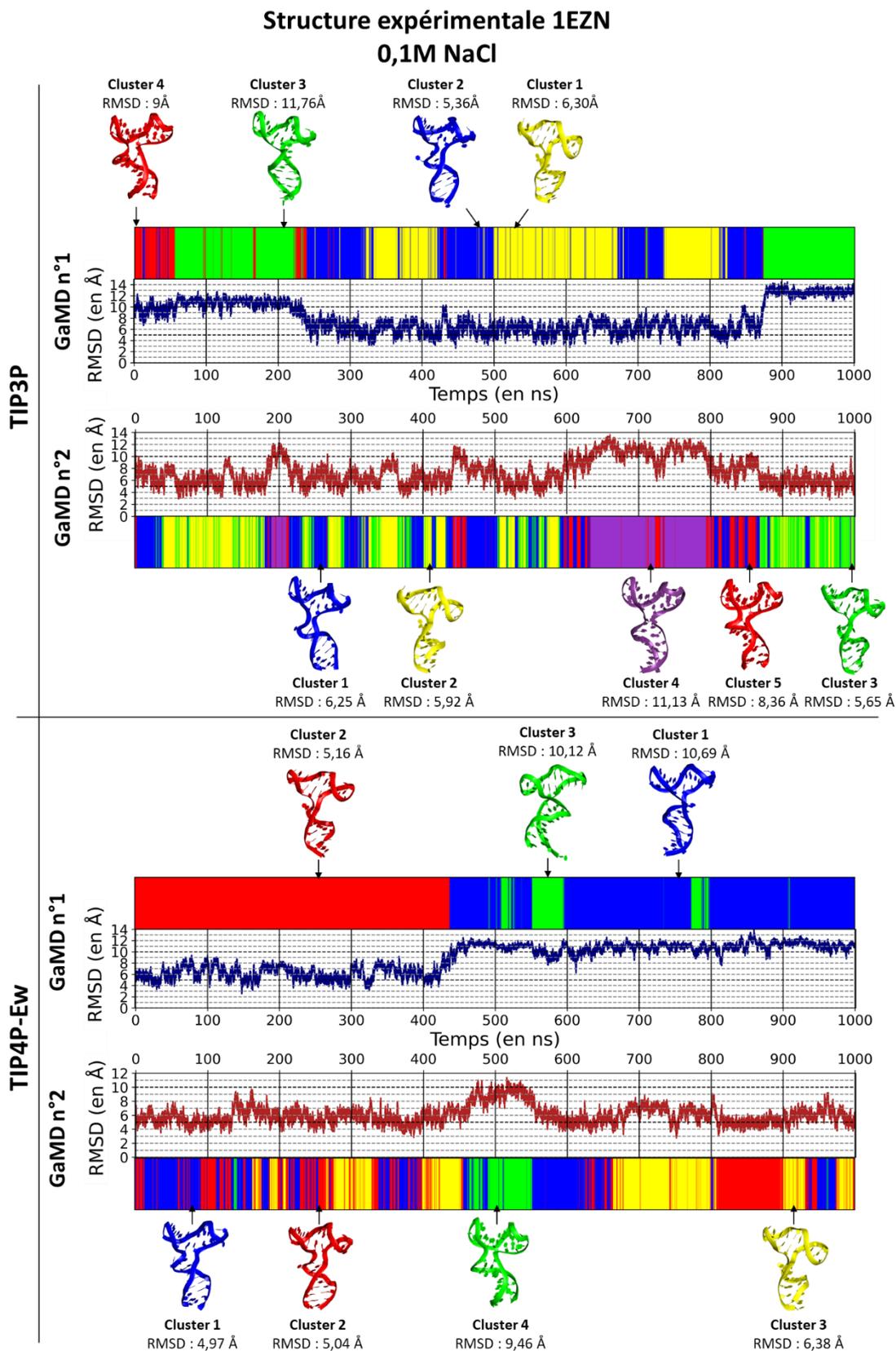


Figure 3-20. Courbe de RMSD et distribution des clusters et des conformations associées au cours des GaMD simulés dans un système concentré à 0,1 M en NaCl avec différents modèles d'eau (TIP3P ou TIP4P-Ew) appliqués sur la structure expérimentale de 1EZN. Chaque couleur représente un cluster différent. Le numéro du cluster est assigné selon son pourcentage d'occurrence dans la simulation.

simulations faites sous ces conditions, chaque conformation échantillonnée est stable sur une longue durée avant d'observer un changement conformationnel (Figure 3-20).

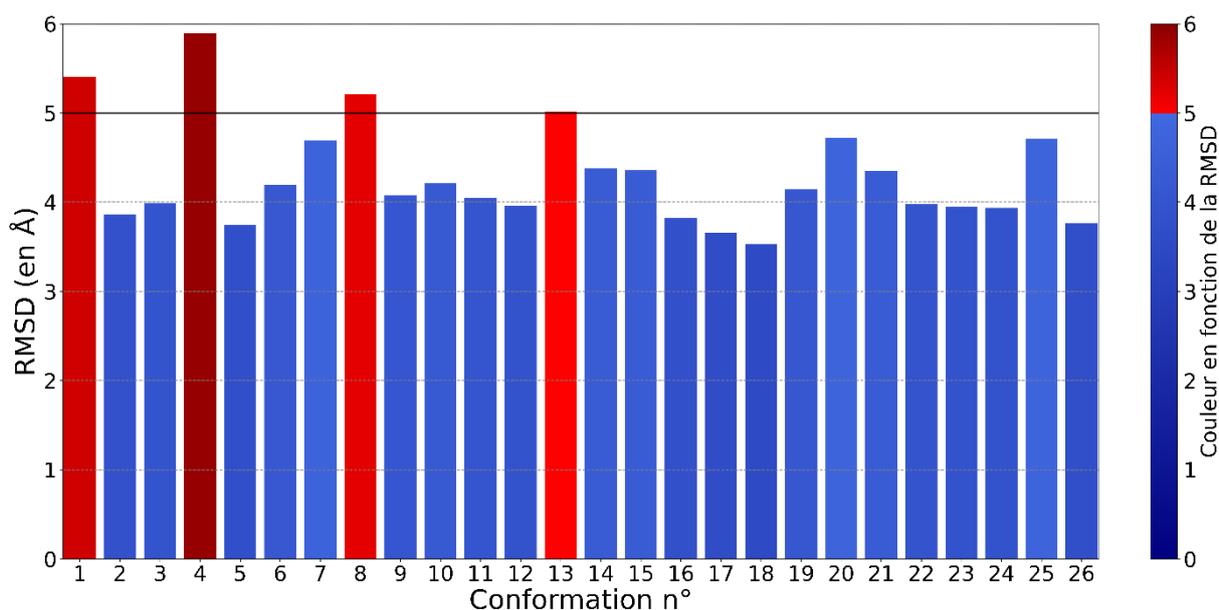


Figure 3-21. RMSD calculée (en Å) entre les différentes conformations de 1EZN expérimentale et le cluster 1.2 de la GaMD n°1 sur structure expérimentale dans un système incluant une concentration de 0,1 M en NaCl et modèle d'eau TIP3P. Le gradient de couleur est décroissant avec les valeurs de RMSD : rouge pour les valeurs les plus élevées, bleu pour les valeurs les plus faibles.

Pour les simulations GaMD effectuées avec le modèle TIP4P-Ew et 0,1 M NaCl, seul un cluster présente une structure représentative avec une  $RMSD \leq 5 \text{ \AA}$  par rapport à l'expérimental, et deux autres clusters franchissent ce seuil avec un dépassement inférieur à 5 % (Annexe 18). Le comportement de ces répliques est similaire au comportement observé en TIP3P. En GaMD n°1, le cluster 1.2, proche de la structure expérimentale est maintenu, sur environ 430 ns avant un changement de conformation vers le cluster 1.1 différent de la structure expérimentale ( $RMSD = 10,69 \text{ \AA}$ ). En GaMD n°2, les clusters 2.1 et 2.2 sont représentatifs de 62 % de la simulation et ont une structure représentative proche de la structure expérimentale ( $\sim 5 \text{ \AA}$ ). Le cluster 2.3 groupe les conformations les plus distantes, et il est toujours suivi ou précédé du cluster 2.1 qui représente un cluster de transition (Figure 3-20). Les simulations appliquées sur l'oligonucléotide associé à 1EZN ont toutes démontré sa forte mobilité qui coïncide avec les données expérimentales.

#### 4.4.2.2 Simulations effectuées à partir du modèle de RNAdenovo depuis la structure secondaire expérimentale

Comme pour la structure expérimentale, le modèle obtenu par RNAdenovo en utilisant la structure secondaire expérimentale a été simulé en aMD et GaMD. Pour rappel, la structure de départ avait une RMSD de 12,9 Å par rapport au premier conformère de l'ensemble obtenu par RMN.

En aMD, les simulations en système neutralisé en absence NaCl ne varient pas entre les deux modèles d'eau testés. En TIP3P comme en TIP4P-Ew, l'aMD n°1 ne présente aucune conformation proche de la structure résolue expérimentalement. L'aMD n°2, en revanche, permet d'identifier des clusters majoritaires proches de la structure expérimentale. Ainsi, l'aMD n°2 en TIP3P présente un cluster majoritaire (cluster 2.1) avec une structure représentative ayant une RMSD proche de 4 Å par rapport à la structure expérimentale (Annexe 11). En TIP4P-Ew, aucun des clusters identifiés par TTclust sur l'aMD n°2 ne présente de similitude avec la structure de référence au premier abord (Annexe 12), mais l'inclusion des 25 conformères obtenus expérimentalement permet de souligner la proximité du cluster 2.2 avec 21 sur 26 conformations ( $\text{RMSD} \leq 5 \text{ \AA}$ ) (Figure 3-22).

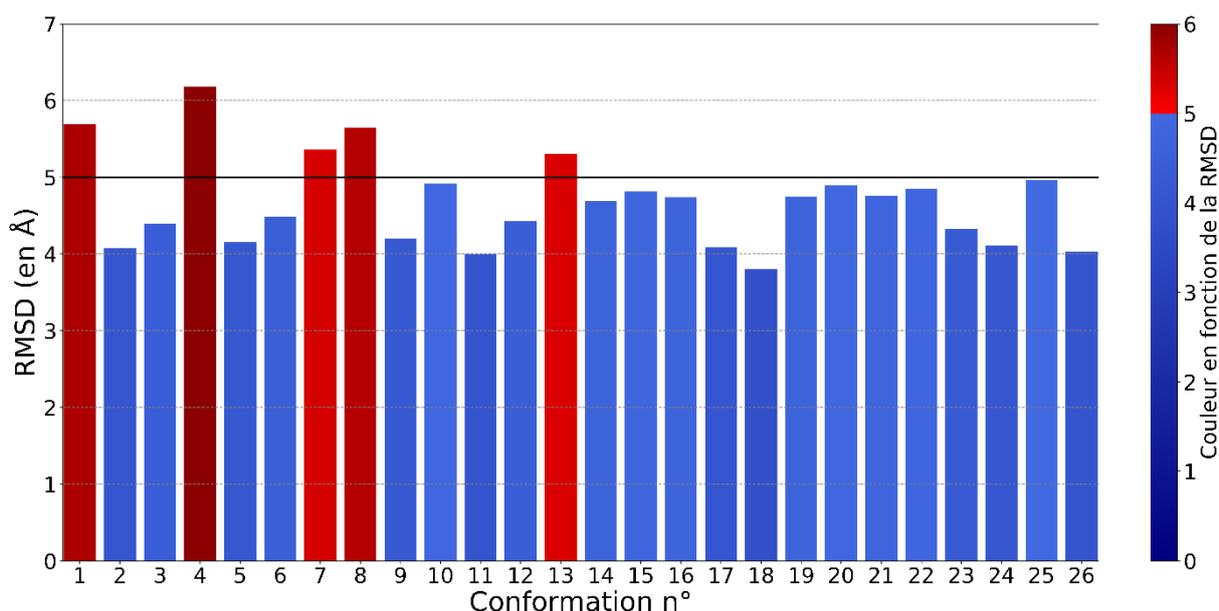


Figure 3-22. RMSD calculée (en Å) entre les différentes conformations de 1EZN expérimentale et le cluster 2.2 du modèle RNAdenovo de 1EZN, obtenue à partir de la structure secondaire expérimentale, simulé en aMD dans un système neutralisé sans NaCl et modèle d'eau TIP4P-Ew. Le gradient de couleur est décroissant avec les valeurs de RMSD : rouge pour les valeurs les plus élevées, bleu pour les valeurs les plus faibles.

Comme observé pour les simulations effectuées à partir de la structure expérimentale, pour les simulations en système incluant une concentration de 0,1 M en NaCl, indépendamment du modèle d'eau, aucun cluster majoritaire avec une structure représentative proche de la structure résolue expérimentalement ( $\text{RMSD} \leq 5 \text{ \AA}$ ) n'a pu être identifié. Néanmoins, en TIP3P, la structure représentative du cluster 1.3 est proche de 5 Å de RMSD face à la structure résolue expérimentalement. Considérant la proximité du cluster 1.3 en TIP3P à la structure expérimentale et intégrant les 25 conformations alternatives dans la comparaison (Figure 3-23), on observe une majorité de  $\text{RMSD} \leq 5 \text{ \AA}$ , soit 18 conformations sur 26. En TIP4P-Ew, aucun des clusters n'approche les 5 Å de RMSD (Annexe 16).

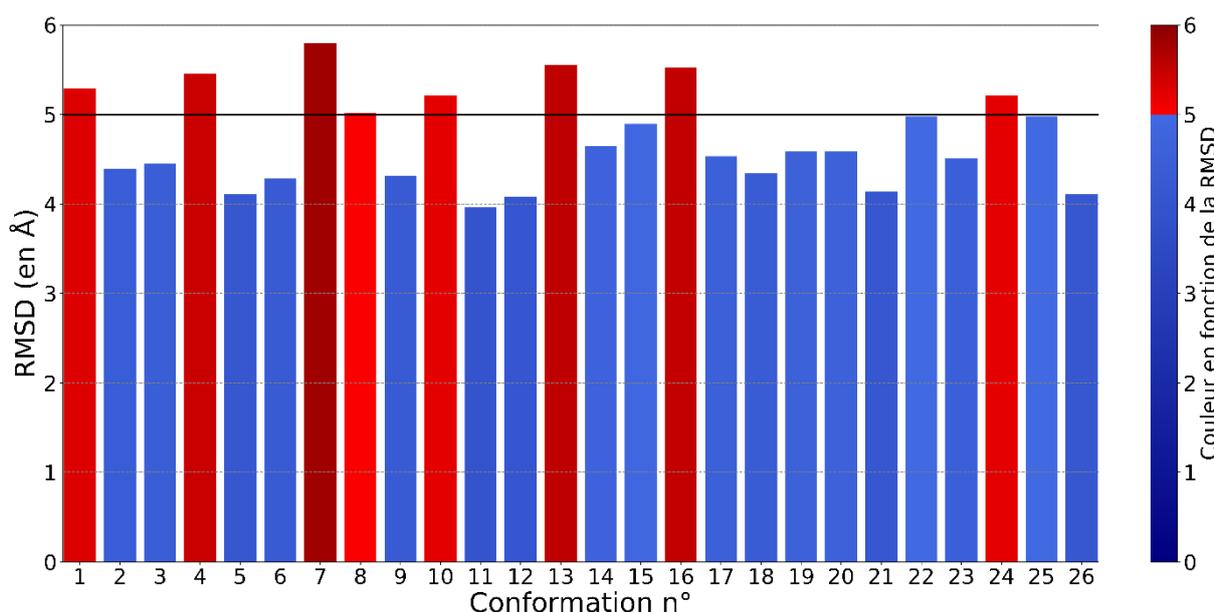


Figure 3-23. RMSD calculée (en Å) entre les différentes conformations de 1EZN expérimentale et le cluster 1.3 du modèle RNAdenovo de 1EZN, obtenue à partir de la structure secondaire expérimentale, simulé en aMD dans un système intégrant une concentration de 0,1M de NaCl et modèle d'eau TIP3P. Le gradient de couleur est décroissant avec les valeurs de RMSD : rouge pour les valeurs les plus élevées, bleu pour les valeurs les plus faibles.

L'analyse des répliques de GaMD n'a pas permis d'identifier de cluster majoritaire dont la structure représentative atteint une  $\text{RMSD} \leq 5 \text{ \AA}$  par rapport à l'expérimentale en utilisant le modèle d'eau TIP3P ou TIP4P-Ew (Annexes 18, 20). Quelques conformations tendent à s'approcher de ce seuil, notamment les clusters 2.2 et 2.3 avec des RMSD de respectivement 5,87 et 6,17 Å face à la référence. En intégrant les 25 conformations alternatives provenant de la structure RMN à la comparaison avec le cluster 2.2, la proximité est vérifiée avec 14 de ces conformations montrant une  $\text{RMSD} < 5 \text{ \AA}$  (Figure 3-24). Les autres clusters identifiés au cours des 2 GaMD montrent cependant des similitudes avec plusieurs clusters provenant des

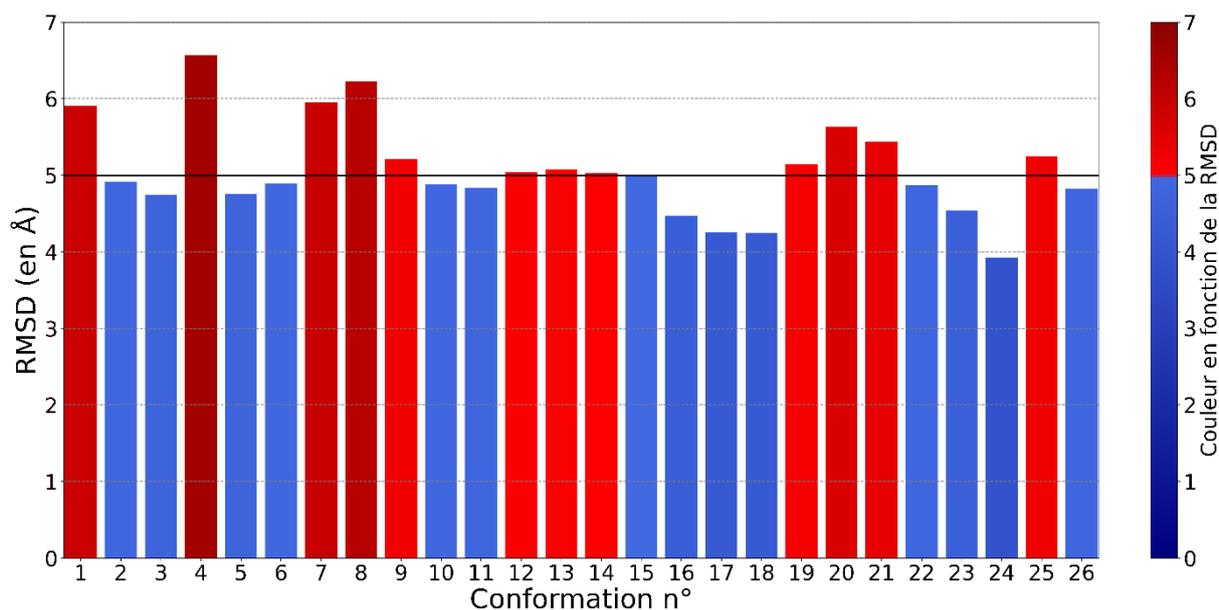


Figure 3-24. RMSD calculée (en Å) entre les différentes conformations de 1EZN expérimentale et le cluster 2.2 du modèle RNAdenovo de 1EZN, obtenue à partir de la structure secondaire expérimentale, simulé en GaMD dans un système incluant une concentration de 0,1M de NaCl et le modèle d'eau TIP3P. Le gradient de couleur est décroissant avec les valeurs de RMSD : rouge pour les valeurs les plus élevées, bleu pour les valeurs les plus faibles.

**RNAdenovo**

		GaMD n° 1				GaMD n° 2				
		Ref	C1	C2	C3	C1	C2	C3	C4	
Expérimentale	Ref	0	11,64	10,99	9,92	11,70	5,87	6,17	8,15	
	GaMD n° 1	C1	6,30	11,14	11,84	9,72	10,64	6,09	3,30	9,23
	C2	5,36	9,50	10,72	7,23	8,74	3,41	5,53	5,08	
	C3	11,76	3,28	5,57	4,71	4,06	10,20	11,53	9,41	
	C4	9,01	7,68	9,27	5,90	7,90	7,11	10,51	3,97	
	GaMD n° 2	C1	6,25	9,60	11,10	7,45	8,76	2,77	6,49	4,09
	C2	5,93	11,05	11,66	9,63	10,56	6,21	3,26	9,31	
	C3	5,65	9,79	11,26	7,75	8,97	4,56	4,18	7,21	
	C4	11,13	4,84	6,11	4,21	5,24	9,62	12,05	7,47	
	C5	8,36	8,62	9,98	6,82	8,57	6,09	9,96	3,55	

RMSD (Å)

Figure 3-25. Mesure de la RMSD entre clusters suggéré par TTclust provenant des simulations de GaMD en système incluant une concentration de 0,1 M de NaCl pour les simulations effectuées sur 1EZN expérimentale et 1EZN modélisé avec RNAdenovo en utilisant la structure secondaire expérimentale. Les clusters obtenus avec TTclust sont numérotés par prédominance dans les GaMD concernée. La structure expérimentale désignée 'Ref' est également incluse à titre comparatif.

simulations faites à partir de la structure expérimentale. Cette proximité est indiquée dans la Figure 3-25 au travers des cellules colorées en vert, indicatif d'une RMSD  $\leq 5$  Å. Par exemple, la structure représentative du cluster 1.1 montre une RMSD de 3,279 Å par rapport à la structure représentative du cluster 1.3 des simulations effectuées à partir de la structure expérimentale. Tous deux ont été identifiés comme conformations distantes de la référence ( $\sim 12$  Å).

En second exemple, le cluster 2.2 identifié dans les simulations faites sur le modèle RNAdenovo obtenu à partir de la structure secondaire expérimentale est proche des clusters 1.2, 2.1 et 2.3 extraits des simulations faites en partant de la structure expérimentale. A l'exception du cluster 1.2, chaque cluster identifié dans les simulations sur le modèle RNAdenovo généré à partir de la structure secondaire exacte dispose d'au moins un cluster proche dans les simulations sur structure expérimentale. Cette proximité identifiée durant les dynamiques moléculaire, combinée aux similitudes observées également avec les autres conformations obtenues en RMN, amène à la conclusion que les simulations ont exploré les mêmes conformations, qui malgré une certaine distance par rapport à la structure expérimentale ne sont que la représentation de la mobilité de l'oligonucléotide.

#### 4.4.2.3 Simulations effectuées à partir du modèle de RNAdenovo depuis la structure secondaire prédite par mfold

Après le modèle RNAdenovo obtenu utilisant la structure secondaire expérimentale, les simulations ont été effectuées en partant du modèle RNAdenovo généré avec la structure secondaire prédite par mfold. Sur cette prédiction de structure secondaire la différence  $Apta_D$  identifiée est faible, mais le modèle RNAdenovo obtenu et utilisé comme structure de départ pour les simulations a montré une RMSD de 12,6 Å face à la structure expérimentale. Donc, des changements conformationnels sont attendus pendant les GaMD pour pouvoir s'approcher de la structure expérimentale.

Parmi les 7 clusters obtenus dans les deux GaMD en TIP3P, 2 sont proches de la structure de référence ( $\leq 5$  Å) soit les clusters 1.2 (4,78 Å) et 2.2 (4,86 Å) (Annexe 20). Ces deux clusters obtenus séparément dans les GaMD n°1 et n°2 sont quasi-identiques avec une RMSD de 1,6 Å entre eux. En GaMD n°1, les 4 clusters obtenus sont distribués équitablement, mais seul le cluster 1.1 se maintient dans la dynamique au regard de l'association cluster et conformations

associées (Figure 3-26). Les clusters 1.2, et 1.4 se manifestent de manière éparse mais sont toujours adjacents dans la simulation et ils sont représentatifs de conformations relativement proches de l'expérimental (RMSD = 4,78 Å et 5,64 Å, respectivement). Le cluster 1.3 est, quant à lui, représentatif des états intermédiaires, avec une RMSD plus élevée, qui séparent notamment les clusters 1.1, 1.2 et 1.4 et qui se manifestent plusieurs fois dans la dynamique (0-100 ns, 340-400 ns, 470-500 ns, 650-690 ns, 980 ns -1  $\mu$ s). En GaMD n°2, le cluster 1.1, qui représente ~50 % de la dynamique, est légèrement différent de l'expérimental (RMSD = 6,16 Å) et apparaît après un changement de conformation du cluster 2.2, proche de l'expérimental (RMSD = 4,86 Å).

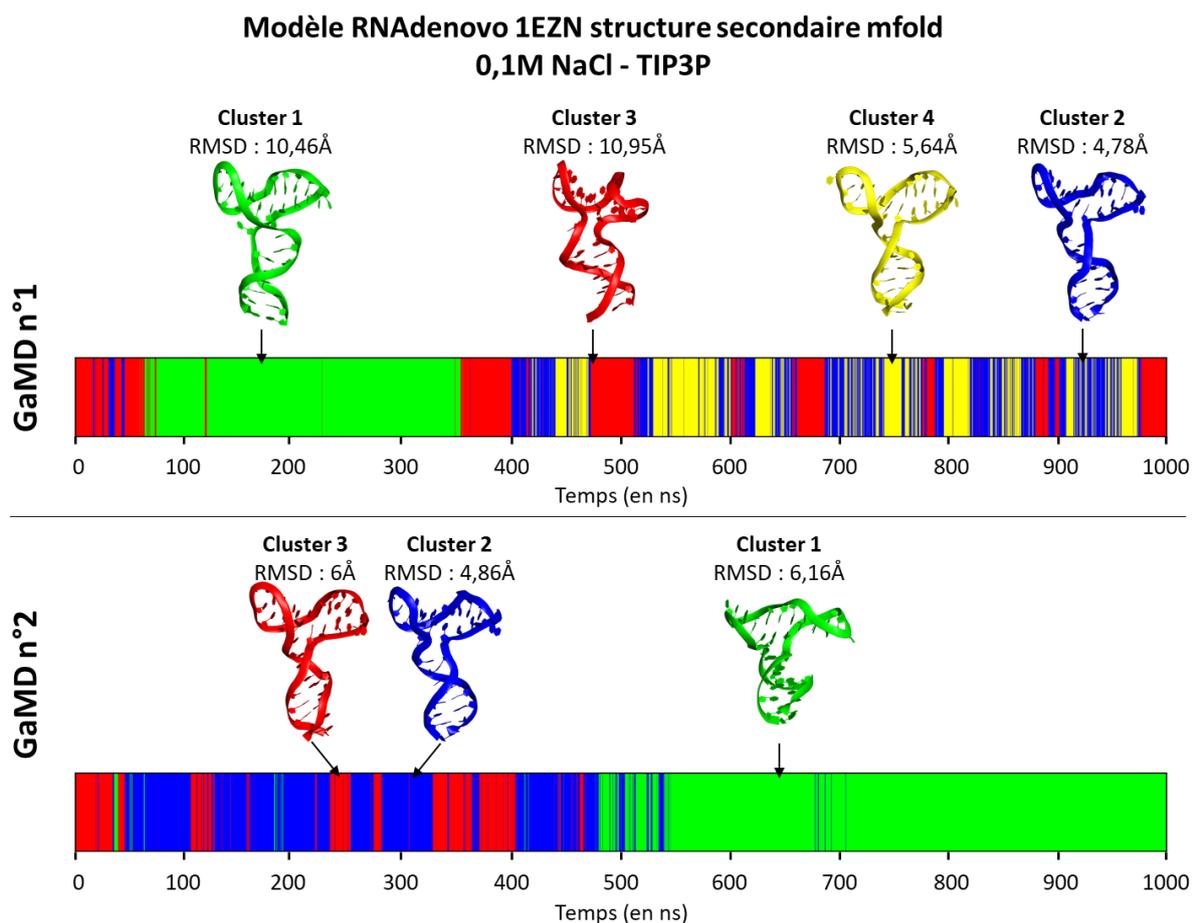


Figure 3-26. Distribution des clusters et des conformations associées au cours des GaMD simulés avec le modèle d'eau TIP3P sur le modèle RNAdenovo de 1EZN utilisant la structure secondaire de mfold. Chaque couleur représente un cluster différent. Le numéro du cluster est assigné selon son pourcentage d'occurrence dans la simulation. En haut est représentée la GaMD n°1, en bas la GaMD n°2.

Les modèles obtenus par RNAdenovo à partir soit de la structure secondaire correcte soit de celle prédite par mfold étaient très éloignés de la structure expérimentale (RMSD = 12,9 Å et 12,61 Å, respectivement). Cependant, les résultats détaillés ici ont souligné, notamment pour le modèle d'eau TIP3P, la capacité du protocole d'échantillonner des conformations proches de l'expérimentale, mais aussi d'explorer différentes conformations avec plusieurs clusters différents (RMSD > 5Å) qui se maintiennent. Les simulations effectuées en TIP4P-Ew ne permettent pas d'identifier de proximité avec la structure expérimentale car l'ensemble des clusters présente une RMSD > 5 Å (Annexe 22). Le cluster 1.4 est le plus proche identifié mais n'est pas représentatif d'une conformation stable (4 % de la dynamique). La moindre fluidité du modèle d'eau TIP4P-Ew et la différente distribution des charges ponctuelles par rapport au modèle TIP3P et la présence d'une concentration de NaCl supérieure à celle expérimentale, dans ce cas, peuvent avoir limité l'exploration de conformations proches de la structure expérimentale.

#### 4.4.3 3HXO

3HXO représente la structure cristallographique d'un aptamère d'ADN en complexe avec sa cible, le facteur de Von Willebrand A1. Elle se présente comme une structure de 40 nucléotides composée d'une jonction à trois branches. La structure expérimentale provenant d'un cristal, une seule conformation est disponible pour 3HXO. Cependant, comme pour 1EZN, on s'attend à avoir plusieurs conformations accessibles, au vu de la nature mobile des jonctions à trois branches (Lescoute & Westhof, 2006). La prédiction par mfold de la structure secondaire avait donné une distance AptaMat > 10, par conséquent 3 structures ont été utilisées comme départ des simulations : la structure cristallographique, le modèle RNAdenovo réalisé à partir de la structure secondaire expérimentale et le modèle RNAdenovo obtenu à partir de la structure secondaire prédite.

Cluster		aMD n°1				aMD n°2				
		n°1	n°2	n°3	n°4	n°1	n°2	n°3	n°4	n°5
NaCl 0M	TIP3P	 5,49 Å <sup>a</sup> 50%	 4,58 Å <sup>a</sup> 27%	 6,85 Å <sup>a</sup> 15%	 10,83 Å <sup>a</sup> 8%	 3,23 Å <sup>a</sup> 39%	 6,08 Å <sup>a</sup> 35%	 4,02 Å <sup>a</sup> 26%		
	TIP4P-EW	 3,04 Å <sup>a</sup> 40%	 2,36 Å <sup>a</sup> 31%	 3,88 Å <sup>a</sup> 18%	 3,43 Å <sup>a</sup> 11%	 2,89 Å <sup>a</sup> 41%	 3,08 Å <sup>a</sup> 34%	 4,55 Å <sup>a</sup> 25%		
NaCl 0,1 M	TIP3P	 6,76 Å <sup>a</sup> 52%	 3,46 Å <sup>a</sup> 48%			 5,10 Å <sup>a</sup> 53%	 6,60 Å <sup>a</sup> 43%	 7,59 Å <sup>a</sup> 3%		
	TIP4P-EW	 3,80 Å <sup>a</sup> 62%	 2,70 Å <sup>a</sup> 38%			 3,84 Å <sup>a</sup> 32%	 5,36 Å <sup>a</sup> 24%	 4,73 Å <sup>a</sup> 20%	 3,35 Å <sup>a</sup> 14%	 4,47 Å <sup>a</sup> 9%

Figure 3-27. Alignement des clusters des différentes simulations aMD effectuées en partant de la structure résolue expérimentalement de 3HXO en système neutralisé avec et sans concentration de 0,1 M de NaCl testé avec les modèles d'eau TIP3P ou TIP4P-Ew. <sup>a</sup> La RMSD obtenue face à la structure expérimentale est précisée pour chaque alignement.

#### 4.4.3.1 Simulations effectuées à partir de la structure expérimentale

Dans un premier temps, les simulations en aMD à partir de la structure expérimentale de 3HXO ont été effectuées (Figure 3-27). Les simulations avec le modèle d'eau TIP3P démontrent la stabilité des conformations qui sont proches de la structure résolue expérimentalement ainsi que l'existence de conformations alternatives plus éloignées de la structure expérimentale, que ce soit en présence ou pas de 0,1 M de NaCl (Figure 3-27, Annexes 9, 13).

Le modèle d'eau TIP4P-Ew appliqué aux systèmes avec et sans 0,1 M NaCl montre une forte stabilité de la structure expérimentale au travers des différents clusters identifiés par TTclust, dont les structures représentatives ont toutes une faible RMSD (Annexe 10, 14). Cela n'est pas surprenant, car les caractéristiques du modèle d'eau TIP4P-Ew semblent impacter négativement l'échantillonnage du système (Figure 3-27).

En effet, les différences de résultats entre les deux modèles se vérifient également pour les simulations GaMD faites à partir de la structure expérimentale (Figure 3-28). La première GaMD de 3HXO avec le modèle TIP3P explore des conformations différentes de la référence avec une RMSD oscillant entre 7 et 8 Å, ainsi qu'une conformation relativement proche de l'expérimental (cluster 1.4), dont la structure représentative a une RMSD de 5,16 Å par rapport à l'expérimental. La seconde GaMD échantillonne uniquement des conformations proches de la référence, avec la totalité des clusters ayant une structure représentative dont le RMSD est sous le seuil de 5 Å (Annexe 17). On se pose donc la question de l'efficacité de cette deuxième simulation, car les biais énergétiques introduits ne semblent pas être suffisants pour l'exploration de l'espace conformationnel de 3HXO, comme c'était le cas pour la GaMD n°1. Cela démontre la nécessité d'effectuer des répliques indépendantes d'une même simulation.

En TIP4P-Ew, les mêmes observations faites pour les simulations aMD sont valables et seule la seconde GaMD est capable d'identifier une conformation dont la distance avec la référence est  $\sim 5$  Å, se situant donc à la limite du seuil fixé (Annexe 18). Toutes les autres conformations échantillonnées sont très proches de l'expérimental (Figure 3-28).

Le modèle RNAdenovo de 3HXO obtenu en partant de la structure secondaire expérimentale a une RMSD par rapport à la structure cristallographique de 12,3 Å, qui indique un modèle très éloigné de l'expérimentale.

#### 4.4.3.2 Simulations effectuées à partir du modèle RNAdenovo obtenu depuis la structure secondaire expérimentale

Les simulations en aMD ont abouti à des résultats différents en fonction de la combinaison de paramètres. Plus précisément, pour les simulations effectuées en système neutralisé sans NaCl et avec le modèle d'eau TIP3P, on observe quelques clusters proches de la structure résolue expérimentalement ( $\text{RMSD} \leq 5$  Å, soit les clusters 1.4, 2.1 et 2.3, 2.5). Ces excellents résultats montrent que les simulations aMD sont capables d'échantillonner de manière significative ( $\sim 26$  % et  $\sim 19$  % de la trajectoire pour les clusters 2.1 et 2.3 respectivement) des conformations proches de l'expérimental, bien que la structure de départ soit très éloignée de cette dernière.

## Structure expérimentale 3HXO 0,1M NaCl

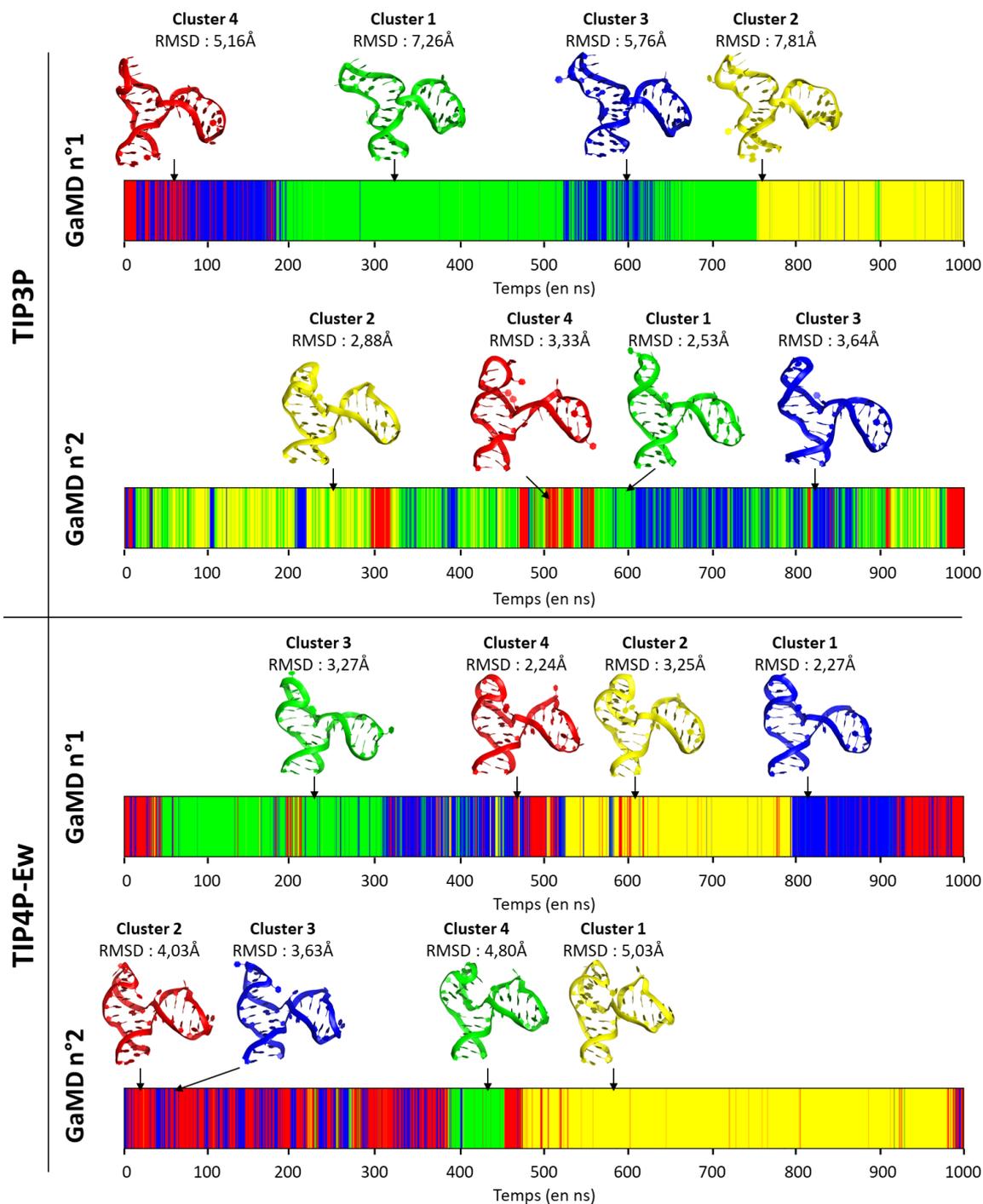


Figure 3-28. Distribution des clusters et des conformations associées au cours des GaMD simulées dans un système concentré à 0,1 M en NaCl avec différents modèles d'eau (TIP3P ou TIP4P-Ew) appliqués sur la structure expérimentale de 3HXO. Chaque couleur représente un cluster différent. Le numéro du cluster est assigné selon son pourcentage d'occurrence dans la simulation.

		aMD n°1					aMD n°2				
Cluster		n°1	n°2	n°3	n°4	n°5	n°1	n°2	n°3	n°4	n°5
NaCl 0M	TIP3P	 6,06 Å <sup>a</sup> 41%	 11,11 Å <sup>a</sup> 35%	 10,58 Å <sup>a</sup> 13%	 3,84 Å <sup>a</sup> 11%		 3,53 Å <sup>a</sup> 26%	 5,48 Å <sup>a</sup> 23%	 4,53 Å <sup>a</sup> 19%	 6,82 Å <sup>a</sup> 18%	 3,17 Å <sup>a</sup> 14%
	TIP4P-EW	 11,41 Å <sup>a</sup> 54%	 5,51 Å <sup>a</sup> 18%	 6,37 Å <sup>a</sup> 14%	 11,87 Å <sup>a</sup> 8%	 12,48 Å <sup>a</sup> 6%		 3,48 Å <sup>a</sup> 41%	 2,74 Å <sup>a</sup> 26%	 8,78 Å <sup>a</sup> 24%	 8,65 Å <sup>a</sup> 9%
NaCl 0,1 M	TIP3P	 4,02 Å <sup>a</sup> 46%	 6,45 Å <sup>a</sup> 19%	 6,25 Å <sup>a</sup> 16%	 4,35 Å <sup>a</sup> 16%	 6,55 Å <sup>a</sup> 3%	 7,99 Å <sup>a</sup> 40%	 7,74 Å <sup>a</sup> 27%	 4,39 Å <sup>a</sup> 18%	 3,97 Å <sup>a</sup> 15%	
	TIP4P-EW	 7,51 Å <sup>a</sup> 43%	 6,13 Å <sup>a</sup> 30%	 6,85 Å <sup>a</sup> 17%	 10,06 Å <sup>a</sup> 10%			 10,28 Å <sup>a</sup> 41%	 11,91 Å <sup>a</sup> 28%	 11,11 Å <sup>a</sup> 28%	 10,17 Å <sup>a</sup> 3%

Figure 3-29. Alignement des clusters des différentes simulations aMD effectuées en partant du modèle RNAdenovo de 3HXO obtenu à partir de la structure secondaire expérimentale en système neutralisé avec et sans concentration de 0,1 M de NaCl testé avec les modèles d'eau TIP3P ou TIP4P-Ew. <sup>a</sup> La RMSD obtenue face à la structure expérimentale est précisée pour chaque alignement.

De plus, la présence de clusters plus éloignés et avec une occurrence supérieure démontre également la capacité à explorer l'espace conformationnel, avec par exemple le cluster 1.1, qui a une structure représentative avec une RMSD de 6 Å et qui est représentatif de ~40 % de l'aMD n°1. On note un comportement similaire des simulations aMD en TIP4P-Ew qui montrent d'un côté la stabilité des conformations proche de la structure résolue expérimentalement et de l'autre la capacité de la technique à explorer l'espace conformationnel (Figure 3-29, Annexe 12).

L'ajout de 0,1 M de NaCl dans le système ne change pas fondamentalement le comportement dynamique du système en TIP3P par rapport au système seulement neutralisé. En revanche, aucune des répliques indépendantes des aMD avec le modèle d'eau TIP4P-Ew n'a permis de retrouver une conformation proche de la structure obtenue expérimentalement, avec pour chaque cluster identifié une structure représentative ayant une RMSD > 6 Å face à cette structure (Annexe 12 et 14).

## Modèle RNAdenovo 3HXO structure secondaire expérimentale 0,1M NaCl - TIP3P

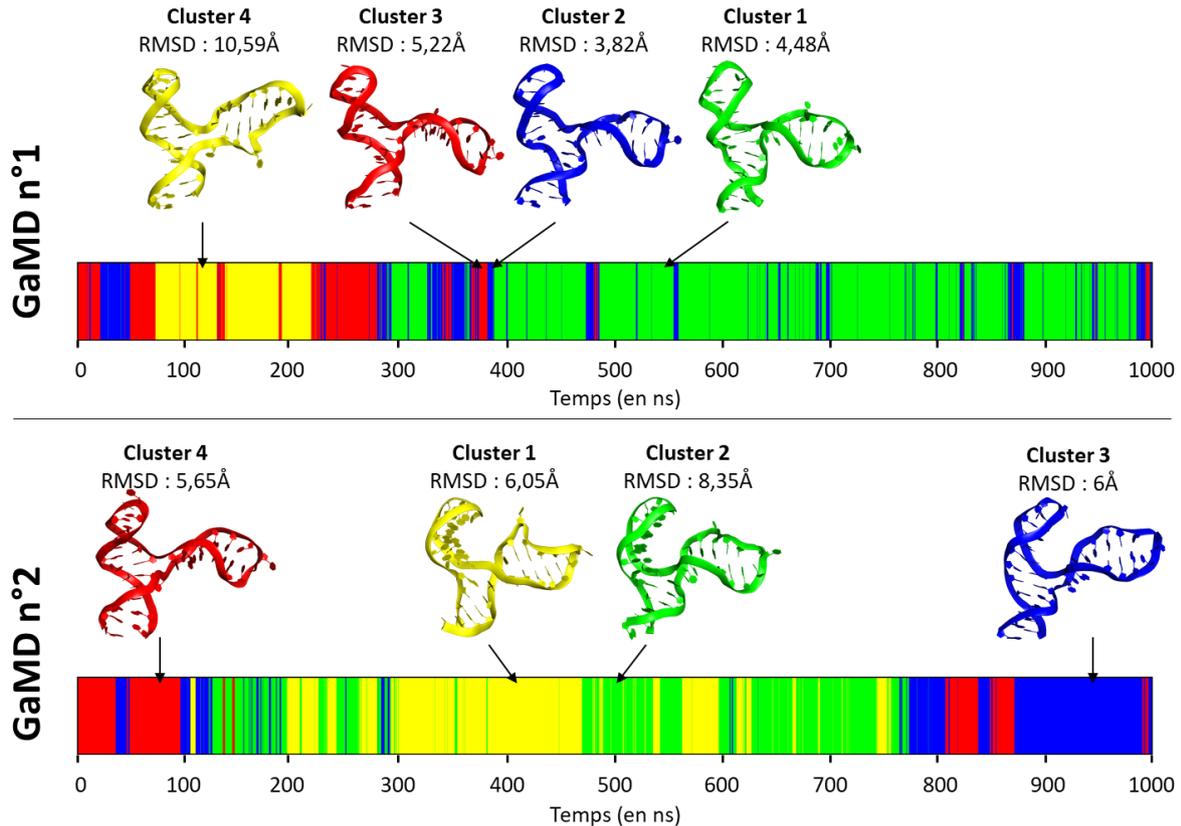


Figure 3-30. Distribution des clusters et des conformations associées au cours des GaMD simulés dans un système concentré à 0,1 M en NaCl avec le modèle d'eau TIP3P appliqué sur le modèle RNAdenovo de 3HXO utilisant la structure secondaire expérimentale. Chaque couleur représente un cluster différent. Le numéro du cluster est assigné selon son pourcentage d'occurrence dans la simulation. En haut est représentée la GaMD n°1, en bas la GaMD n°2.

Pour continuer l'analyse, les simulations suivantes ont été effectuées en GaMD sur le système avec une concentration de NaCl de 0,1 M pour vérifier la validité des observations faites précédemment. Les résultats obtenus sont similaires aux simulations en aMD dans les mêmes conditions. La première simulation GaMD faite avec le modèle TIP3P montre une conformation majoritaire proche de la structure de référence et présente dans le cluster 1.1, représentatif de 56 % de la dynamique (RMSD ~4,5 Å) et dans le cluster 1.2 regroupant 16 % de la trajectoire (~3,7 Å). Les autres clusters avoisinent 15 % d'occupation et ne semblent pas se maintenir au regard de la Figure 3-30. La présence du cluster 1.1 est marquée entre 400 ns jusqu'à la fin de la simulation, avec quelques instants de prédominance du cluster 1.2. La seconde GaMD explore quelques conformations additionnelles, dont l'augmentation de la

RMSD est majoritairement influencée par la mobilité des épingles à cheveux G6-C19 et G22-T35 formant la jonction à deux branches.

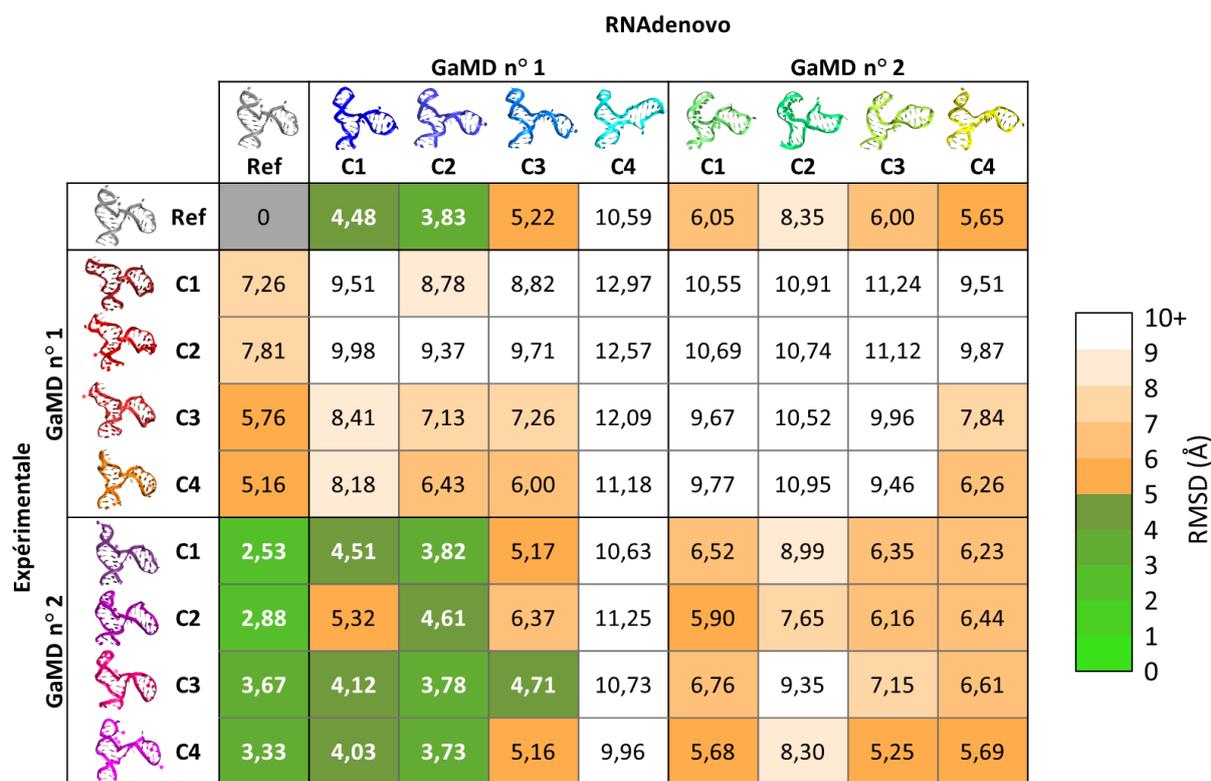


Figure 3-31. Mesure de la RMSD entre clusters suggéré par TTclust de 3HXO expérimentale et 3HXO modélisé avec RNAdenovo obtenu en utilisant la structure secondaire expérimentale simulés en GaMD incluant une concentration de 0,1 M de NaCl avec le modèle d'eau TIP3P. Les clusters obtenus avec TTclust sont numérotés par prédominance dans les GaMD concernée. La structure expérimentale désignée 'Ref' est également incluse à titre comparatif.

Comme attendu au vu de la structure de départ et des caractéristiques du modèle d'eau, les simulations GaMD avec le modèle TIP4-Ew sont majoritairement éloignées de la structure résolue expérimentalement. Néanmoins, il faut remarquer que le cluster 2.2 s'approche du seuil de 5 Å (Annexes 21).

Globalement, il est possible de conclure que, malgré une structure initiale très éloignée de la structure expérimentale, les simulations aMD et GaMD sont capables de retrouver cette dernière si le modèle solvant utilisé est le TIP3P et que l'effet des ions est moindre par rapport à l'effet du modèle d'eau. Il est probable que des simulations plus longues ou avec un biais plus élevé soient nécessaires si le modèle de solvant utilisé est le TIP4P-Ew.

L'échantillonnage des conformations alternatives devrait être similaire entre les simulations GaMD à partir de la structure expérimentale et du modèle généré par RNAde novo pour démontrer la fiabilité du protocole. Dans ce contexte, on remarque la proximité de certains clusters obtenus des simulations conduites à partir du modèle de RNAde novo à d'autres obtenus des simulations faites à partir de structures expérimentales, soit les clusters 1.1 et 1.2 des GaMD depuis le modèle de RNAde novo, proches de l'ensemble des clusters en GaMD n°2 depuis la structure expérimentale ( $\text{RMSD} \leq 5 \text{ \AA}$ ) (Figure 3-31). Cependant les clusters 1.1 et 1.2 sont déjà des conformations similaires à la structure expérimentale (Annexe 19), et n'apportent pas de nouvelle information. Aucun autre cluster obtenu des simulations faites à partir du modèle de RNAde novo ne montre de proximité avec ceux obtenus en commençant les simulations de la structure expérimentale, ce qui indique que l'espace conformationnel exploré n'est pas le même entre les deux groupes de simulations. Cela ouvre la question sur l'exploration complète de l'espace conformationnel de 3HXO et sur la nécessité d'effectuer de plus longues simulations et d'augmenter le biais utilisé pour atteindre la convergence entre les simulations indépendamment des coordonnées de la structure de départ.

#### 4.4.3.3 Simulations effectuées à partir du modèle RNAde novo obtenu depuis la structure secondaire prédite par mfold

Cela est aussi questionné au vu des simulations GaMD effectuées à partir du modèle RNAde novo obtenu depuis la structure secondaire prédite par mfold, qui a une RMSD de 16,6 Å par rapport à la structure expérimentale (Figure 3-14). Dans ce cas, afin de s'approcher de la structure de référence, il faut observer un changement conformationnel encore plus important que celui nécessaire pendant les simulations faites à partir du modèle RNAde novo obtenu à partir de la structure secondaire expérimentale. La distribution de la RMSD au cours des 2 dynamiques GaMD ne permet pas d'identifier d'événement relatif à un changement de conformation de cette ampleur, que ce soit en utilisant le modèle d'eau TIP3P ou TIP4P-Ew (Figure 3-33). De plus la totalité des conformations explorées au cours des simulations présentent des valeurs de  $\text{RMSD} > 15 \text{ \AA}$  par rapport à l'expérimental (Annexes 20 et 22). Ainsi les clusters les plus importants de chaque GaMD en TIP3P et TIP4P-Ew ne s'alignent pas avec la référence comme montré en Figure 3-32.

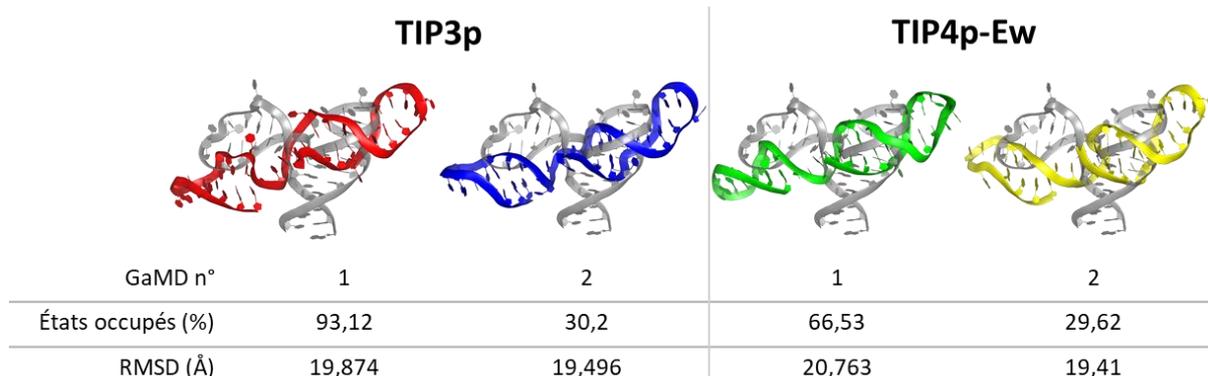


Figure 3-32. Alignements de la structure de 3HXO résolue expérimentalement (en gris) avec les clusters n°1 obtenues avec TTclust provenant des GaMD sur 3HXO modélisé avec RNAdenovo en utilisant la structure secondaire de mfold (en rouge, bleu, vert ou jaune). Sont différenciés les résultats de GaMD avec le modèle d'eau TIP3P ou TIP4P-Ew. Le pourcentage de temps d'occupation dans la GaMD ainsi que la RMSD face à la structure résolue expérimentalement sont indiqués pour chaque structure.

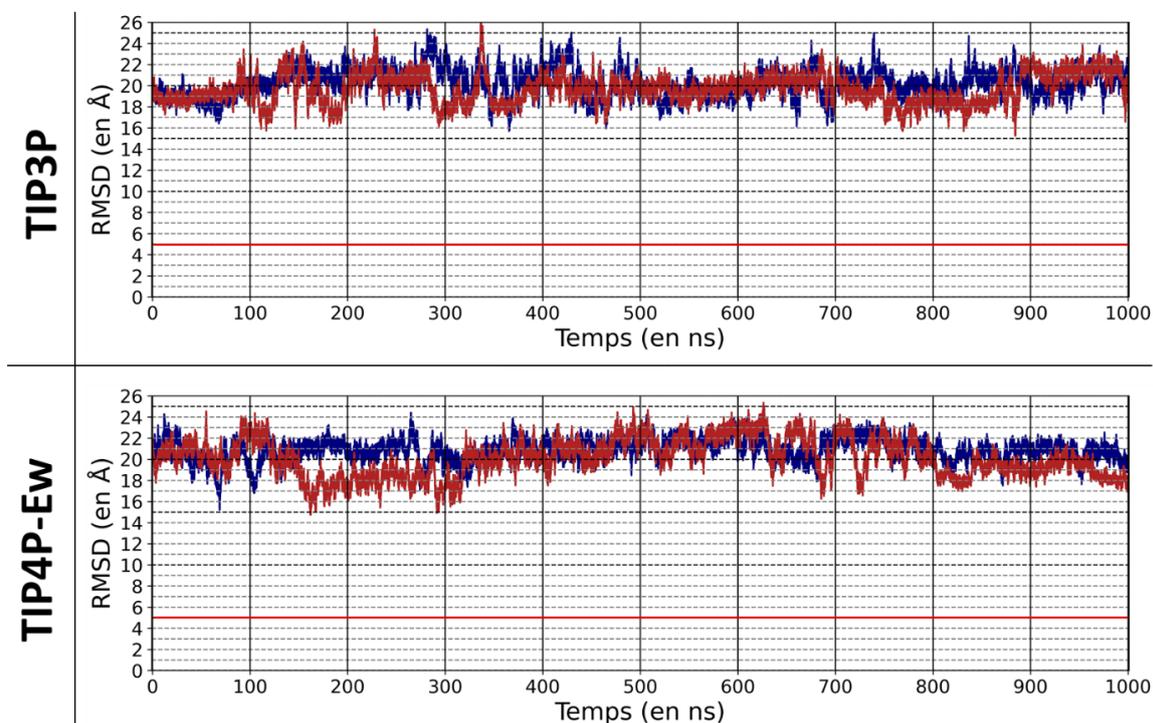


Figure 3-33. RMSD de 3HXO modélisé avec RNAdenovo utilisant la structure secondaire de mfold mesurée pour les deux modèles d'eau TIP3P et TIP4P-Ew, face à la référence expérimentale le long des dynamiques de GaMD n°1 (bleu) et GaMD n°2 (rouge).

#### 4.4.4 3THW

3THW correspond au complexe entre un oligonucléotide d'ADN simple brin et la MutS $\beta$  humaine (Gupta et al., 2012) dont la structure a été obtenue par cristallographie aux rayons X. Cet oligonucléotide est composé de 53 nucléotides structurés en une longue hélice interrompue par un bourgeon responsable d'une flexion de la structure. Sa structure

secondaire est bien prédite, mais le modèle de RNAdenovo qui en résulte en une RMSD de 10,3 Å par rapport à l'expérimentale. Cette grande différence conformationnelle est due à la différence d'orientation relative des deux régions hélicoïdales interrompues par le bourgeon (Figure 3-14). Les simulations ont donc été conduites à partir soit de la structure cristallographique soit du modèle obtenu par RNAdenovo.

#### 4.4.4.1 Simulations effectuées à partir de la structure expérimentale

		aMD n°1					aMD n°2					
Cluster		n°1	n°2	n°3	n°4	n°5	n°6	n°1	n°2	n°3	n°4	n°5
NaCl 0M	TIP3P	 6,63 Å <sup>a</sup> 62%	 4,79 Å <sup>a</sup> 18%	 10,84 Å <sup>a</sup> 11%	 8,32 Å <sup>a</sup> 9%			 5,39 Å <sup>a</sup> 38%	 7,53 Å <sup>a</sup> 37%	 3,92 Å <sup>a</sup> 25%		
	TIP4P-EW	 7,10 Å <sup>a</sup> 31%	 6,88 Å <sup>a</sup> 24%	 5,66 Å <sup>a</sup> 23%	 8,76 Å <sup>a</sup> 11%	 10,22 Å <sup>a</sup> 11%			 4,25 Å <sup>a</sup> 52%	 5,04 Å <sup>a</sup> 31%	 5,51 Å <sup>a</sup> 9%	 7,65 Å <sup>a</sup> 8%
NaCl 0,1 M	TIP3P	 6,33 Å <sup>a</sup> 43%	 6,29 Å <sup>a</sup> 25%	 8,13 Å <sup>a</sup> 15%	 4,74 Å <sup>a</sup> 9%	 10,44 Å <sup>a</sup> 7%	 13,68 Å <sup>a</sup> 1%	 9,37 Å <sup>a</sup> 28%	 9,09 Å <sup>a</sup> 26%	 4,72 Å <sup>a</sup> 19%	 3,91 Å <sup>a</sup> 14%	 11,82 Å <sup>a</sup> 13%
	TIP4P-EW	 4,09 Å <sup>a</sup> 39%	 5,82 Å <sup>a</sup> 30%	 3,97 Å <sup>a</sup> 20%	 8,84 Å <sup>a</sup> 11%			 5,30 Å <sup>a</sup> 33%	 4,41 Å <sup>a</sup> 31%	 10,15 Å <sup>a</sup> 20%	 3,47 Å <sup>a</sup> 16%	

Figure 3-34. Alignement des clusters des différentes simulations aMD effectuées en partant de la structure résolue expérimentalement de 3THW en système neutralisé avec et sans concentration de 0,1 M de NaCl avec les modèles d'eau TIP3P ou TIP4P-Ew. <sup>a</sup> La RMSD obtenue face à la structure expérimentale est précisée pour chaque alignement ainsi que le pourcentage d'occupation dans la simulation.

Pour ce qui concerne les simulations aMD effectuées à partir de la structure expérimentale, indépendamment des conditions utilisées, au moins un cluster majoritaire représentatif de la structure résolue expérimentalement (RMSD ≤ 5 Å) a été identifié (Figure 3-34, Annexes 9, 10, 11, 12). Seul l'aMD n°1 incluant une concentration de 0,1 M de NaCl avec le modèle d'eau TIP3P ne montre pas de cluster majoritaire et stable proche de la structure expérimentale, mais, en contrepartie, l'aMD n°2 maintient 2 conformations proches de l'expérimentale (Annexe 9). Les autres clusters plus éloignés de la structure de référence (RMSD > 5 Å) présents dans ces dynamiques sont de potentielles conformations faisant partie de l'espace

conformationnel de cet oligonucléotide, impliquant la flexibilité au niveau du bourgeon. Ces résultats semblent être cohérents avec le fait que la flexion des hélices due au bourgeon semble être stabilisée par la présence de la protéine (Figure 3-15).

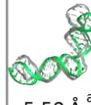
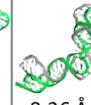
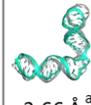
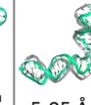
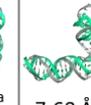
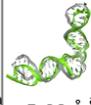
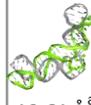
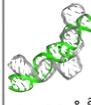
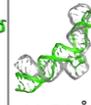
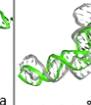
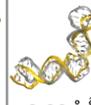
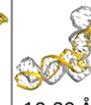
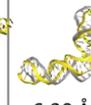
		GaMD n°1					GaMD n°2					
Cluster		n°1	n°2	n°3	n°4	n°5	n°1	n°2	n°3	n°4	n°5	n°6
Structure expérimentale	TIP3P	 5,80 Å <sup>a</sup> 40%	 8,64 Å <sup>a</sup> 38%	 2,98 Å <sup>a</sup> 22%			 3,70 Å <sup>a</sup> 56%	 6,49 Å <sup>a</sup> 16%	 10,91 Å <sup>a</sup> 14%	 13,18 Å <sup>a</sup> 14%		
	TIP4P-EW	 3,43 Å <sup>a</sup> 35%	 3,34 Å <sup>a</sup> 22%	 4,29 Å <sup>a</sup> 20%	 5,58 Å <sup>a</sup> 15%	 8,26 Å <sup>a</sup> 8%	 2,55 Å <sup>a</sup> 39%	 3,52 Å <sup>a</sup> 26%	 3,66 Å <sup>a</sup> 22%	 6,14 Å <sup>a</sup> 5%	 5,05 Å <sup>a</sup> 5%	 7,68 Å <sup>a</sup> 3%
RNAde novo	TIP3P	 12,67 Å <sup>a</sup> 40%	 3,13 Å <sup>a</sup> 22%	 5,09 Å <sup>a</sup> 21%	 10,61 Å <sup>a</sup> 17%			 12,34 Å <sup>a</sup> 31%	 10,11 Å <sup>a</sup> 24%	 11,89 Å <sup>a</sup> 21%	 10,50 Å <sup>a</sup> 14%	 10,61 Å <sup>a</sup> 10%
	TIP4P-EW	 3,98 Å <sup>a</sup> 34%	 4,18 Å <sup>a</sup> 32%	 5,27 Å <sup>a</sup> 26%	 6,66 Å <sup>a</sup> 6%	 10,89 Å <sup>a</sup> 2%	 8,15 Å <sup>a</sup> 36%	 7,00 Å <sup>a</sup> 32%	 10,77 Å <sup>a</sup> 29%	 6,29 Å <sup>a</sup> 3%		

Figure 3-35. Alignement des clusters des différentes simulations GaMD effectuées en partant de la structure résolue expérimentalement ou le modèle RNAde novo obtenu à partir de la structure secondaire expérimentale de 3THW. Toutes ces simulations ont été effectuées en système avec concentration de 0,1 M de NaCl testé avec les modèles d'eau TIP3P ou TIP4P-EW. <sup>a</sup> La RMSD obtenue face à la structure expérimentale est précisée pour chaque alignement ainsi que le pourcentage d'occupation dans la simulation.

Les simulations GaMD avec le modèle d'eau TIP3P à partir de la structure expérimentale montrent le même comportement. Ainsi, on identifie les clusters majoritaires proches de la structure résolue expérimentalement, soit les cluster 1.3 et cluster 2.1, et d'autres conformations représentatives de l'espace conformationnel accessible (Figure 3-35, Annexe 17). En revanche, les simulations conduites en utilisant le modèle TIP4P-EW ont permis d'identifier une majorité de clusters proches de la structure expérimentale avec des valeurs de RMSD inférieurs ou proche de 5 Å et seul un nombre de clusters minoritaires représentant moins de 10 % des dynamiques a montré des déviations plus significatives par rapport à la structure expérimentale (Figure 3-35, Annexe 18).

#### 4.4.4.2 Simulations effectuées à partir du modèle RNAdenovo généré depuis la structure secondaire expérimentale

Les simulations effectuées en partant du modèle RNAdenovo de 3THW basé sur la structure secondaire expérimentale ont permis de souligner des différences entre aMD et GaMD. En aMD il semble que les systèmes neutralisés, sans ajout de NaCl, indépendamment du modèle d'eau, ne montrent aucun cluster proche de la structure résolue expérimentalement ( $\leq 5 \text{ \AA}$ ) (Annexes 11, 12). De manière similaire, les simulations incluant une concentration de 0,1 M de NaCl ne présentent aucun cluster majoritaire reproduisant la référence, bien que les simulations parviennent à explorer quelques conformations proches de la structure expérimentale. Ainsi le cluster 2.3 pour les simulations en modèle d'eau TIP3P montre une RMSD de 4,59 Å avec la structure expérimentale, et le cluster 1.4 pour les simulations en modèle d'eau TIP4P-Ew montre une RMSD de 3,99 Å avec la structure expérimentale. Ces deux clusters ne sont cependant pas maintenus dans la simulation et ne peuvent pas être représentatifs d'une conformation stable, suggérant la nécessité de simulations plus longues.

La GaMD n°1 effectuée avec le modèle TIP3P a permis de retrouver la conformation expérimentale au travers de 2 clusters importants qui représentent chacun  $\sim 21 \%$  de la GaMD, soit les clusters 1.2 et 1.3 présentant une RMSD relativement faible avec la structure de référence, respectivement de 3,13 Å et 5,09 Å (Figure 3-35). Les clusters 1.1 et 1.4, plus éloignés de la structure de référence, montrent une proximité mutuelle ( $\sim 3 \text{ \AA}$ ) ainsi qu'avec les clusters de la GaMD n°2, comme les clusters 2.2 ou 2.4 à hauteur de 3 à 5 Å de RMSD (Figure 3-36). Leur occurrence durant la dynamique montre que ces conformations distantes se maintiennent pendant les 600 premières ns, puis un brusque changement conformationnel visible sur le graphique de RMSD en Figure 3-37, aboutit aux conformations des clusters 2.2 et 2.3, similaires à la référence. Avec le modèle TIP4P-Ew, on observe au cours de la GaMD n°1, des clusters montrant une forte proximité avec la structure expérimentale, soit les clusters 1.1 et 1.2 de RMSD  $< 5 \text{ \AA}$  et le cluster 3 avec une RMSD = 5,3 Å. Dans la GaMD n°2, la simulation semble très instable, bien qu'une conformation avec une RMSD de l'expérimentale de 10,77 Å parvienne à se stabiliser entre 375 et 630 ns représentée dans le cluster 2.3.

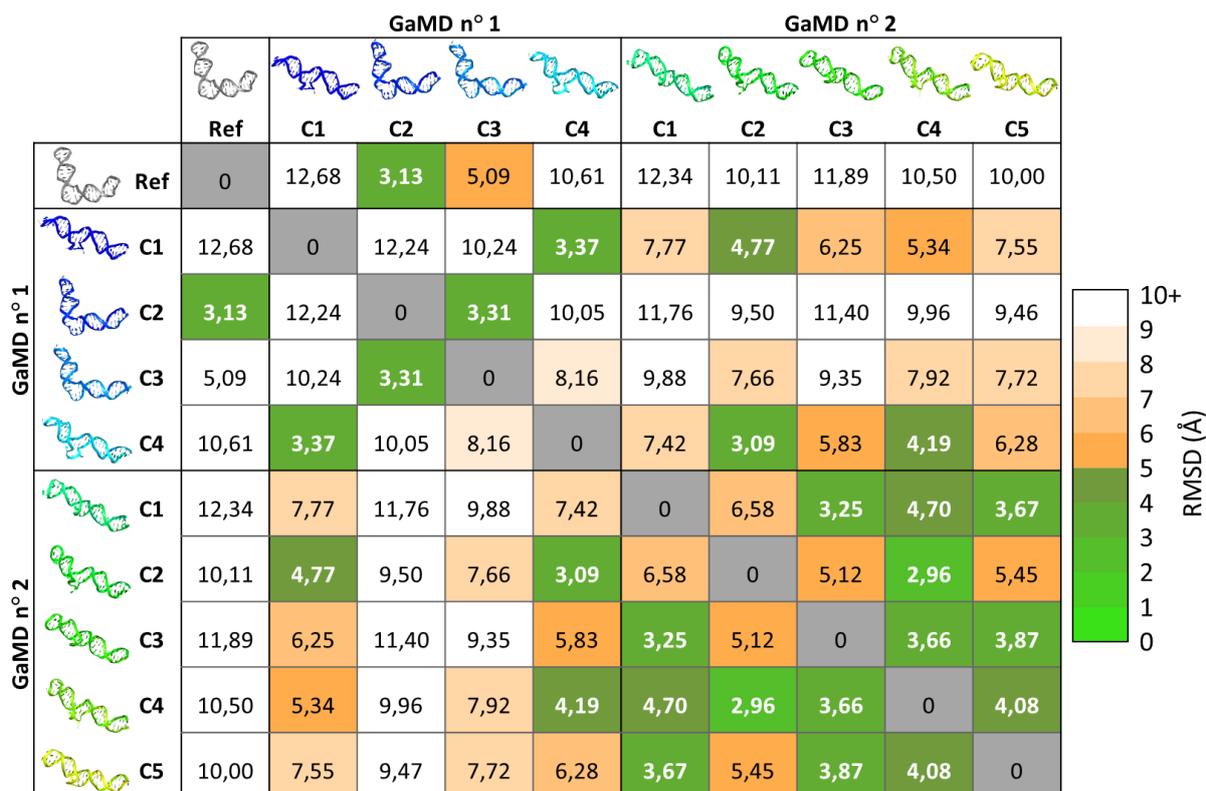


Figure 3-36. Mesure de la RMSD entre clusters suggéré par TTclust de 3THW, modélisé avec RNAdenovo en utilisant la structure secondaire expérimentale, simulé en GaMD incluant une concentration de 0,1 M de NaCl avec le modèle d'eau TIP3P. Les clusters obtenus avec TTclust sont numérotés par prédominance dans les GaMD concernée (Annexe 17). La structure expérimentale désignée 'Ref' est également incluses à titre comparatif.

Cette conformation est présente également dans les simulations en TIP3P mais n'est pas retrouvée parmi les conformations explorées en simulation sur structure expérimentale TIP3P et TIP4P-Ew.

De manière analogue à 3HXO, la diversité de conformation explorée durant les simulations en GaMD sur 3THW peut-être commune entre les simulations faites à partir de la structure résolue expérimentalement et celles à partir du modèle de RNAdenovo (Figure 3-38). Notamment, les clusters 1.2 et 1.3 extraits des simulations sur RNAdenovo sont similaires au clusters 1.3 ou 2.1 extraits des simulations depuis la structure expérimentale. Cependant, ces 4 clusters sont représentatifs de conformations similaires à la structure résolue expérimentalement. L'espace conformationnel exploré est donc différent entre ces simulations bien que toutes soient parvenues à retrouver la conformation expérimentale.

Modèle RNAdenovo 3THW structure secondaire expérimentale  
0,1M NaCl

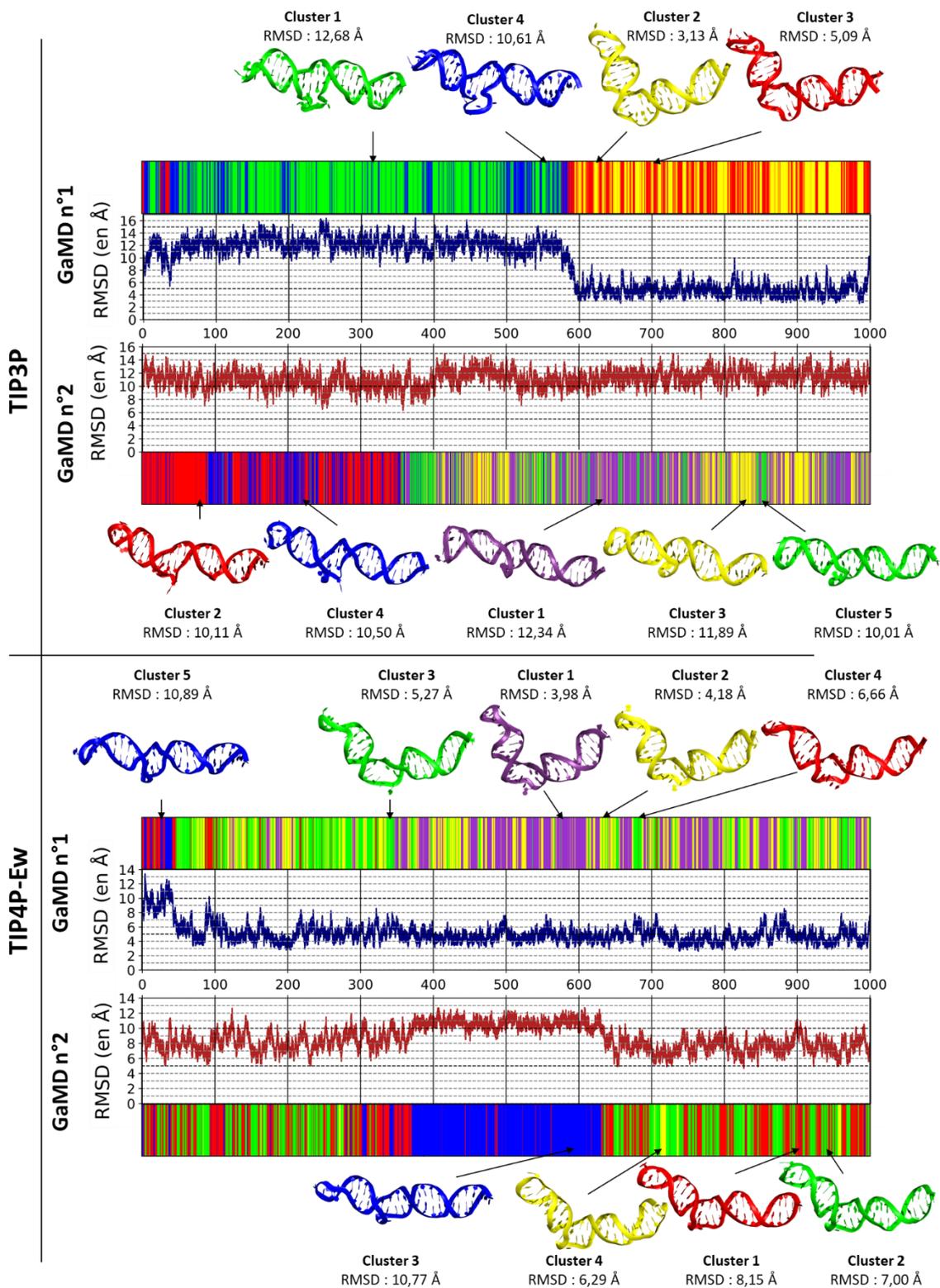


Figure 3-37. Courbe de RMSD et distribution des clusters et des conformations associées au cours des GaMD simulés dans un système concentré à 0,1 M en NaCl avec différents modèles d'eau (TIP3P ou TIP4P-Ew) appliqués sur le modèle RNAdenovo de 3THW en utilisant la structure secondaire expérimentale. Chaque couleur représente un cluster différent. Le numéro du cluster est assigné selon son pourcentage d'occurrence dans la simulation.

Pour cela et au vu des différences obtenues entre les deux simulations GaMD indépendantes, il faut questionner la convergence de ce type de simulation avec les temps de simulation et les biais choisis.

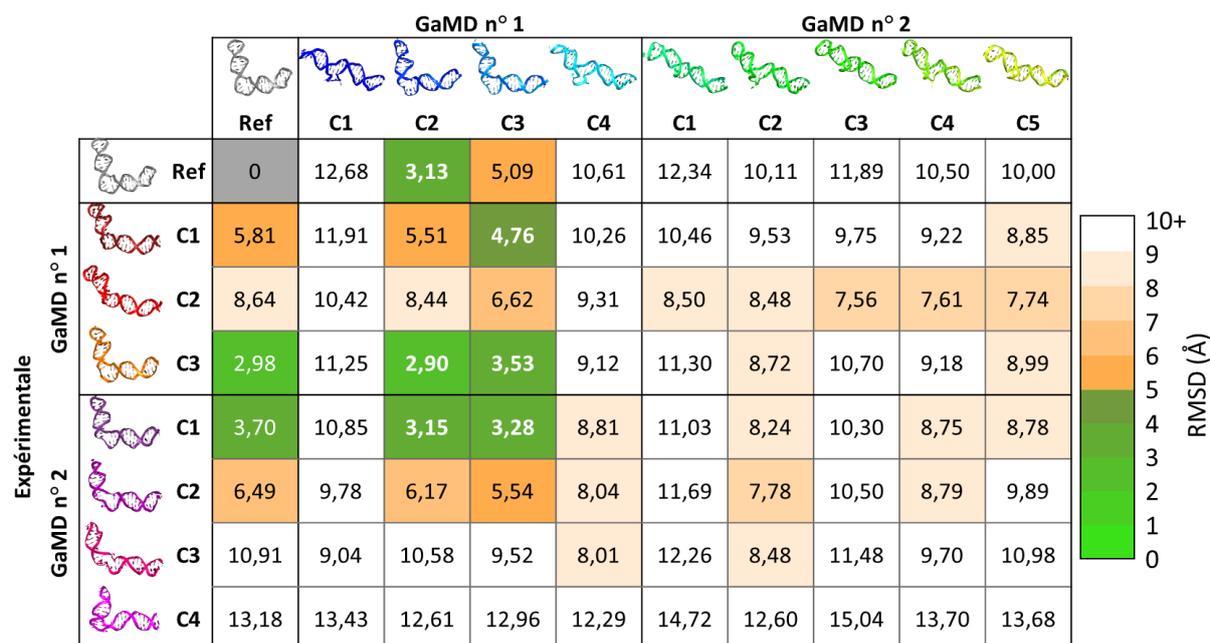


Figure 3-38. Mesure de la RMSD entre clusters suggéré par TTclust de 3THW expérimentale et 3THW modélisé avec RNAdenovo en utilisant la structure secondaire expérimentale simulés en GaMD incluant une concentration de 0,1 M de NaCl en TIP3P. Les clusters obtenus avec TTclust sont numérotés par prédominance dans les GaMD concernée. La structure expérimentale désignée 'Ref' est également incluse à titre comparatif.

#### 4.4.5 5HTO

5HTO représente le complexe entre un aptamère d'ADN et la LDH de *Plasmodium vivax* obtenu par cristallographie à rayons X (Choi & Ban, 2016). L'interaction entre la protéine et l'aptamère fait intervenir une série de nucléotides non-appariés dont la position dans l'espace est médiée par la présence d'un repliement type pseudonœud. Les simulations ont été donc effectuées en partant de i) la structure expérimentale, ii) le modèle de RNAdenovo obtenu à partir de la structure secondaire expérimentale et ayant une RMSD de 5,0 Å par rapport à celle-ci, et iii) le modèle de RNAdenovo obtenu à partir de la structure secondaire prédite par mfold et ayant une RMSD face à l'expérimentale de 7,8 Å. Mfold n'étant pas capable de prédire les interactions à longue distance, le pseudonœud caractéristique de 5HTO n'est pas présent dans ce deuxième modèle. Il est intéressant de vérifier si l'échantillonnage des dynamiques moléculaires permet de récupérer les interactions à longue distance.

#### 4.4.5.1 Simulations effectuées à partir de la structure expérimentale

La structure expérimentale de 5HTO a été simulée en aMD ou GaMD dans les différentes conditions de concentration en sel NaCl choisies (0 ou 0,1 M) ainsi qu'avec les deux modèles d'eau, TIP3P et TIP4P-Ew. Toutes les simulations ont montré la même stabilité, car aucun cluster identifié par TTclust ne dépasse les 3,5 Å de RMSD par rapport à la structure obtenue expérimentalement. Ces dynamiques n'incluent pas l'ion  $Mg^{2+}$  utilisé durant la cristallisation pour maintenir la formation du pseudonœud. Néanmoins, ces simulations semblent indiquer la stabilité de la conformation expérimentale sur des simulations de 1  $\mu s$ . Il est à déplorer qu'aucune autre conformation n'ait été identifiée, questionnant la capacité exploratoire de l'espace conformationnel de l'oligonucléotide de ces simulations quand la structure de départ est très stable. Néanmoins, l'absence de structures expérimentales pour cet oligonucléotide avec des autres conformations ne permet pas de mieux répondre à cette question.

#### 4.4.5.2 Simulations effectuées à partir du modèle RNAdenovo généré depuis la structure secondaire expérimentale

Les simulations aMD effectuées sur le modèle RNAdenovo de 5HTO obtenu à partir de la structure secondaire expérimentale montrent des différences de comportement entre les conditions appliquées. Quatre comportements différents peuvent être identifiés et semblent dépendants des combinaisons de conditions (TIP3P/0M NaCl ; TIP3P/0,1M NaCl ; TIP4P-Ew/0,1M NaCl; TIP4P-Ew/0M NaCl) (Figure 3-39). Dans le cas des simulations effectuées sur le système seulement neutralisé avec les ions  $Na^+$ , les dynamiques appliquant le modèle TIP3P favorisent l'apparition de conformations différentes et légèrement éloignées de la référence et ne parviennent pas à retrouver une conformation majoritaire proche de la référence. Seules des conformations très peu échantillonnées semblent s'en rapprocher.

		aMD n°1						aMD n°2					
Cluster		n°1	n°2	n°3	n°4	n°5	n°6	n°1	n°2	n°3	n°4	n°5	
NaCl 0M	TIP3P	 7,33 Å <sup>a</sup> 63%	 10,93 Å <sup>a</sup> 31%	 4,51 Å <sup>a</sup> 6%				 6,27 Å <sup>a</sup> 43%	 7,15 Å <sup>a</sup> 34%	 9,35 Å <sup>a</sup> 13%	 7,53 Å <sup>a</sup> 10%		
	TIP4P-EW	 5,29 Å <sup>a</sup> 49%	 4,14 Å <sup>a</sup> 20%	 5,16 Å <sup>a</sup> 17%				 3,39 Å <sup>a</sup> 14%	 4,27 Å <sup>a</sup> 45%	 3,66 Å <sup>a</sup> 25%	 3,72 Å <sup>a</sup> 17%		 3,18 Å <sup>a</sup> 7%
NaCl 0,1 M	TIP3P	 9,57 Å <sup>a</sup> 34%	 7,83 Å <sup>a</sup> 26%	 8,88 Å <sup>a</sup> 20%	 8,02 Å <sup>a</sup> 20%				 8,34 Å <sup>a</sup> 43%	 9,69 Å <sup>a</sup> 26%	 9,29 Å <sup>a</sup> 20%	 5,22 Å <sup>a</sup> 11%	
	TIP4P-EW	 13,91 Å <sup>a</sup> 32%	 6,28 Å <sup>a</sup> 20%	 14,98 Å <sup>a</sup> 18%	 13,83 Å <sup>a</sup> 11%				 14,73 Å <sup>a</sup> 11%	 17,93 Å <sup>a</sup> 8%	 14,07 Å <sup>a</sup> 33%	 14,01 Å <sup>a</sup> 25%	

Figure 3-39. Alignement des clusters des différentes simulations aMD effectuées en partant du modèle RNAdenovo de 5HTO obtenu à partir de la structure secondaire expérimentale en système neutralisé avec et sans concentration de 0,1 M de NaCl testé avec les modèles d'eau TIP3P ou TIP4P-Ew. <sup>a</sup> La RMSD obtenue face à la structure expérimentale est précisée pour chaque alignement ainsi que le pourcentage d'occupation dans la simulation.

En passant au modèle TIP4P-Ew, très peu de changements conformationnels sont observés et la structure se maintient dans une conformation proche de la structure de référence (Figure 3-39). Les clusters majoritaires sont représentatifs de ~40 à 50 % de leur simulation en aMD respective et sont très similaires à la référence avec des RMSD proches de 5 Å (aMD n°1) ou inférieures (aMD n°2) (Figure 3-39). Cette proximité est au détriment de l'exploration d'autres conformations. Ce n'est pas surprenant car le modèle d'eau ralentit les changements conformationnels du soluté.

Le système incluant une concentration de 0,1 M de NaCl et avec le modèle d'eau TIP3P permet d'explorer des conformations différentes de la référence mais ne présente aucune conformation stable similaire à la référence (Annexe 15). Au regard de l'évolution de la RMSD par rapport à la structure expérimentale en fonction du temps, il semblerait que la structure, au départ assez proche de celle de référence, s'en éloigne très rapidement.

Enfin, le passage au modèle d'eau TIP4P-Ew dans le système avec une concentration de NaCl de 0,1 M désorganise fortement la structure tertiaire et les clusters majoritaire en Figure 3-39 montrent des RMSD dépassant les 10 Å par rapport à la référence (Annexe 16) et des repliements très différents de ceux attendus.

Les simulations en GaMD à partir du modèle RNAdenovo généré sur la base de la structure secondaire expérimentale ont été effectuées exclusivement sur le système avec une concentration de 0,1 M de NaCl, et ont permis d'obtenir des résultats intéressants avec l'observation de conformations proches de la structure expérimentale, ainsi que d'autres conformations non explorées durant les simulations faites à partir de la structure expérimentale, mais qui pour la plupart dérivent probablement du fait que la structure initiale était éloignée de la structure de référence.

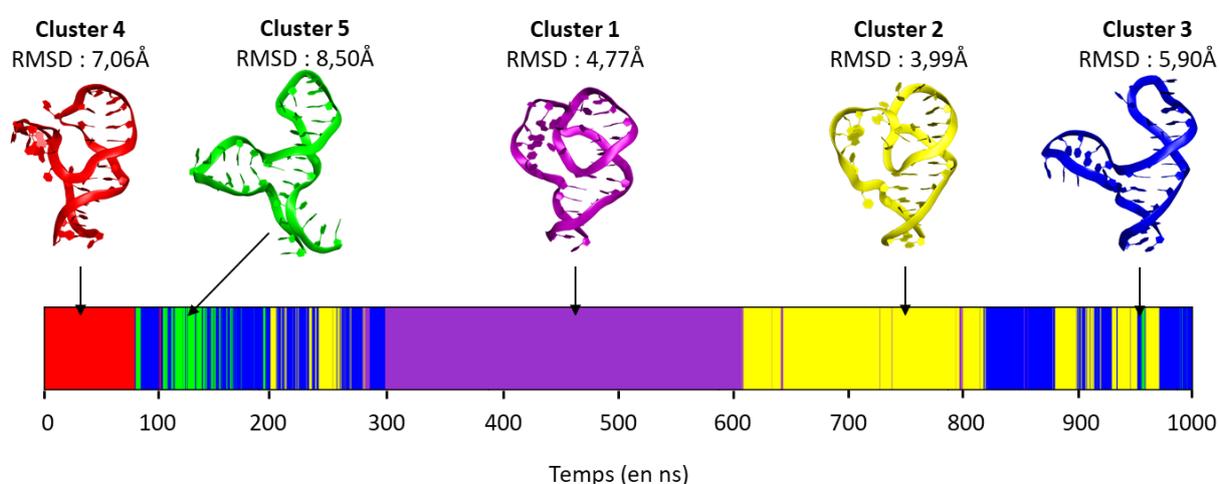


Figure 3-40. Distribution des clusters et des conformations associées au cours de la GaMD n°1 simulée dans un système concentré à 0,1 M en NaCl avec le modèle d'eau TIP3P sur le modèle RNAdenovo de 5HTO utilisant la structure secondaire expérimentale. Chaque couleur représente un cluster différent. Le numéro du cluster est assigné selon son pourcentage d'occurrence dans la simulation.

Au contraire des simulations aMD avec une concentration de NaCl de 0,1 M et le modèle de solvant TIP3P, la première GaMD faite dans les mêmes conditions montre un cluster majoritaire (1.1, ~32 % de la simulation) ayant une structure représentative avec une RMSD par rapport à l'expérimental de 4,77 Å (Figure 3-40) et qui reste stable pendant environ 300 ns. Le second cluster majoritaire (1.2, ~30 % de la simulation) présente une structure représentative avec une RMSD encore plus faible, égale à 3,99 Å, témoignant de la proximité de ces configurations avec les observations expérimentales et la stabilité au cours de la

dynamique. Une autre conformation, représentative de 25 % de la dynamique n°1 (cluster 1.3), montre une RMSD de 5,9 Å face à la référence, suggérant un changement structural important visible en Figure 3-40 et qui consiste en la perte des interactions caractérisant le pseudonœud. En calculant la RMSD entre les structures représentatives du cluster 1.3 et des clusters 1.2 et 1.5, les valeurs (4,19 Å et 5,49 Å, respectivement) indiquent des structures relativement proches. Cela semble indiquer que le cluster 1.3 est un état intermédiaire vers d'autres conformations plus éloignées. Ces observations sont appuyées avec la Figure 3-40 qui présente la répartition des clusters de conformations au cours de la GaMD n°1. Ainsi, entre 100 et 300 ns, de fréquents changements d'appartenance des conformations sont observés entre le cluster 1.5, le cluster 1.3 et le cluster 1.2. De la même façon, le cluster 1.5 semble être un état transitoire entre le cluster 1.4 stable en début de dynamique, et le cluster 1.3. On remarque également que les conformations du cluster 1.3 se retrouvent en fin de dynamique, indiquant la possibilité d'un autre changement de conformation et donc que la simulation n'a pas convergé.

Avec le modèle d'eau TIP4P-Ew, les clusters représentatifs des simulations de GaMD identifient plus de conformations similaires à la structure expérimentale, avec l'intégralité des clusters de GaMD n°2 en dessous de 5 Å de RMSD par rapport à la référence indiquant un certain maintien des repliements déjà présents au départ de la simulation (Annexe 21). Dans la GaMD n°1, le cluster le plus proche de la référence représente 20 % de la dynamique et est majoritairement retrouvé entre 0 et 120 ns de la dynamique (Figure 3-41)

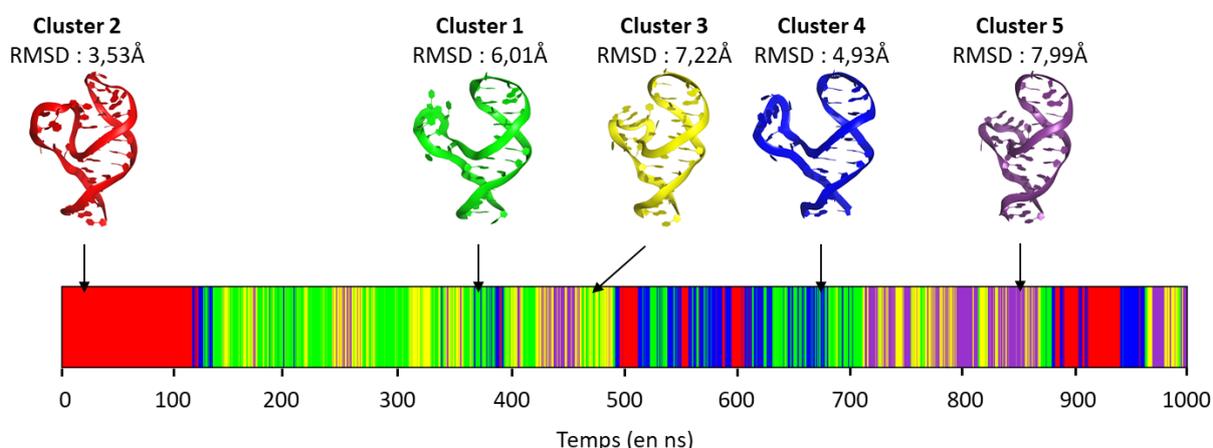


Figure 3-41. Distribution des clusters et des conformations associées au cours de la GaMD n°1 simulée dans un système concentré à 0,1 M en NaCl avec le modèle d'eau TIP4P-Ew sur le modèle RNAdenovo de 5HTO utilisant la structure secondaire expérimentale. Chaque couleur représente un cluster différent. Le numéro du cluster est assigné selon son pourcentage d'occurrence dans la simulation.

#### 4.4.5.3 Simulations effectuées à partir du modèle RNAdenovo généré depuis la structure secondaire prédite par mfold

Enfin, les simulations effectuées à partir du modèle de RNAdenovo obtenu depuis la structure secondaire prédite par mfold sont une démonstration des potentialités de la dynamique moléculaire à retrouver la structure expérimentale d'un oligonucléotide caractérisé par des interactions à longue distance qui ne sont pas indiquées dans la séquence dot-bracket générée par mfold. En effet, pour 5HTO, le modèle de départ montre quelques variations de structure en raison de l'absence de prédiction des interactions longues distance par mfold, impliquant les paires T11-A21 et T12-A20. Les nucléotides T6-T14 sont ainsi positionnés différemment par rapport à la structure résolue expérimentalement.

Les simulations faites avec le modèle d'eau TIP3P montrent 2 comportements différents en fonction de la réplique indépendante. TTclust assigne aux différentes conformations de la GaMD n°1 uniquement 2 clusters. Le premier est représentatif des conformations de 275 ns jusqu'à 900 ns (63 % de la simulation) et il présente une structure représentative avec une RMSD de 4,21 Å face à la structure expérimentale, contrairement au second cluster qui représente majoritairement le début de la dynamique (37 % de la simulation) et diffère de la structure résolue expérimentalement (RMSD = 5,4 Å), car les interactions à longue distance du pseudonœud manquent. Il existe donc un changement conformationnel entre la conformation de départ, le cluster 1.1 et le cluster 1.2 et la dynamique favorise l'apparition d'une conformation proche de l'expérimental, permettant la formation des interactions à longue distance, comme recherché. En revanche, en GaMD n°2, le comportement est très différent et les conformations explorées sont toutes distantes de la structure résolue expérimentalement (Annexe 20). Cela, comme déjà observé par des autres systèmes, montre la nécessité d'effectuer plusieurs répliques indépendantes de la simulation et de vérifier la longueur et le biais des simulations.

Le changement de modèle de solvant en TIP4P-Ew ne permet pas d'atteindre le repliement attendu et la structure semble alterner entre une conformation avec les nucléotides T6-T14 proches de la boucle G15-C29, et une avec ces deux régions éloignées (Figure 3-42). Toutes ces conformations sont ainsi très éloignées de la structure résolue expérimentalement (Annexe 21). Ainsi, le réplica n'a pas validé le comportement de la GaMD n°1 et suggère un manque d'échantillonnage.

## Modèle RNAdenovo 5HTO structure secondaire mfold 0,1M NaCl

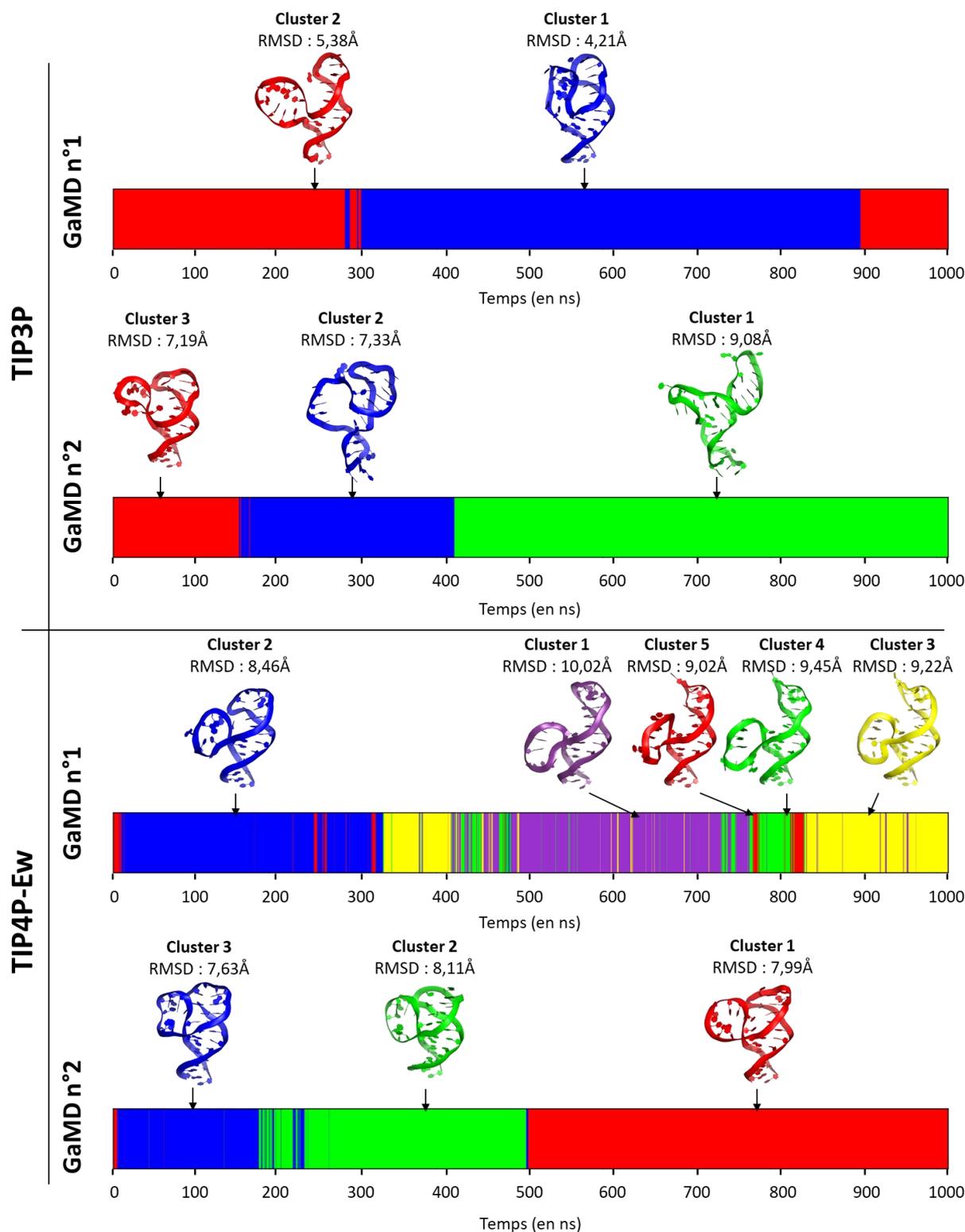


Figure 3-42. Distribution des clusters et des conformations associées au cours des GaMD simulés dans un système concentré à 0,1 M en NaCl avec différents modèles d'eaux (TIP3P ou TIP4P-Ew) appliqués sur le modèle RNAdenovo de 5HTO utilisant la structure secondaire de mfold. Chaque couleur représente un cluster différent. Le numéro du cluster est assigné selon son pourcentage d'occurrence dans la simulation.

## 4.5. Discussion

Le but premier de la dynamique moléculaire dans ces travaux est de mettre en lumière la capacité à retrouver les conformations obtenues expérimentalement lorsque celle-ci n'a pas pu être déterminée avec les approches de prédiction 3D ce qui permet de démontrer le potentiel de la dynamique moléculaire dans la modélisation des oligonucléotides. Au regard des résultats obtenus, ce dernier a été globalement vérifié. En effet, dans la plupart des cas il a été possible d'échantillonner de manière significative des conformations proches de l'expérimentale, même si la structure de départ en était très éloignée. Un exemple très intéressant est celui représenté par 5HTO, qui est caractérisé par la présence d'un pseudonœud que *mfold* n'est pas capable de prédire. Sous certaines conditions, avec les dynamiques moléculaires à échantillonnage intensif, il a été possible d'obtenir une conformation proche de l'expérimentale et d'observer la formation des interactions à longue distance caractéristique des pseudonœuds.

Les résultats prometteurs dispensés par ce protocole permettent d'envisager cette approche de modélisation tridimensionnelles des oligonucléotides comme support aux méthodes standards d'obtention de structures (Rayons X, RMN, cryo-EM) et pallier le manque de structures disponibles sur les bases de données. Néanmoins, il est bon d'envisager quelques optimisations du protocole afin de i) maximiser les similitudes entre structure expérimentale et conformations explorées ii) couvrir l'ensemble de l'espace conformationnel durant les dynamiques moléculaires.

Le point de départ de la dynamique moléculaire est déterminant pour les simulations et influe sur les conformations explorées. De trop fortes divergences peuvent rendre inatteignables les conformations accessibles expérimentalement sur des simulations de l'ordre de la nanoseconde, comme avec 3HXO dont le modèle RNAdenovo à partir de la structure secondaire obtenue avec *mfold* présente trop de variations par rapport à la structure attendue. Au contraire, les structures comme 1EZN et 5HTO dont les structures secondaires varient mais demeurent sous le seuil de  $Apta_D \leq 5$  ont permis de retrouver des conformations proches de l'expérimental. La structure secondaire de départ se doit alors d'être la plus proche possible de celle attendue pour s'attendre à une certaine représentativité de la dynamique moléculaire en termes de conformations explorées, comme l'ensemble des simulations

effectuées sur les structures obtenues avec RNAAdenovo utilisant la structure secondaire de référence.

En outre, au regard des observations sur les différents systèmes, il semble que certaines conditions de simulation affectent le comportement des systèmes en dynamique moléculaire.

L'effet du modèle d'eau est clairement identifiable à partir de l'analyse des simulations ayant comme coordonnées initiales celles de la structure expérimentale. Par exemple, la structure expérimentale de 3HXO simulé avec le modèle d'eau TIP3P permet d'observer plusieurs conformations alternatives, que ce soit en aMD ou GaMD indépendamment de la concentration en NaCl. Les simulations en TIP4P-Ew offrent des dynamiques beaucoup moins favorables à l'exploration de conformations alternatives, car l'ensemble des clusters observés montrent une RMSD avec la structure résolue expérimentalement inférieur à 5,5 Å (dont 2 clusters légèrement supérieur à 5 Å). Ce constat est également possible sur les simulations GaMD de 1EZN et 5HTO avec le modèle TIP4P-Ew effectuées à partir des modèles RNAAdenovo générés depuis la structure secondaire prédite par mfold. Sur ces simulations aucun cluster proche de l'expérimentale n'a pu être identifié malgré les deux répliques indépendantes de 1 µs, suggérant un manque d'échantillonnage sous ces conditions qui empêche d'observer des changements conformationnels menant à la conformation souhaitée. En revanche, les simulations GaMD avec le modèle TIP3P sont parvenues à trouver plusieurs conformations proches de l'expérimental dans les 2 répliques de 1 µs. Cela peut être dû à la nature moins fluide du modèle TIP4P-Ew par rapport au modèle TIP3P et la différente distribution des charges ponctuelles qui peut impacter l'échantillonnage d'un système très chargé, comme celui oligonucléotidique. Les déviations de ce comportement sont probablement imputables à un manque de convergence de simulations. Par exemple, les simulations en aMD faites sur le modèle RNAAdenovo de 5HTO obtenue à partir de la structure secondaire expérimentale avec le modèle d'eau TIP3P n'aboutissent à l'exploration d'aucune conformation proche de 5HTO expérimental. Avec le modèle TIP4P-Ew, les simulations aMD ont amené la structure à des repliements inattendus. Ces deux problèmes n'ont cependant pas été observés en GaMD.

L'effet d'une concentration 0,1 M de NaCl est moins clair et évident. Par exemple, pour 3HXO, le modèle TIP4P-Ew en présence de NaCl donne des simulations pendant lesquelles plus aucune conformation proche de la structure expérimentale n'est explorée, contrairement aux

systèmes sans NaCl avec le modèle d'eau TIP4P-Ew. Il en va de même pour 5HTO, avec en plus une totale variation de la structure tertiaire comme montré en Figure 3-39. En revanche, pour les simulations effectuées sur 3THW à partir du modèle de RNAdenovo, les résultats soulignent une absence totale de conformations proches de la structure obtenue expérimentalement pour les dynamiques sans NaCl, et seulement quelques conformations proches pour les simulations avec 0,1 M de NaCl en TIP3P (cluster 2.3) ou TIP4P-Ew (cluster 1.4). L'effet variable de la présence d'une certaine concentration de sel est de quelque façon attendu : en effet, on ne connaît pas l'effet de la présence de 0,1 M de NaCl sur la structure des oligonucléotides qui ont été résolus par cristallographie aux rayons X et, pour les deux autres, les conditions utilisées pour les expériences de RMN ne sont pas forcément les mêmes que celles des simulations, qui ont été choisies pour pouvoir comparer plus facilement les résultats des différents systèmes simulés.



## Conclusion Générale

Les oligonucléotides à simple brin jouent un rôle très important dans le fonctionnement cellulaire et ils ont aussi un fort potentiel biotechnologique. Leur fonction dépend fortement de leur structure tridimensionnelle. Ainsi, pouvoir prédire leur repliement peut être un point clé pour la compréhension de leur rôle et pour concevoir des oligonucléotides à visée diagnostique et thérapeutique. L'inconvénient des approches *in silico* pour la modélisation réside dans leur dépendance vis-à-vis des connaissances actuelles. Ce domaine reste encore limité à cause de l'intérêt plutôt récent apporté aux acides nucléiques pour des applications alternatives au transport de l'information génétique.

Dans le contexte de la prédiction de la structure des oligonucléotides, la flexibilité de ce type de molécule occupe une place importante et rend difficile la possibilité de prédire le repliement tridimensionnel avec exactitude. La structure secondaire, malgré quelques variations possibles au sein d'un même oligonucléotide, compile une partie des informations de repliement essentielles pour aider à la prédiction de la structure tridimensionnelle. La plupart des méthodes de prédiction de structures tridimensionnelles incluent une étape visant l'exploration des conformations. RNAdenovo de Rosetta rend possible la génération de multiples prédictions d'énergie minimisée qui font office de conformations ou états transitoires de la mobilité des oligonucléotides. De la même façon, SimRNA utilise l'approche Monte Carlo pour fournir différentes conformations au travers d'un échantillonnage aléatoire, au même titre que la dynamique moléculaire permet de reproduire le comportement dynamique des oligonucléotides en solution au travers des lois de la mécanique Newtoniennes. Néanmoins, dans la Partie 3.3 de ce manuscrit, il a été montré que cela n'est pas suffisant pour presque la moitié des oligonucléotides étudiés.

Une autre approche possible pour l'obtention de la structure tridimensionnelle des oligonucléotides est la dynamique moléculaire, mais, sans aucune information préliminaire sur le possible repliement, cette méthode nécessite de longs temps de simulation pour procéder au repliement des oligonucléotides.

Par conséquent, l'utilisation combinée des outils de prédiction tridimensionnelles et des approches de dynamiques moléculaires peuvent contrebalancer leurs défauts. Ensemble, ces

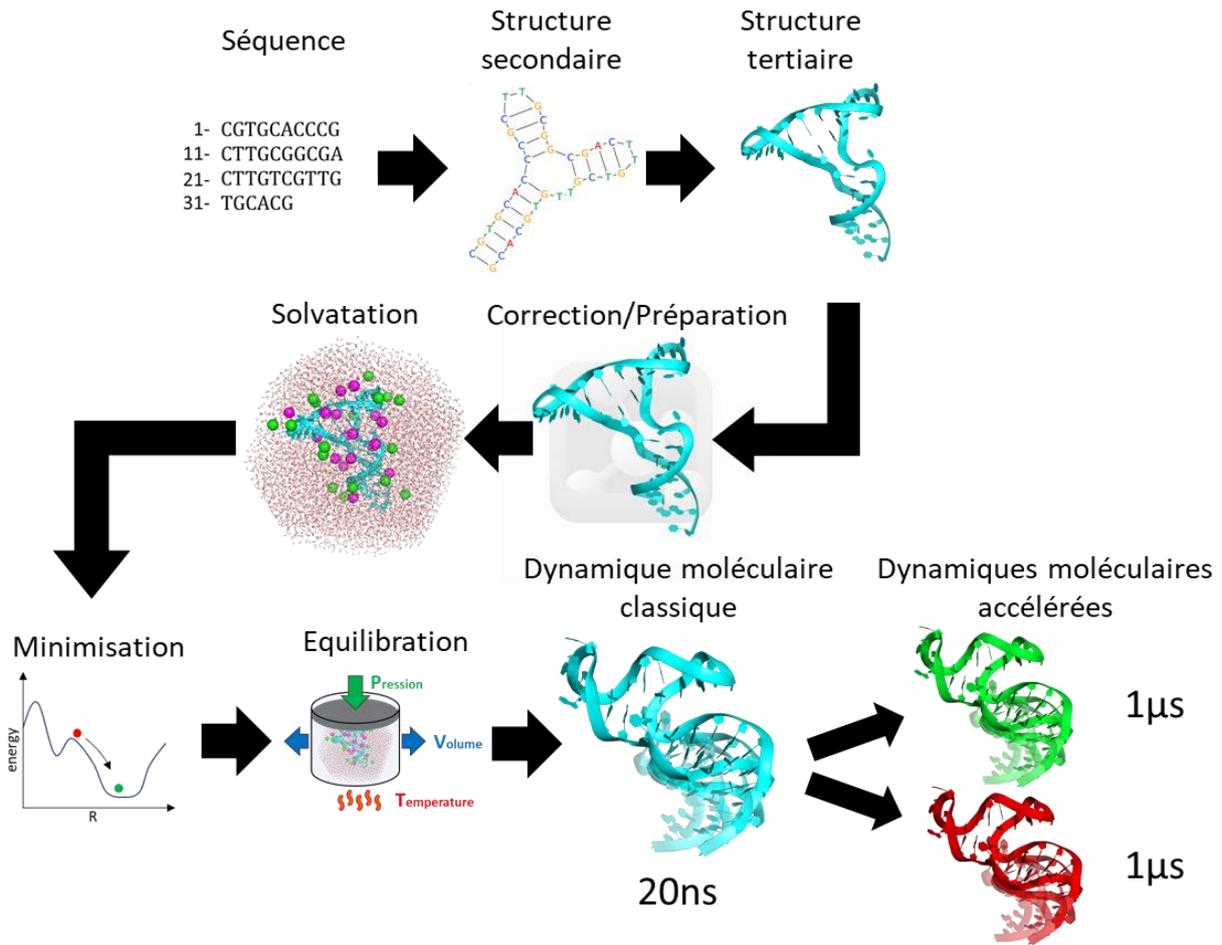


Figure C-0-1. Protocole complet de prédiction de structure des oligonucléotides.

approches ont présenté des résultats prometteurs dans l'échantillonnage des conformations des oligonucléotides et il a été possible pour l'ensemble des structures testées de retrouver les repliements attendus en ne partant que de la séquence et de la prédiction de la structure secondaire.

Pour obtenir cette dernière, le benchmark des outils de prédiction 2D a permis d'identifier les outils les plus intéressants pour des prédictions fiables bien que quelques améliorations puissent être apportées pour pallier ce problème, notamment l'ajout de la prédiction des G-quadruplexes. Il en résulte une certaine prédominance des approches de minimisation d'énergie libre dans la prédiction de structure secondaire fiable, bien que les approches modernes intégrant le *deep-learning* présentent des performances prometteuses. Au cours de ce travail, le développement et l'utilisation d'AptaMat pour la comparaison des structures a été nécessaire au regard des limites des métriques de comparaison des structures secondaires couramment utilisées. Son application pour la classification en familles

fonctionnelles des oligonucléotides issus de la base de données Rfam a su ultérieurement démontrer l'efficacité de ce nouvel outil. Son potentiel de comparaison peut être exploité dans l'étude des relations entre structure, fonction et évolution.

Les approches de prédictions réunies dans un protocole de modélisation *de novo* (Figure C-0-1) ont su montrer leur potentiel à retrouver les informations structurales proches des observations expérimentales. Malgré un besoin d'optimisations, le protocole s'est montré capable de retrouver les repliements tridimensionnels d'une séquence d'ADN à simple brin. Cette caractéristique a le potentiel d'être appliqué dans la recherche de molécules actives contre une cible sélectionnée car il peut épauler les approches *in vitro* limitées par l'incertitude dans l'obtention de résultats malgré les ressources déployées, en plus de combler les manques dans les bases de données de structure.

Pour améliorer ces résultats, il serait envisageable d'optimiser, sur plusieurs aspects, le protocole. Tout d'abord, pour faciliter l'échantillonnage des conformations expérimentales, les modèles issus de SimRNA pourraient être utilisés comme point de départ, car cet outil est légèrement plus efficace que RNAde novo. En alternative, plusieurs conformations produites par RNAde novo pourraient être utilisées comme point de départ des simulations.

Pour ce qui concerne les simulations, l'idéal est de se focaliser sur l'optimisation du protocole GaMD, avec comme objectif d'obtenir non seulement les conformations majoritaires mais aussi le profil énergétique non biaisé associé. Cela permettrait de comprendre comment le changement conformationnel a lieu. Il faudrait donc vérifier la longueur et le biais introduit dans les simulations et, potentiellement, rajouter une troisième réplique. Pour ce qui concerne l'environnement de simulation, un choix cas-par-cas des sels à rajouter et de leur concentration est envisageable. Malgré ces optimisations à faire, les résultats sont encourageants.

Enfin, ce protocole à long terme peut servir pour le développement d'une procédure de conception *de novo* d'aptamères ciblant des protéines d'intérêt. Il est ainsi possible de continuer ces études en utilisant des approches de *docking* pour l'identification d'un site d'interaction avec une cible, et par des approches avancées de dynamique moléculaire pour confirmer la stabilité d'un potentiel complexe.



## Références

- Aartsma-Rus, A., Straub, V., Hemmings, R., Haas, M., Schlosser-Weber, G., Stoyanova-Beninska, V., Mercuri, E., Muntoni, F., Sepodes, B., Vroom, E., & Balabanov, P. (2017). Development of Exon Skipping Therapies for Duchenne Muscular Dystrophy: A Critical Review and a Perspective on the Outstanding Issues. *Nucleic Acid Therapeutics*, *27*(5), 251-259.
- Adcock, S. A., & McCammon, J. A. (2006). Molecular dynamics: Survey of methods for simulating the activity of proteins. *Chemical Reviews*, *106*(5), 1589-1615.
- Afanasyeva, A., Nagao, C., & Mizuguchi, K. (2019). Prediction of the secondary structure of short DNA aptamers. *Biophysics and physcobiology*, *16*(0), 287-294.
- Agius, P., Bennett, K. P., & Zuker, M. (2010). Comparing RNA secondary structures using a relaxed base-pair score. *Rna*, *16*(5), 865-878.
- Akiyama, M., Sato, K., & Sakakibara, Y. (2018). A max-margin training of RNA secondary structure prediction integrated with the thermodynamic model. *Journal of Bioinformatics and Computational Biology*, *16*(6), 1840025.
- Alonso, D., & Mondragón, A. (2021). Mechanisms of catalytic RNA molecules. *Biochemical Society Transactions*, *49*(4), 1529-1535.
- Amann, R., & Fuchs, B. M. (2008). Single-cell identification in microbial communities by improved fluorescence in situ hybridization techniques. *Nature Reviews Microbiology*, *6*(5), 339-348.
- Avalle-Bihan, B., Loussouarn, C., Isber, H., Friboulet, A., & Padiolleau, S. (2015). *Stat5 inhibitors and use of same (Word Wide patent No. WO 2015/140479 A1)*.
- Bartel, D. P. (2009). MicroRNAs: Target Recognition and Regulatory Functions. *Cell*, *136*(2), 215-233.
- Bartys, N., Kierzek, R., & Lisowiec-Wachnicka, J. (2019). The regulation properties of RNA secondary structure in alternative splicing. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, *1862*(11-12), 194401.
- Bell, D. R., Weber, J. K., Yin, W., Huynh, T., Duan, W., & Zhou, R. (2020). In silico design and validation of high-affinity RNA aptamers targeting epithelial cellular adhesion molecule dimers. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(15), 8486-8493.
- Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. D., & Zardecki, C. (2002). The protein data bank. *Acta Crystallographica Section D: Biological Crystallography*, *58*(6 l), 899-907.
- 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. *Biophysical Journal*, *63*(3), 751-759.

- Binet, T., Avalle, B., Dávila Felipe, M., & Maffucci, I. (2023). AptaMat: a matrix-based algorithm to compare single-stranded oligonucleotides secondary structures. *Bioinformatics*, *39*(1), 2022.05.04.490414.
- Boiziau, C., Kurfurst, R., Cazenave, C., Roig, V., Thuong, N. T., & Toulmé, J.-J. (1991). Inhibition of translation initiation by antisense oligonucleotides via an RNase-H independent mechanism. *Nucleic Acids Research*, *19*(5), 1113-1119.
- Boniecki, M. J., Lach, G., Dawson, W. K., Tomala, K., Lukasz, P., Soltysinski, T., Rother, K. M., & Bujnicki, J. M. (2015). SimRNA: A coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic Acids Research*, *44*(7), e63.
- Breaker, R. R. (2011). Prospects for Riboswitch Discovery and Analysis. *Molecular Cell*, *43*(6), 867-879.
- Buratti, E., & Baralle, F. E. (2004). Influence of RNA Secondary Structure on the Pre-mRNA Splicing Process. *Molecular and Cellular Biology*, *24*(24), 10505-10514.
- Burge, S., Parkinson, G. N., Hazel, P., Todd, A. K., & Neidle, S. (2006). Quadruplex DNA: Sequence, topology and structure. *Nucleic Acids Research*, *34*(19), 5402-5415.
- Cai, Z., Cao, C., Ji, L., Ye, R., Wang, D., Xia, C., Wang, S., Du, Z., Hu, N., Yu, X., Chen, J., Wang, L., Yang, X., He, S. & Xue Y. (2020) RIC-seq for global in situ profiling of RNA–RNA spatial interactions. *Nature* *582*, 432–437.
- Caliński, T., & Harabasz, J. (1974). A Dendrite Method For Cluster Analysis. *Communications in Statistics*, *3*(1), 1-27.
- Chełkowska-Pauszek, A., Kosiński, J. G., Marciniak, K., Wysocka, M., Bakowska-żywicka, K., & Żywicki, M. (2021). The role of rna secondary structure in regulation of gene expression in bacteria. *International Journal of Molecular Sciences*, *22*(15).
- Cheng, C. Y., Chou, F. C., & Das, R. (2015). Modeling complex RNA tertiary folds with Rosetta. *Methods in Enzymology*, *553*, 35-64.
- Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 6.
- Cho, E. J., Collett, J. R., Szafranska, A. E., & Ellington, A. D. (2006). Optimization of aptamer microarray technology for multiple protein targets. *Analytica Chimica Acta*, *564*(1), 82-90.
- Choi, S.-J., & Ban, C. (2016). Crystal structure of a DNA aptamer bound to PvLDH elucidates novel single-stranded DNA structural elements for folding and recognition. *Scientific Reports*, *6*(1), 34998.
- Chowdhury, S., Maris, C., Allain, F. H. T., & Narberhaus, F. (2006). Molecular basis for temperature sensing by an RNA thermometer. *EMBO Journal*, *25*(11), 2487-2497.
- Chung, N. C., Miasojedow, B. Z., Startek, M., & Gambin, A. (2019). Jaccard/Tanimoto similarity test and estimation methods for biological presence-absence data. *BMC Bioinformatics*, *20*(S15), 644.
- Collie, G. W., & Parkinson, G. N. (2011). The application of DNA and RNA G-quadruplexes to therapeutic medicines. *Chemical Society Reviews*, *40*(12), 5867-5892.

- Crick, F. H. C. (1954). The Complementary Structure of Dna. *Proceedings of the National Academy of Sciences*, 40(8), 756-758.
- Crooke, S. T. (2017). Molecular Mechanisms of Antisense Oligonucleotides. *Nucleic Acid Therapeutics*, 27(2), 70-77.
- D.A. Case, K. Belfon, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V. W. D. C., T.A. Darden, R.E. Duke, G. Giambasu, M.K. Gilson, H. Gohlke, A.W. Goetz, R Harris, S. Izadi, S.A. Izmailov, K. Kasavajhala, A. Kovalenko, R. Krasny, T. Kur, P. A. K., D.A. Case, K. Belfon, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V. W. D. C., T.A. Darden, R.E. Duke, G. Giambasu, M.K. Gilson, H. Gohlke, A.W. Goetz, R Harris, S. Izadi, S. A. I., K. Kasavajhala, A. Kovalenko, R. Krasny, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. L., T. Luchko, R. Luo, V. Man, K.M. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, A. Onufriev, F., Pan, S. Pantano, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, N. R., Skrynnikov, J. Smith, J. Swails, R.C. Walker, J. Wang, L. Wilson, R.M. Wolf, X. Wu, Y. Xiong, Y. X., & D.M. York and P.A. Kollman (2020), AMBER 2020, University of California, S. F. (2020). Amber 2020. *Journal of Chemical Information and Modeling*, 53(9), 1689-1699.
- Danaee, P., Rouches, M., Wiley, M., Deng, D., Huang, L., & Hendrix, D. (2018). bpRNA: large-scale automated annotation and analysis of RNA secondary structure. *Nucleic Acids Research*, 46(11), 5381-5394.
- Darty, K., Denise, A., & Ponty, Y. (2009). VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, 25(15), 1974-1975.
- Das, R., & Baker, D. (2007). Automated de novo prediction of native-like RNA tertiary structures. *Proceedings of the National Academy of Sciences*, 104(37), 14664-14669.
- Davies, D. R., Gelinas, A. D., Zhang, C., Rohloff, J. C., Carter, J. D., O'Connell, D., Waugh, S. M., Wolk, S. K., Mayfield, W. S., Burgin, A. B., Edwards, T. E., Stewart, L. J., Gold, L., Janjic, N., & Jarvis, T. C. (2012). Unique motifs and hydrophobic interactions shape the binding of modified DNA ligands to protein targets. *Proceedings of the National Academy of Sciences*, 109(49), 19971-19976.
- Davlieva, M., Donarski, J., Wang, J., Shamo, Y., & Nikonowicz, E. P. (2014). Structure analysis of free and bound states of an RNA aptamer against ribosomal protein S8 from *Bacillus anthracis*. *Nucleic Acids Research*, 42(16), 10795-10808.
- De Beauchene, I. C., De Vries, S. J., & Zacharias, M. (2016). Fragment-based modelling of single stranded RNA bound to RNA recognition motif containing proteins. *Nucleic Acids Research*, 44(10), 4565-4580.
- Deigan, K. E., Li, T. W., Mathews, D. H., & Weeks, K. M. (2009). Accurate SHAPE-directed RNA structure determination. *Proceedings of the National Academy of Sciences*, 106(1), 97-102.
- Dias, N., & Stein, C. A. (2002). Antisense oligonucleotides: basic concepts and mechanisms. *Molecular cancer therapeutics*, 1(5), 347-355.
- Do, C. B., Woods, D. A., & Batzoglou, S. (2006). CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, 22(14), e90-e98.

- Dougan, H., Lyster, D. M., Vo, C. V, Stafford, A., Weitz, J. I., & Hobbs, J. B. (2000). Extending the lifetime of anticoagulant oligodeoxynucleotide aptamers in blood. *Nuclear Medicine and Biology*, 27(3), 289-297.
- Draper, D. E., Grilley, D., & Soto, A. M. (2005). Ions and RNA folding. *Annual Review of Biophysics and Biomolecular Structure*, 34, 221-243.
- Eichhorn, C. D., & Al-Hashimi, H. M. (2014). Structural dynamics of a single-stranded RNA–helix junction using NMR. *RNA*, 20(6), 782-791.
- Elayadi, A. N., Braasch, D. A., & Corey, D. R. (2002). Implications of High-Affinity Hybridization by Locked Nucleic Acid Oligomers for Inhibition of Human Telomerase. *Biochemistry*, 41(31), 9973-9981.
- Felden, B. (2007). RNA structure: experimental analysis. *Current Opinion in Microbiology*, 10(3), 286-291.
- Floege, J., Ostendorf, T., Janssen, U., Burg, M., Radeke, H. H., Vargeese, C., Gill, S. C., Green, L. S., & Janjic, N. (1999). Novel Approach to Specific Growth Factor Inhibition in Vivo. *The American Journal of Pathology*, 154(1), 169-179.
- Flores, S. C., Wan, Y., Russell, R., & Altman, R. B. (2010). Predicting RNA structure by multiple template homology modeling. *Pacific Symposium on Biocomputing*, 216-227.
- Fohrer, J., Hennig, M., & Carlomagno, T. (2006). Influence of the 2'-Hydroxyl Group Conformation on the Stability of A-form Helices in RNA. *Journal of Molecular Biology*, 356(2), 280-287.
- Frank, J. (2002). Single-particle imaging of macromolecules by cryo-electron microscopy. *Annual Review of Biophysics and Biomolecular Structure*, 31(1), 303-319.
- Fu, L., Cao, Y., Wu, J., Peng, Q., Nie, Q., & Xie, X. (2022). Ufold: Fast and accurate RNA secondary structure prediction with deep learning. *Nucleic Acids Research*, 50(3), E14.
- Galindo-Murillo, R., Robertson, J. C., Zgarbová, M., Šponer, J., Otyepka, M., Jurečka, P., & Cheatham, T. E. (2016). Assessing the Current State of Amber Force Field Modifications for DNA. *Journal of Chemical Theory and Computation*, 12(8), 4114-4127.
- Ganser, L. R., Kelly, M. L., Herschlag, D., & Al-Hashimi, H. M. (2019). The roles of structural dynamics in the cellular functions of RNAs. *Nature Reviews Molecular Cell Biology*, 20(8), 474-489.
- Golas, M. M., Sander, B., Will, C. L., Lührmann, R., & Stark, H. (2005). Major Conformational Change in the Complex SF3b upon Integration into the Spliceosomal U11/U12 di-snRNP as Revealed by Electron Cryomicroscopy. *Molecular Cell*, 17(6), 869-883.
- Gong, M. (2023). *Developing a new tool to purify methylated peptides from bacteria in order to study bacterial mechanosensing*. Université de Technologie de Compiègne.
- Gorzelnik, K. V., & Zhang, J. (2021). Cryo-EM reveals infection steps of single-stranded RNA bacteriophages. *Progress in Biophysics and Molecular Biology*, 160, 76-83.
- Grasby, J. A., & Gait, M. J. (1994). Synthetic oligoribonucleotides carrying site-specific modifications for RNA structure-function analysis. *Biochimie*, 76(12), 1223-1234.

- Gruber, A. R., Lorenz, R., Bernhart, S. H., Neuböck, R., & Hofacker, I. L. (2008). The Vienna RNA websuite. *Nucleic Acids Research*, *36*, W70-W74.
- Guo, P., & Lam, S. L. (2016). Minidumbbell: A New Form of Native DNA Structure. *Journal of the American Chemical Society*, *138*(38), 12534-12540.
- Gupta, S., Gellert, M., & Yang, W. (2012). Mechanism of mismatch recognition revealed by human MutS $\beta$  bound to unpaired DNA loops. *Nature Structural and Molecular Biology*, *19*(1), 72-79.
- Gyi, J. I., Lane, A. N., Conn, G. L., & Brown, T. (1998). The orientation and dynamics of the C2'-OH and hydration of RNA and DNA RNA hybrids. *Nucleic Acids Research*, *26*(13), 3104-3110.
- Hajdin, C. E., Ding, F., Dokholyan, N. V., & Weeks, K. M. (2010). On the significance of an RNA tertiary structure prediction. *RNA*, *16*(7), 1340-1349.
- Hale, S. P., Poole, L. B., & Gerlt, J. A. (1993). Mechanism of the reaction catalyzed by staphylococcal nuclease: Identification of the rate-determining step. *Biochemistry*, *32*(29), 7479-7487.
- Haller, A., Soulière, M. F., & Micura, R. (2011). The dynamic nature of RNA as key to understanding riboswitch mechanisms. *Accounts of Chemical Research*, *44*(12), 1339-1348.
- Hamelberg, D., Mongan, J., & McCammon, J. A. (2004). Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *Journal of Chemical Physics*, *120*(24), 11919-11929.
- Havrila, M., Stadlbauer, P., Kührová, P., Banáš, P., Mergny, J. L., Otyepka, M., & Šponer, J. (2018). Structural dynamics of propeller loop: Towards folding of RNA G-quadruplex. *Nucleic Acids Research*, *46*(17), 8754-8771.
- Haynie, D. T. (2008). Biological thermodynamics: Second edition. In *Biological Thermodynamics: Second Edition*. Cambridge University Press.
- Healy, J. M., Lewis, S. D., Kurz, M., Boomer, R. M., Thompson, K. M., Wilson, C., & McCauley, T. G. (2004). Pharmacokinetics and Biodistribution of Novel Aptamer Compositions. *Pharmaceutical Research*, *21*(12), 2234-2246.
- Hendrix, C., Rosemeyer, H., Verheggen, I., Van Aerschot, A., Seela, F., & Herdewijn, P. (1997). 1', 5' -Anhydrohexitol Oligonucleotides: Synthesis, Base Pairing and Recognition by Regular Oligodeoxyribonucleotides and Oligoribonucleotides. *Chemistry - A European Journal*, *3*(1), 110-120.
- Hénon, M. (1971). The Monte Carlo method. *Astrophysics and Space Science*, *14*(1), 151-167.
- Herschlag, D., Allred, B. E., & Gowrishankar, S. (2015). From static to dynamic: The need for structural ensembles and a predictive model of RNA folding and function. *Current Opinion in Structural Biology*, *30*, 125-133.
- Hoetzel, J., & Suess, B. (2022). Structural Changes in Aptamers are Essential for Synthetic Riboswitch Engineering: Synthetic riboswitch engineering. *Journal of Molecular Biology*, *434*(18), 167631.

- Hofacker, I. L. (2003). Vienna RNA secondary structure server. *Nucleic Acids Research*, 31(13), 3429-3431.
- Hofacker, I. L. (2014). Energy-directed RNA structure prediction. *Methods in Molecular Biology*, 1097, 71-84.
- Holbrook, S. R., Cheong, C., Tinoco, I., & Kim, S. H. (1991). Crystal structure of an RNA double helix incorporating a track of non-Watson-Crick base pairs. *Nature*, 353(6344), 579-581.
- Horn, H. W., Swope, W. C., Pitera, J. W., Madura, J. D., Dick, T. J., Hura, G. L., & Head-Gordon, T. (2004). Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *Journal of Chemical Physics*, 120(20), 9665-9678.
- Hoshika, S., Minakawa, N., & Matsuda, A. (2004). Synthesis and physical and physiological properties of 4'-thioRNA: Application to post-modification of RNA aptamer toward NF- $\kappa$ B. *Nucleic Acids Research*, 32(13), 3815-3825.
- Houseley, J., & Tollervey, D. (2009). The Many Pathways of RNA Degradation. *Cell*, 136(4), 763-776.
- Hu, W.-P., Lin, H.-T., Tsai, J. J. P., & Chen, W.-Y. (2017). Investigating interactions between proteins and nucleic acids by computational approaches. *Computational Methods with Applications in Bioinformatics Analysis*, 98-117.
- Huang, L., Zhang, H., Deng, D., Zhao, K., Liu, K., Hendrix, D. A., & Mathews, D. H. (2019). LinearFold: linear-time approximate RNA folding by 5'-to-3' dynamic programming and beam search. *Bioinformatics*, 35(14), i295-i304.
- Huang, R.-H., Fremont, D. H., Diener, J. L., Schaub, R. G., & Sadler, J. E. (2009). A Structural Explanation for the Antithrombotic Activity of ARC1172, a DNA Aptamer that Binds von Willebrand Factor Domain A1. *Structure*, 17(11), 1476-1484.
- Ivani, I., Dans, P. D., Noy, A., Pérez, A., Faustino, I., Hospital, A., Walther, J., Andrio, P., Goñi, R., Balaceanu, A., Portella, G., Battistini, F., Gelpí, J. L., González, C., Vendruscolo, M., Laughton, C. A., Harris, S. A., Case, D. A., & Orozco, M. (2015). Parmbsc1: A refined force field for DNA simulations. *Nature Methods*, 13(1), 55-58.
- Ivry, T., Michal, S., Avihoo, A., Sapiro, G., & Barash, D. (2009). An image processing approach to computing distances between RNA secondary structures dot plots. *Algorithms for Molecular Biology*, 4(1), 4.
- Jabbari, H., Wark, I., Montemagno, C., & Will, S. (2018). Knotty: efficient and accurate prediction of complex RNA pseudoknot structures. *Bioinformatics*, 34(22), 3849-3856.
- Jenison, R. D., Gill, S. C., Pardi, A., & Polisky, B. (1994). High-resolution molecular discrimination by RNA. *Science*, 263(5152), 1425-1429.
- Ji, H., Deng, H., Lu, H., & Zhang, Z. (2020). Predicting a Molecular Fingerprint from an Electron Ionization Mass Spectrum with Deep Neural Networks. *Analytical Chemistry*, 92(13), 8649-8653.
- Johnson, S. C., Sherrill, C. B., Marshall, D. J., Moser, M. J., & Prudent, J. R. (2004). A third base pair for the polymerase chain reaction: Inserting isoC and isoG. *Nucleic Acids Research*, 32(6), 1937-1941.

- Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., & Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2), 926-935.
- Joseph L. Kim, Dimitar B. Nikolov, & Stephen K. Burley. (1993). Co-crystal structure of TBP recognizing the minor groove of a TATA element. *Nature*, 365, 520-527.
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583-589.
- Kalvari, I., Nawrocki, E. P., Ontiveros-Palacios, N., Argasinska, J., Lamkiewicz, K., Marz, M., Griffiths-Jones, S., Toffano-Nioche, C., Gautheret, D., Weinberg, Z., Rivas, E., Eddy, S. R., Finn, R. D., Bateman, A., & Petrov, A. I. (2021). Rfam 14: Expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Research*, 49(D1), D192-D200.
- Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley & Sons Inc.
- Kaur, H., Li, J. J., Bay, B.-H., & Yung, L.-Y. L. (2013). Investigating the Antiproliferative Activity of High Affinity DNA Aptamer on Cancer Cells. *PLoS ONE*, 8(1), e50964.
- Ke, A., & Doudna, J. A. (2004). Crystallization of RNA and RNA-protein complexes. *Methods*, 34(3), 408-414.
- Khvorova, A., & Watts, J. K. (2017). The chemical evolution of oligonucleotide therapies of clinical utility. *Nature Biotechnology*, 35(3), 238-248.
- Kim, S. (1978). Three-Dimensional Structure of Transfer Rna and Its Functional Implications. *Advances in Enzymology and Related Areas of Molecular Biology*, 46, 279-315.
- Krüger, A., Zimbres, F. M., Kronenberger, T., & Wrenger, C. (2018). Molecular modeling applied to nucleic acid-based molecule development. *Biomolecules*, 8(3), 83.
- Kusi-Appauh, N., Ralph, S. F., van Oijen, A. M., & Spenkeliink, L. M. (2023). Understanding G-Quadruplex Biology and Stability Using Single-Molecule Techniques. *Journal of Physical Chemistry B*, 127(25), 5521-5540.
- Lacroix, A., Edwardson, T. G. W., Hancock, M. A., Dore, M. D., & Sleiman, H. F. (2017). Development of DNA Nanostructures for High-Affinity Binding to Human Serum Albumin. *Journal of the American Chemical Society*, 139(21), 7355-7362.
- Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P. D., Smith, C. A., Sheffler, W., Davis, I. W., Cooper, S., Treuille, A., Mandell, D. J., Richter, F., Ban, Y. E. A., Fleishman, S. J., Corn, J. E., Kim, D. E., ... Bradley, P. (2011). Rosetta3: An object-oriented software suite for the simulation and design of macromolecules. *Methods in Enzymology*, 487(C), 545-574.
- Lee, I., & Berdis, A. J. (2010). Non-natural nucleotides as probes for the mechanism and fidelity of DNA polymerases. *Biochimica et Biophysica Acta - Proteins and Proteomics*, 1804(5), 1064-1080.
- Lee, R. C., Feinbaum, R. L., & Ambros, V. (1993). The C. elegans heterochronic gene lin-4

- encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5), 843-854.
- Leigh, K. E., Navarro, P. P., Scaramuzza, S., Chen, W., Zhang, Y., Castaño-Díez, D., & Kudryashev, M. (2019). Subtomogram averaging from cryo-electron tomograms. *Methods in Cell Biology*, 152, 217-259.
- Leontis, N. B., & Westhof, E. (2001). Geometric nomenclature and classification of RNA base pairs. *Rna*, 7(4), 499-512.
- Lescoute, A., & Westhof, E. (2006). Topology of three-way junctions in folded RNAs. *Rna*, 12(1), 83-93.
- Lin, C. H., & Patel, D. J. (1997). Structural basis of DNA folding and recognition in an AMP-DNA aptamer complex: Distinct architectures but common recognition motifs for DNA and RNA aptamers complexed to AMP. *Chemistry and Biology*, 4(11), 817-832.
- Lindert, S., Meiler, J., & McCammon, J. A. (2013). Iterative molecular dynamics - Rosetta protein structure refinement protocol to improve model quality. *Journal of Chemical Theory and Computation*, 9(8), 3843-3847.
- Loussouarn, C. (2014). *Sélection et caractérisation d'aptamères oligonucléotidiques régulateurs de la protéine STAT5B, impliquée dans les leucémies*. Université de Technologie de Compiègne.
- Lu, X. J. (2020). DSSR-enabled innovative schematics of 3D nucleic acid structures with PyMOL. *Nucleic Acids Research*, 48(13), e74-e74.
- Luo, Y., Zhou, J., Watt, S. K., Lee, V. T., Dayie, T. K., & Sintim, H. O. (2012). Differential binding of 2'-biotinylated analogs of c-di-GMP with c-di-GMP riboswitches and binding proteins. *Molecular BioSystems*, 8(3), 772-778.
- Mackerrill, A. D. (1997). Influence of Magnesium Ions on Duplex DNA Structural, Dynamic, and Solvation Properties. *The Journal of Physical Chemistry B*, 101(4), 646-650.
- Manoharan, M. (2002). Oligonucleotide Conjugates as Potential Antisense Drugs with Improved Uptake, Biodistribution, Targeted Delivery, and Mechanism of Action. *Antisense and Nucleic Acid Drug Development*, 12(2), 103-128.
- Marchand, B., Will, S., Berkemer, S. J., Bulteau, L., & Ponty, Y. (2022). Automated Design of Dynamic Programming Schemes for RNA Folding with Pseudoknots. *Algorithms for molecular biology : AMB*, 18(1), 18.
- Marušič, M., Toplišek, M., & Plavec, J. (2023). NMR of RNA - Structure and interactions. *Current Opinion in Structural Biology*, 79, 102532.
- Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, M., & Turner, D. H. (2004). Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proceedings of the National Academy of Sciences of the United States of America*, 101(19), 7287-7292.
- Mathews, D. H., Sabina, J., Zuker, M., & Turner, D. H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology*, 288(5), 911-940.

- Miao, Y., Nichols, S. E., Gasper, P. M., Metzger, V. T., & McCammon, J. A. (2013). Activation and dynamic network of the M2 muscarinic receptor. *Proceedings of the National Academy of Sciences*, *110*(27), 10982-10987.
- Mishra, S. K., Jain, N., Shankar, U., Tawani, A., Sharma, T. K., & Kumar, A. (2019). Characterization of highly conserved G-quadruplex motifs as potential drug targets in *Streptococcus pneumoniae*. *Scientific Reports*, *9*(1).
- Mitrovich, Q. M., & Guthrie, C. (2007). Evolution of small nuclear RNAs in *S. cerevisiae*, *C. albicans*, and other hemiascomycetous yeasts. *RNA*, *13*(12), 2066-2080.
- Mooers, B. H. M. (2009). Crystallographic studies of DNA and RNA. *Methods*, *47*(3), 168-176.
- Moumné, L., Marie, A.-C., & Crouvezier, N. (2022). Oligonucleotide Therapeutics: From Discovery and Development to Patentability. *Pharmaceutics*, *14*(2), 260.
- Murray, L. J. W., Arendall, W. B., Richardson, D. C., & Richardson, J. S. (2003). RNA backbone is rotameric. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(24), 13904-13909.
- Neidle, S. (2021). Beyond the double helix: DNA structural diversity and the PDB. *The Journal of Biological Chemistry*, *296*, 100553.
- Ni, S., Zhuo, Z., Pan, Y., Yu, Y., Li, F., Liu, J., Wang, L., Wu, X., Li, D., Wan, Y., Zhang, L., Yang, Z., Zhang, B.-T., Lu, A., & Zhang, G. (2021). Recent Progress in Aptamer Discoveries and Modifications for Therapeutic Applications. *ACS Applied Materials & Interfaces*, *13*(8), 9500-9519.
- Noble, A. J., Dandey, V. P., Wei, H., Brasch, J., Chase, J., Acharya, P., Tan, Y. Z., Zhang, Z., Kim, L. Y., Scapin, G., Rapp, M., Eng, E. T., Rice, W. J., Cheng, A., Negro, C. J., Shapiro, L., Kwong, P. D., Jeruzalmi, D., des Georges, A., ... Carragher, B. (2018). Routine single particle CryoEM sample and grid characterization by tomography. *eLife*, *7*.
- Nowakowski, J., & Tinoco, I. (1997). RNA Structure and Stability. *Seminars in Virology*, *8*(3), 153-165.
- Oweida, T. J., Kim, H. S., Donald, J. M., Singh, A., & Yingling, Y. G. (2021). Assessment of AMBER Force Fields for Simulations of ssDNA. *Journal of Chemical Theory and Computation*, *17*(2), 1208-1217.
- Parisien, M., & Major, F. (2008). The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, *452*(7183), 51-55.
- Pérez, A., Marchán, I., Svozil, D., Sponer, J., Cheatham, T. E., Laughton, C. A., & Orozco, M. (2007). Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of  $\alpha/\gamma$  Conformers. *Biophysical Journal*, *92*(11), 3817-3829.
- Petrov, A. I., Zirbel, C. L., & Leontis, N. B. (2013). Automated classification of RNA 3D motifs and the RNA 3D Motif Atlas. *RNA*, *19*(10), 1327-1340.
- Popenda, M., Miskiewicz, J., Sarzynska, J., Zok, T., & Szachniuk, M. (2020). Topology-based classification of tetrads and quadruplex structures. *Bioinformatics*, *36*(4), 1129-1134.
- Popenda, M., Szachniuk, M., Antczak, M., Purzycka, K. J., Lukasiak, P., Bartol, N., Blazewicz, J.,

- & Adamiak, R. W. (2012). Automated 3D structure composition for large RNAs. *Nucleic Acids Research*, *40*(14), e112-e112.
- Popenda, M., Szachniuk, M., Blazewicz, M., Wasik, S., Burke, E. K., Blazewicz, J., & Adamiak, R. W. (2010). RNA FRABASE 2.0: an advanced web-accessible database with the capacity to search the three-dimensional fragments within RNA structures. *BMC Bioinformatics*, *11*(1), 231.
- Reidys, C. M., Huang, F. W. D., Andersen, J. E., Penner, R. C., Stadler, P. F., & Nebel, M. E. (2011). Topology and prediction of RNA pseudoknots. *Bioinformatics*, *27*(8), 1076-1085.
- Reynolds, M. A., Hogrefe, R. I., Jaeger, J. A., Schwartz, D. A., Riley, T. A., Marvin, W. B., Daily, W. J., Vaghefi, M. M., Beck, T. A., Knowles, S. K., Klem, R. E., & Arnold, L. J. (1996). Synthesis and Thermodynamics of Oligonucleotides Containing Chirally Pure RP Methylphosphonate Linkages. *Nucleic Acids Research*, *24*(22), 4584-4591.
- Roberts, T. C., Langer, R., & Wood, M. J. A. (2020). Advances in oligonucleotide drug delivery. *Nature reviews. Drug discovery*, *19*(10), 673-694.
- Rocher, V., Genais, M., Nassereddine, E., & Mourad, R. (2021). DeepG4: A deep learning approach to predict cell-type specific active G-quadruplex regions. *PLoS Computational Biology*, *17*(8), e1009308.
- Roehr, B. (1998). Fomivirsen approved for CMV retinitis. *Journal of the International Association of Physicians in AIDS Care*, *4*(10), 14-16.
- Ropii, B., Bethasari, M., Anshori, I., Koesoema, A. P., Shalannanda, W., Satriawan, A., Setianingsih, C., Akbar, M. R., & Aditama, R. (2023). The assessment of molecular dynamics results of three-dimensional RNA aptamer structure prediction. *PLoS ONE*, *18*(7 July), e0288684.
- Rother, M., Rother, K., Puton, T., & Bujnicki, J. M. (2011). ModeRNA: a tool for comparative modeling of RNA 3D structure. *Nucleic Acids Research*, *39*(10), 4007-4022.
- Rould, M. A., Perona, J. J., Söll, D., & Steitz, T. A. (1989). Structure of E. coli Glutamyl-tRNA Synthetase Complexed with tRNA Gln and ATP at 2.8 Å Resolution. *Science*, *246*(4934), 1135-1142.
- Ruigrok, V. J. B., Levisson, M., Hekelaar, J., Smidt, H., Dijkstra, B. W., & van der Oost, J. (2012). Characterization of aptamer-protein complexes by x-ray crystallography and alternative approaches. *International Journal of Molecular Sciences*, *13*(8), 10537-10552.
- SantaLucia, J. (1998). A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(4), 1460-1465.
- Sato, K., Akiyama, M., & Sakakibara, Y. (2021). RNA secondary structure prediction using deep learning with thermodynamic integration. *Nature Communications*, *12*(1), 941.
- Sato, K., Hamada, M., Asai, K., & Mituyama, T. (2009). CentroidFold: A web server for RNA secondary structure prediction. *Nucleic Acids Research*, *37*, W277–W280.
- Schnieders, R., Knezic, B., Zetsche, H., Sudakov, A., Matzel, T., Richter, C., Hengesbach, M., Schwalbe, H., & Fürtig, B. (2020). NMR Spectroscopy of Large Functional RNAs: From

- Sample Preparation to Low-Gamma Detection. *Current Protocols in Nucleic Acid Chemistry*, 82(1).
- Schrödinger, L. (2015). *The PyMOL Molecular Graphics System, Version 1.8*.
- Shen, T., & Hamelberg, D. (2008). A statistical analysis of the precision of reweighting-based simulations. *Journal of Chemical Physics*, 129(3), 034103.
- Shiflett, P. R., Taylor-McCabe, K. J., Michalczyk, R., Silks, L. A. "Pete", & Gupta, G. (2003). Structural Studies on the Hairpins at the 3' Untranslated Region of an Anthrax Toxin Gene. *Biochemistry*, 42(20), 6078-6089.
- Singh, J., Hanson, J., Paliwal, K., & Zhou, Y. (2019). RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nature Communications*, 10(1), 5407.
- Sloma, M. F., & Mathews, D. H. (2016). Exact calculation of loop formation probability identifies folding motifs in RNA secondary structures. *RNA*, 22(12), 1808-1818.
- Smola, M., Rice, G., Busan, S., Siegfried, A. N. & Weeks, M. K (2015). Selective 2'-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. *Nat Protoc*, 10, 1643–1669.
- Sosnick, T. R., Fang, X., & Shelton, V. M. (2000). Application of circular dichroism to study RNA folding transitions. In *Methods in Enzymology*, 317, 393-409.
- Staple, D. W., & Butcher, S. E. (2005). Pseudoknots: RNA structures with diverse functions. *PLoS Biology*, 3(6), 0956-0959.
- Swaminathan, V., Sundaralingam, M., & Bau, R. (1979). The Crystal Structures of Metal Complexes of Nucleic Acids and Their Constituent. *CRC Critical Reviews in Biochemistry*, 6(3), 245-336.
- Szeberenyi, J., Roy, M. K., Vaidya, H. C., & Apirion, D. (1984). 7S RNA, containing 5S ribosomal RNA and the termination stem, is a specific substrate for the two RNA processing enzymes RNase III and RNase E. *Biochemistry*, 23(13), 2952-2957.
- Tan, Z., Fu, Y., Sharma, G., & Mathews, D. H. (2017). TurboFold II: RNA structural alignment and secondary structure prediction informed by multiple homologs. *Nucleic Acids Research*, 45(20), 11570-11581.
- Tang, K., Lu, Y. Y., & Sun, F. (2018). Background adjusted alignment-free dissimilarity measures improve the detection of horizontal gene transfer. *Frontiers in Microbiology*, 9(APR).
- Tubiana, T., Carvaille, J. C., Boulard, Y., & Bressanelli, S. (2018). TTClust: A Versatile Molecular Simulation Trajectory Clustering Program with Graphical Summaries. *Journal of Chemical Information and Modeling*, 58(11), 2178-2182.
- Ussery, D. W. (2002). DNA Structure: A-, B- and Z-DNA Helix Families. In *eLS* (Ed.). Wiley.
- Van Buuren, B. N. M., Overmars, F. J. J., Ippel, J. H., Altona, C., & Wijmenga, S. S. (2000). Solution structure of a DNA three-way junction containing two unpaired thymidine bases. Identification of sequence features that decide conformer selection. *Journal of Molecular Biology*, 304(3), 371-383.

- Van Dongen, M. J. P., Doreleijers, J. F., Van der Marel, G. A., Van Boom, J. H., Hilbers, C. W., & Wijmenga, S. S. (1999). Structure and mechanism of formation of the H-γ5 isomer of an intramolecular DNA triple helix. *Nature Structural Biology*, 6(9), 854-859.
- Wang, G., Zhang, Z. T., Jiang, B., Zhang, X., Li, C., & Liu, M. (2014). Recent advances in protein NMR spectroscopy and their implications in protein therapeutics research. *Analytical and Bioanalytical Chemistry*, 406(9-10), 2279-2288.
- Wang, J., Arantes, P. R., Bhattarai, A., Hsu, R. V., Pawnikar, S., Huang, Y. M., Palermo, G., & Miao, Y. (2021). Gaussian accelerated molecular dynamics: Principles and applications. *WIREs Computational Molecular Science*, 11(5).
- Wang, J., Cieplak, P., & Kollman, P. A. (2000). How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *Journal of Computational Chemistry*, 21(12), 1049.
- Watkins, A. M., Rangan, R., & Das, R. (2020). FARFAR2: Improved De Novo Rosetta Prediction of Complex Global RNA Folds. *Structure (London, England : 1993)*, 28(8), 963-976.e6.
- Westhof, E., & Leontis, N. B. (2021). An RNA-centric historical narrative around the Protein Data Bank. *The Journal of Biological Chemistry*, 296.
- Win, M. N., Klein, J. S., & Smolke, C. D. (2006). Codeine-binding RNA aptamers and rapid determination of their binding constants using a direct coupling surface plasmon resonance assay. *Nucleic Acids Research*, 34(19), 5670-5682.
- Xu, G., Zhao, J., Liu, N., Yang, M., Zhao, Q., Li, C., & Liu, M. (2019). Structure-guided post-SELEX optimization of an ochratoxin A aptamer. *Nucleic Acids Research*, 47(11), 5963-5972.
- Yan, S., Ilgu, M., Nilsen-Hamilton, M., & Lamm, M. H. (2022). Computational Modeling of RNA Aptamers: Structure Prediction of the Apo State. *Journal of Physical Chemistry B*.
- Yang, X.-L., Sugiyama, H., Ikeda, S., Saito, I., & Wang, A. H.-J. (1998). Structural Studies of a Stable Parallel-Stranded DNA Duplex Incorporating Isoguanine:Cytosine and Isocytosine:Guanine Basepairs by Nuclear Magnetic Resonance Spectroscopy. *Biophysical Journal*, 75(3), 1163-1171.
- Zgarbová, M., Šponer, J., Otyepka, M., Cheatham, T. E., Galindo-Murillo, R., & Jurečka, P. (2015). Refinement of the Sugar-Phosphate Backbone Torsion Beta for AMBER Force Fields Improves the Description of Z- and B-DNA. *Journal of Chemical Theory and Computation*, 11(12), 5723-5736.
- Zhang, W., Yang, F., Ou, D., Lin, G., Huang, A., Liu, N., & Li, P. (2019). Prediction, docking study and molecular simulation of 3D DNA aptamers to their targets of endocrine disrupting chemicals. *Journal of Biomolecular Structure and Dynamics*, 37(16), 4274-4282.
- Zhang, X., Lee, I., & Berdis, A. J. (2005). The Use of Nonnatural Nucleotides to Probe the Contributions of Shape Complementarity and  $\pi$ -Electron Surface Area during DNA Polymerization. *Biochemistry*, 44(39), 13101-13110.
- Zheng, Y. Y., Wu, Y., Begley, T. J., & Sheng, J. (2021). Sulfur modification in natural RNA and therapeutic oligonucleotides. *RSC Chemical Biology*, 2(4), 990-1003.
- Zubradt, M., Gupta, P., Persad, S., Lambowitz, M. A., Weissman, S. J. & Rouskin, S. (2017) DMS-

- MaPseq for genome-wide or targeted RNA structure probing in vivo. *Nat Methods*, 14, 75–82.
- Zuker, M. (1989). On finding all suboptimal foldings of an RNA molecule. *Science*, 244(4900), 48-52.
- Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31(13), 3406-3415.
- Zuker, M., Jaeger, J. A., & Turner, D. H. (1991). A comparison of optimal and suboptimal RNA secondary structures predicted by free energy minimization with structures determined by phylogenetic comparison. *Nucleic Acids Research*, 19(10), 2707-2714.



## Annexes

Annexe 1. Jeu de donnée ADN obtenu de la PDB. Plusieurs informations complémentaires sont reportées : séquence, méthode de résolution, taille, présence de G-quadruplexes, et si l'oligonucléotide est extrait d'un complexe.

<b>PDB</b>	<b>Séquence</b>	<b>Obtention</b>	<b>Taille</b>	<b>Quadruplexes</b>	<b>En Complexe</b>
1PQT	GCGAAGC	NMR	7	Non	Non
5GWL	CCTGCCTG	NMR	8	Non	Non
5GWQ	TTTATTTA	NMR	8	Non	Non
6J37	CTTGCTTG	NMR	8	Non	Non
6MOB	CTTGCGTG	NMR	8	Non	Non
6MOC	CTTGCATG	NMR	8	Non	Non
2K71	GCGAAAGC	NMR	8	Non	Non
5OND	GCGGTAGTC	X-RAY	9	Non	Oui
1ZHU	CAATGCAATG	NMR	10	Non	Non
6IY5	ATTCTATTCT	NMR	10	Non	Non
2LO8	GCCGCAGTGC	NMR	10	Non	Non
2A0I	TGGGGTGTGG	X-RAY	10	Non	Oui
3WPD	CCTGGATGGG	X-RAY	10	Non	Oui
1BJH	GTACAAAGTAC	NMR	11	Non	Non
3WPG	CCTGGATGGGA	X-RAY	11	Non	Oui
2LO5	GGCCGCAGTGCC	NMR	12	Non	Non
6FKE	GTAGCGAAGCTA	X-RAY	12	Non	Oui
3WPH	CCTGGATGGGAA	X-RAY	12	Non	Oui
1LA8	CGCGGTGTCCGCG	NMR	13	Non	Non
1P0U	GCATCGACGATGC	NMR	13	Non	Non
5F55	GATGTACGCTAGGC	X-RAY	14	Non	Oui
2EXF	GTCCTGTTCGGGC	NMR	14	Non	Oui
2JZW	GTCCTGTTCGGGC	NMR	14	Non	Oui
6FK5	CTAGCGAAGCTAGA	X-RAY	14	Non	Oui
2M8Y	CGCGAAGCATTCGCG	NMR	15	Non	Non
1UUT	CAGCTCTTTGAGCTG	X-RAY	15	Non	Oui
1AC7	ATCCTAGTTATAGGAT	NMR	16	Non	Non
6FK4	GGCTAGCGAAGCTAGA	X-RAY	16	Non	Oui
1XUE	GTGGAATGCAATGGAAC	NMR	17	Non	Non
1EN1	GTCCTGTTCGGGCGCCA	NMR	18	Non	Non
4KB1	GGCCCTCTTTAGGGCCTC	X-RAY	18	Non	Oui
4KB0	GGCCCTCTTTAGGGCCCC	X-RAY	18	Non	Oui
1ECU	GCGCGAAACTGTTTCGCG C	NMR	19	Non	Non
3Q0A	GTCAAAGAAGCGGAGCT TC	X-RAY	20	Non	Oui
4FF1	TTCAAAGAAGCGGAGCT TC	X-RAY	20	Non	Oui

3C46	TCCAAAAGAAGCGGAGCT TCT	X-RAY	21	Non	Oui
3Q23	ATCCAAAAGAAGCGGAGC TTCT	X-RAY	22	Non	Oui
3Q24	ATCCAAAAGAAGCGGAGC TTCT	X-RAY	22	Non	Oui
2A60	CCCCTAGCTTTAGCTATG GGGA	X-RAY	22	Non	Oui
2L5K	CAGTTGATCCTTTGGATA CCCTG	NMR	23	Non	Non
3DSD	GCACAAGCTTTTGCTTGT GGATA	X-RAY	23	Non	Oui
2VHG	CCCCTAGCTTTAGCTATG GGGAGT	X-RAY	24	Non	Oui
1OSB	GCGCACCGAAAGGTGCGT ATTGTCT	X-RAY	25	Non	Oui
1ZM5	GCGCACCGAAAGGTGCGT ATTGTCT	X-RAY	25	Non	Oui
2CDM	CGCACCGAAAGGTGCGTA TTGTCTA	X-RAY	25	Non	Oui
2VIC	AAAGCCCCTAGCTTTTAG CTATGGGG	X-RAY	26	Non	Oui
5N2Q	ACTTTATGAAAATAAAGT ATAGTGTG	X-RAY	26	Non	Oui
1NGU	CTCTCCTTGTATTTCTTA CAAAAAGAG	NMR	27	Non	Non
1NGO	CTCTTTTTGTAAAGAAATA CAAGGAGAG	NMR	27	Non	Non
1JVE	CCTAATTATAACGAAGTT ATAATTAGG	NMR	27	Non	Non
3ZH2	CTGGGCGGTAGAACCATA GTGACCCAG	X-RAY	27	Non	Oui
4HT4	CGCGAACGGAACGTTTCGC ATAAGTGCGC	X-RAY	28	Non	Oui
1YTB	GTATATAAACGGGTGGC GTTTTATATAC	X-RAY	29	Non	Oui
1B4Y	TCTTCCTTTTCTTCTCC CGAGAAGTTTT	NMR	30	Non	Non
4ER8	GTAGGACGGATAAGGCGT TTACGCCGCATCCG	X-RAY	32	Non	Oui
6SEI	AATCGGCAATGACCTTTG GTCATTCAGCAGAT	X-RAY	32	Non	Oui
5HRU	TCGATTGGATTGTGCCGG AAGTGCTGGCTCGA	X-RAY	32	Non	Oui
4F41	CATAATAACAATATCTTG ATATTGTTATTATG	X-RAY	32	Non	Oui
4F43	CATAATAACAATATCAAG ATATTGTTATTATG	X-RAY	32	Non	Oui
5HTO	TTCGATTGGATTGTGCCG GAAGTGCTGGCTCGAA	X-RAY	34	Non	Oui
2VJU	GAATCCCCTAGCTTTAGC TATGGGGAGTATGTCAA	X-RAY	35	Non	Oui
1EZN	CGTGCACCCGCTTGCGGC GACTTGTCGTTGTGCACG	NMR	36	Non	Non
1SNJ	CGTGCAGCGGCTTGCCGG CACTTGCTTCTGCACG	NMR	36	Non	Non

6U82	GCTAATCTAATCAACCGC AGGTTGATTAGCCCATTA GC	X-RAY	38	Non	Oui
3HXO	GGCGTGCAGTGCCTTCGG CCGTGCGGTGCCTCCGTC ACGC	X-RAY	40	Non	Oui
2N8A	GCTGGCTTCGTAAGAAGC CAGCTCGCGGTCAGCTTG CTGACCGCG	NMR	45	Non	Oui
3THW	CCTCTATCTGAAGCCGAT CGATGAAGCATCGATCGC ACAGCTTCAGATAGAGG	X-RAY	53	Non	Oui
148D	GGTTGGTGTGGTTGG	NMR	15	Oui	Non
1HAO	GGTTGGTGTGGTTGG	X-RAY	15	Oui	Oui
2N21	TTGGGTGGGTGGGTGGGT	NMR	18	Oui	Oui
5NYS	TAGGGACGGGCGGGCAGG GT	NMR	20	Oui	Non
1I34	GGTTTTGGCAGGGTTTTG GT	NMR	20	Oui	Non
2KF8	GGGTTAGGGTTAGGGTTA GGGT	NMR	22	Oui	Non
1OZ8	GGAGGAGGAGGAGGAGGA GGAGGA	NMR	24	Oui	Non
5VHE	AGGGTGGGTAGGGTGGGT TTTTTT	X-RAY	24	Oui	Oui
2HY9	AAAGGGTTAGGGTTAGGG TTAGGGAA	NMR	26	Oui	Non
7CV4	GGGAGGGCGCGCCAGCGG GGTCGGGC	NMR	26	Oui	Non
6EVV	CGCCTAGGTTGGGTAGGG TGGTGGCG	X-RAY	26	Oui	Oui
4I7Y	GTCCGTGGTAGGGCAGGT TGGGGTGAC	NMR	27	Oui	Oui
7CV3	GCGGGAGGGCGCGCCAGC GGGGTCGGG	NMR	27	Oui	Non
2M8Z	GGTTGGCGCGAAGCATT CGGGTTGG	NMR	27	Oui	Non
6H1K	GGGAGGCGTGGCCTGGGC GGGACTGGGG	NMR	28	Oui	Non
2M91	GGGAAGGGCGCGAAGCAT TCGCGAGGTAGG	NMR	30	Oui	Non
5CMX	TGACGTAGGTTGGTGTGG TTGGGGCGTCAC	X-RAY	30	Oui	Oui
2M93	TTGGGTGGGCGCGAAGCA TTCGCGGGGTGGGT	NMR	32	Oui	Non
6SUU	AGGGCGGTTTTGGGAAGAG GGAAGAGGGGGAGG	NMR	32	Oui	Non
6T2G	AGGGCGGTGTGGGAAGAG GGAAGATGGGGAGG	NMR	32	Oui	Non
2M90	GCGCGAAGCATTGCGGGG GAGGTGGGGAAGGG	NMR	32	Oui	Non
7CLS	TTGGATCTGAGAATCAGA TGTGGGTGGGTGGGT	NMR	33	Oui	Non
5MTA	CAGGGTTAAGGGTATAAC TTTAGGGGTTAGGGTT	NMR	34	Oui	Non

<i>2M92</i>	AGGGTGGGTGCTGGGGCG CGAAGCATTGCGGAGG	NMR	34	Oui	Non
<i>6ZL9</i>	GATCAGTTTTACTGATCG GGTGGGTAGGGTGGGTA	NMR	35	Oui	Non
<i>6ZL2</i>	TGAGGGTGGGTAGGGTGG GCTAGTCATTTTACTAG	NMR	36	Oui	Non
<i>6ZTE</i>	GATCAGTTTTACTGATCG GGTGGTGGGTGGGAAGG	NMR	36	Oui	Non

Annexe 2. Jeu de donnée ADN obtenu de la PDB. Plusieurs informations complémentaires sont reportées : séquence, méthode de résolution, taille, et si l'oligonucléotide est extrait d'un complexe

<b>PDB</b>	<b>Séquence</b>	<b>Obtention</b>	<b>Taille</b>	<b>En Complexe</b>
2IXZ	GCUGUGCC	NMR	8	Non
2OJ7	GCUGUUGU	NMR	8	Non
1IDV	GGGCGUGCCC	NMR	10	Non
1R4H	GGGCAAGCCC	NMR	10	Non
2MXJ	CCAGAAACGGA	NMR	11	Non
1AFX	GGUGUGAACACC	NMR	12	Non
1RNG	GGCGCUUGCGUC	NMR	12	Non
1ZIF	GGGCGAAAGCCU	NMR	12	Non
1ZIG	GGGCGAGAGCCU	NMR	12	Non
1ZIH	GGGCGCAAGCCU	NMR	12	Non
2F87	GGCUGAAGGGCC	NMR	12	Non
5FMZ	AGUAGUAACAAG	X-RAY	12	Yes
1ESH	GGUGCAUAGCACC	NMR	13	Non
1HS1	GCGUUAACUCGCA	NMR	13	Non
1HS2	GCGUUAAGUCGCA	NMR	13	Non
1HS3	GCGUUAUUCGCA	NMR	13	Non
1HS4	GCGUUAUUCGCA	NMR	13	Non
1HS8	GCGUCAUUCGCA	NMR	13	Non
1I46	GGUGCGUAGCACC	NMR	13	Non
1I4B	GGUGCUUAGCACC	NMR	13	Non
1JZC	GGUGCAUAGCACC	NMR	13	Non
1VOP	GACUGGGGCGGUC	NMR	13	Non
4Z0C	ACGGAAAGACCCC	X-RAY	13	Yes
6FQ3	GGAGUCCAACUCC	X-RAY	13	Yes
6FQL	UGCAUUUAAUGCA	X-RAY	13	Yes
1F85	GGCCUGAUAGGGUC	NMR	14	Non
1FHK	GGCGGUGAAAUGCC	NMR	14	Non
1IK1	GGUACUAUGUACCA	NMR	14	Non
1K4A	GGUUCAGAAGAACC	NMR	14	Non
1K4B	GGUUCAGUUGAACC	NMR	14	Non
1ROQ	GGUAUCACGGUACC	NMR	14	Non
2EVY	GGUAUGCUAGUACC	NMR	14	Non
2KOC	GGCACUUCGGUGCC	NMR	14	Non
2Y95	GGCGCAUCGGCGCC	NMR	14	Non
4Z7L	GCAAAUAACAAGC	X-RAY	14	Yes
1A4T	GCGCUGACAAAGCGC	NMR	15	Yes
1ATW	GCUCCAGAUGGAGCG	NMR	15	Non
1OQ0	GAGAGUUGGGCUCUC	NMR	15	Non
1Q75	GGCUCUCAGUGAGCC	NMR	15	Non
1QFQ	GCCUGAAAAAGGGC	NMR	15	Yes
1XWP	GGAGAUCGCACUCCA	NMR	15	Non

2LPA	GAGGACAUAGUCUUC	NMR	15	Non
4AL7	CUGCCGUUAGGCAG	X-RAY	15	Yes
1JTW	GGGUGCGAGAGCGUCA	NMR	16	Non
1JWC	GGCCUUUUCAGGGCC	NMR	16	Non
1XWU	CGAAACAUAGAUUCGA	NMR	16	Non
2L6I	GAUCUCUUGUAGAUCA	NMR	16	Non
2LP9	GAGGACAUAGAUCUUC	NMR	16	Non
2MNC	CCGUUGAAUCUCACGG	NMR	16	Non
4AL5	ACUGCCGUUAGGCAG	X-RAY	16	Yes
4ILM	GCUAAUCUACUUAAGA	X-RAY	16	Yes
4QIL	UCCCUUCUGUGAAGGG	X-RAY	16	Yes
1ATV	GGGACCAGAAGGUCCCG	NMR	17	Non
1BZ2	UCAGACUUUUAUCUGA	NMR	17	Non
1BZ3	UCAGACUUUUAUCUGA	NMR	17	Non
1KKA	GGGGAUUGAAAUCCCC	NMR	17	Non
1WKS	GGCUUUGGAUAAAAGCC	NMR	17	Non
1YN1	GCGAGUUGACUACUCGC	NMR	17	Non
2JR4	CCUCCCUUACAAGGAGG	NMR	17	Non
2KPC	UGAGCACAGUUUGCUCA	NMR	17	Non
2KPD	UGAGCUCAGUUUGCUCA	NMR	17	Non
2KRP	CUCGGCUACGAACCGAG	NMR	17	Non
2KVN	GGUGAUUGAGUUCACCA	NMR	17	Non
2LAC	GGGGACUGUAAAUCCCC	NMR	17	Non
2LBJ	GGGCCUUGCCAAGGUCC	NMR	17	Non
2LBK	GGGACCUUCCCGGUCUC	NMR	17	Non
2LBL	GGGACCUUCCAAGUCUC	NMR	17	Non
2M4W	GGAAUCGAAAGAUGUCC	NMR	17	Non
4ZLD	UAACUUCUGUGAAGUUG	X-RAY	17	Yes
6CYT	AUCUGAGCCUGGGAGCU	X-RAY	17	Yes
1Z30	GGCGUUCGUUAGAACGUC	NMR	18	Non
2GVO	GAGGUCGGGAUGGAUCUC	NMR	18	Non
2QH4	GGCACAGAGUUAUGUGCC	NMR	18	Non
2Y9H	UCCCCACGCGUGUGGGGA	X-RAY	18	Yes
1ATO	GGCACCUCUCGCGGUGCC	NMR	19	Non
1ESY	GGCGACUGGUGAGUACGCC	NMR	19	Non
1I3X	GGCUGGCUGUUCGCCAGCC	NMR	19	Non
1SLP	UUACCCAAGUUUGAGGUAA	NMR	19	Non
1UUU	GGCGUACGUUUCGUACGCC	NMR	19	Non
2B6G	GGAGGCUCUGGCAGCUUUC	NMR	19	Yes
2B7G	GGAGGCUCUGGCAGCUUUC	NMR	19	Non
2MEQ	GGCCGUAACUUAACGGUC	NMR	19	Non
2MFD	GGCGUUCGCUUAGAACGUC	NMR	19	Non
2RLU	GGUUGCGGGUCUCGCAACC	NMR	19	Non
2Y8Y	UCCCCACGCGUGUGGGGAU	X-RAY	19	Yes
4QI2	AUGUUUUCUGUGAAAACGG	X-RAY	19	Yes
5N5C	AUUCUUAAAUAAGGAGU	NMR	19	Non

6TQB	GGAAAUUAUAUUAAUUUCC	X-RAY	19	Yes
1A1T	GGACUAGCGGAGGCUAGUCC	NMR	20	Yes
1HLX	GGGAUAACUUCGGUUGUCCC	NMR	20	Non
1MFJ	GACAGUCUCUACGGAGACUG	NMR	20	Non
1U2A	GGUCAGUGUAACAACUGACC	NMR	20	Non
2JPP	GGGCUUCACGGAUGAAGCCC	NMR	20	Yes
2O33	GGUUUGCCUUUUGGCUUACC	NMR	20	Non
2RPK	GGGAUCCAUGACAGGAUCCC	NMR	20	Non
2RPT	GGCCCGCCGAAAGGCCGGCC	NMR	20	Non
2Y8W	UCCCCACGCGUGUGGGGAUG	X-RAY	20	Yes
4L8H	AUGCAUGUCUAAGACAGCAU	X-RAY	20	Yes
5F5F	UGACUGCGUUUUAGGAGUUA	X-RAY	20	Yes
6PK9	GGAGGGUAGACUCGCUCUCC	NMR	20	Non
17RA	GGCGUAAGGAUUACCUAUGCC	NMR	21	Non
1D0U	GGGAUCACCAUUAGGGAUCUC	NMR	21	Non
1JOX	GGCGGUGCUGAGAUGCCCGUC	NMR	21	Non
1QWA	GGAUGCCUCCCAGUGCAUCC	NMR	21	Non
1RKJ	GGAUGCCUCCCAGUGCAUCC	NMR	21	Yes
1SZY	GGCAGGGCUCAUAACCCUGCC	NMR	21	Non
2FY1	GGACUGUCCACAAGACAGUCC	NMR	21	Yes
2M21	GGCGAUACACUAUUUAUCGCC	NMR	21	Non
2MFF	GGGAUCGCAGGAAGCGAUCCC	NMR	21	Yes
2MFG	GGGUCAUCAGGACGAUGACCC	NMR	21	Yes
5F5H	CCACACCGUUCUAGGUGCUGG	X-RAY	21	Yes
5ID6	AAUUUCUACUAAGUGUAGAUC	X-RAY	21	Yes
5L1Z	GAUCUGAGCCUGGGAGCUCUC	X-RAY	21	Yes
6XWJ	GGUGCCUAAUAUUUAGGCACC	NMR	21	Non
1F9L	GGCGAAGUCGAAAGAUGGCGCC	NMR	22	Non
1FJE	GGCCGAAAUCCCAGUAGGCC	NMR	22	Yes
1IKD	GGGGCUCUUCGGAGCUCCACCA	NMR	22	Non
1JUR	GGCCUGAGGAGACUCAGAAGCC	NMR	22	Non
1K2G	CAGACUUCGGUCGCAGAGAUGG	NMR	22	Non
1K6G	GGCGUCAUGAGUCCAUGGCGCC	NMR	22	Non
1K6H	GGCGUGUUCAGAAGAACGCGCC	NMR	22	Non
1N66	GGACCUCUCGAAAGAGUUUGUCC	NMR	22	Non
1OSW	GGAGGCGCUACGGCGAGGCUCC	NMR	22	Non
1PJY	GGCCUUCCCACAAGGGAAGGCC	NMR	22	Non
1TJZ	GGUGACGCCGUAAGGCGCAGCC	NMR	22	Non
2G1W	GGGGUGGCUCCCCUAACAGCCG	NMR	22	Non
2GRW	GGACCUCUCGAAAGAGAUGUCC	NMR	22	Non
2GV3	GGCCAGACUCCCAGAUUCUGGCC	NMR	22	Non
2GV4	GGACCUCUCGAAAGAGUGGUCC	NMR	22	Non
2HNS	GGCGUGAUCAAGUGAUCGCGCC	NMR	22	Non
2JSE	GGAGUGGCCGAAAGGCAUCUCC	NMR	22	Non
2JYM	GGCUCGCAGCAGGUCUGGAGUC	NMR	22	Non
2K66	GGAGUAUGUGAAAGCAUACUCC	NMR	22	Non

2KD8	GGAUGGUUGGGUUAGCCAUCC	NMR	22	Non
2M5U	GGCAGAUUCUGGUGAAUCUGCC	NMR	22	Non
2MFC	GGGUGUCGACGGAUAGACACCC	NMR	22	Yes
2MFE	GGGCCAUCAAGGACGAUGGUCC	NMR	22	Yes
2W2H	GCUCAGAUUCUGCGGUCUGAGC	X-RAY	22	Yes
4A4S	GGACCCGGCUCACGCUGGGUCC	NMR	22	Non
6F4H	GGCCGCAUUGCACCUCGCGGCC	X-RAY	22	Yes
6KYV	GGUAGACGCUUCGGCGUUUGCC	X-RAY	22	Yes
1BGZ	GGGAUACUGCUUCGGUAAGUCCC	NMR	23	Non
1BVJ	GGCGACGGUGUAAAAUUCUGCC	NMR	23	Non
1JTJ	CUUGCUGAAGCACGCACGGCAAG	NMR	23	Non
1K5I	GGACCCGGGCUCAACCUGGGUCC	NMR	23	Non
1MFK	GGCGGUUGCAGGUCUGCACCGCC	NMR	23	Non
1OW9	GAGCGAAGACGAAAGUCGAGCUC	NMR	23	Non
1S2F	GGGGAGUGGUUUUGUAUCCUCCC	NMR	23	Non
1TLR	GGCCUAAGACUUCGGUUAUGGCC	NMR	23	Non
2ANN	CGCGGAUCAGUCACCCAAGCGCG	X-RAY	23	Yes
2ES5	GGAGAGGCUCUGGCAGCUUUUCC	NMR	23	Non
2M12	GGGUGUAUUGGAAAUAGACACCC	NMR	23	Non
2M22	GGCAGAUUCUGUAAUAGAACUGCC	NMR	23	Non
2N0R	CUGAGCUCGAAAGAGCAAUGAUG	NMR	23	Non
2N2O	GCAUGUUUUCUGUGAAAACGGUU	NMR	23	Non
2N2P	GCAUGUUUAGUGUCUAAACGGUU	NMR	23	Non
2N3O	GGGACCUGGUCUUUCCAGGUCCC	NMR	23	Yes
2N7X	GGUAGUUUUGGCAUGACUCUACC	NMR	23	Non
2N82	GGUAGUUUUGGCAUGACUCUACC	NMR	23	Yes
2PJP	GGCGGUUGCAGGUCUGCACCGCC	X-RAY	23	Yes
2QH3	GGAGUGCCUACUGUGGCACUCC	NMR	23	Non
2RO2	GGGAGACCUGAAGUGGGUUUCCC	NMR	23	Non
2UW	GGCGUUGCCGGUCUGGCAACGCC	X-RAY	23	Yes
M				
3PHP	GGUCCGAGGGUCAUCGGAACCA	NMR	23	Non
5UF3	GGACAUAAGGAAAACCUAUGUCC	NMR	23	Non
5WQ1	GGACAUCAGAUUUCUGGUGUCC	NMR	23	Non
6GBM	GGCAGAUUACAAUUCUAUUUGCC	NMR	23	Yes
1A9N	CCUGGUAUUGCAGUACCUCCAGGU	X-RAY	24	Yes
1E4P	GUGCGAAGACGAAAGUCCGAGCGC	NMR	24	Non
1KKS	GGAAGGCCCUUUUCAGGGCCACCC	NMR	24	Non
1MT4	GGCGUAACGUUGAAAAGUUACGCC	NMR	24	Non
1NC0	GGUUCCCCUGCAUAAGGAGGAACC	NMR	24	Non
1NYB	GGUUCACCUCUAACCGGGUGAGCC	NMR	24	Yes
1RHT	GGGACUGACGAUCACGCAGUCUAU	NMR	24	Non
1SYZ	GGUUCCCCUGCAUAAGGAUGAACC	NMR	24	Non
1TFN	GGGACUGACGAUCACGCAGUCUAU	NMR	24	Non
2HEM	GGGAAGGCGCUUCGGCGUCGGCCC	NMR	24	Non
2LK3	GGCUUAGAUCAGAAAUGAUCAGCC	NMR	24	Non

2LV0	GGGCUAAUGUUGAAAAAUAGCCC	NMR	24	Non
2QH2	GGAGUGCCUGAGCUGUGGCACUCC	NMR	24	Non
3NVK	AGCUCUGACCGAAAGGCGUGAUGA	X-RAY	24	Yes
5F9F	GAAUAUAAUAGUGAUUUUAUUAUC	X-RAY	24	Yes
5NG6	AAUUUCUACUGUUGUAGAUAGAUU	X-RAY	24	Yes
5UDZ	GGGGUAGUGAUUUUACCCUGGAGA	X-RAY	24	Yes
1M82	GGAAGCAGGCUUCGGCCUUGUUUCC	NMR	25	Non
1QC8	GGCAGUGUGAGUACCUUCACACGUC	NMR	25	Non
6DU5	GGCUGGUGUGGUACAGAGAAGCCAG	X-RAY	25	Yes
6F4G	GCGGCCGUUUGCAGUACCGCGGCC	X-RAY	25	Yes
1QWB	GGACACGAAAUCCCGAAGUAGUGUCC	NMR	26	Non
2L5Z	GAGCUGCAGCACGAAAGUGACGGCUC	NMR	26	Non
4BW0	GGGGGAGCCGAAAGGCGAAGAACCCA	X-RAY	26	Yes
4QOZ	CCAAGGCUCUUUUCAGAGCCACCCA	X-RAY	26	Yes
4TV0	GGCCAAAGGCCCUUUUCAGGGCCACC	X-RAY	26	Yes
1F7F	GGAAGUCCGGUCUUCGGACCGGCUUCC	NMR	27	Non
1FQZ	GCCGAGUAGUGUUGGGUCGCGAAAGGC	NMR	27	Non
1FYO	GGCGUCGCACCUUCGGGUGAAGUCGCC	NMR	27	Non
1XSG	GGAAGACCGGUCUUCGGACCGGCUUCC	NMR	27	Non
1XSH	GGAAGCUCGGUCUUCGGACCGGCUUCC	NMR	27	Non
1YSV	GGUAACAUAUAGCUAAAUGUUGUUACC	NMR	27	Non
2AHT	GGAGCGGGGUGUAAACCUAUCGCUCC	NMR	27	Non
2IXY	GGCCUCCAAGCUGUGCCUUGGGUGGCC	NMR	27	Non
2LDL	GGAUCCAUUCGAUUAGUGAACGGAUCC	NMR	27	Non
2LJJ	GGCCUCAGCACUACCCCAGUGUAGGUC	NMR	27	Non
2LQZ	GGACUCCAUAUUGCUCGGCAAAGUCC	NMR	27	Non
484D	GGUGUCUUGGAGUGCUGAUCGGACACC	NMR	27	Yes
6XH0	GCAGAUCUGAGCCUGGGAGCUCUCUGC	X-RAY	27	Yes
1ZBN	GGCUCGUGUAGCUCAUUAGCUCCGAGC C	NMR	28	Yes
28SP	GGCGUCAGGUCCGGAAGGAAGCAGCGC C	NMR	28	Non
2GIP	GGCCAUCUUCUCUUCGGAGGAUUUGGC C	NMR	28	Non
2KMJ	GGCCAGAUUGAGCUUCGGCUCUCUGGU C	NMR	28	Non
2LUN	GGGCAGUGAUGCUUCGGCAUAUCAGCC C	NMR	28	Non
2NCO	GGGCUGAAGGAUGGAGACGUCUAGGCC C	NMR	28	Non
2NCI	GGACGUUAAAAGGCUUCGGCCUACGUC C	NMR	28	Non
5MOI	GAUAACUGAAUCGAAAGACAUUAUCAC G	X-RAY	28	Yes
6AAS	GGUAAGUGUACUGGAAAGUGCACUUGC C	NMR	28	Non
6SNJ	GGGAUUUCCCCAAAUGUGGGAAACUCC C	NMR	28	Yes
6VZC	GGGCUGUGAUGCUUCGGCAUAUCAGCC C	NMR	28	Non

1ANR	GGCAGAUCUGAGCCUGGGAGCUCUCUG CC	NMR	29	Non
1EBS	GGUGGGCGCAGCUUCGGCUGCGGUACA CC	NMR	29	Non
1F84	GGCCGAGUAGUGUUGGGUCGCGAAAGG CC	NMR	29	Non
1JBT	CGCUCCUCAGUACGAGAGGAACCGGAG CG	X-RAY	29	Yes
1L1C	GGAUUGUUACUGCUACGGCAGGCAAAA CC	NMR	29	Yes
1L1W	GGUGACCUCCTCGGGAGCGGGGACCAC CA	NMR	29	Non
1NBR	GGAGUGCUUCAAACAGUGCUUGGACGCU CC	NMR	29	Non
1OOA	CAUACUUGAAACUGUAAGGUUGGCGUA UG	X-RAY	29	Yes
1SCL	GGGUGCUCAGUACGAGAGGAACCGCAC CC	NMR	29	Non
2GIO	GGCCAUCUUGCUCUUCGGAGGAUUUGG CC	NMR	29	Non
2JWV	GAUACUUGAAACUGUAAGGUUGGCGUA UC	NMR	29	Non
2K5Z	GGUCUACAUUGCUGUUGUCGUGUGUGA CC	NMR	29	Non
2M24	GGAGUAUGUAUUGGCACUGAGCAUACU CC	NMR	29	Non
3SN2	GAUUUUCGGAAGCAGUGCCUCCAUA UC	X-RAY	29	Yes
5LM7	CUCUUUAACAUAAGCCCUGAAGAAGG GC	X-RAY	29	Yes
5LSN	CCUCGUGGUGGUUGUGAACCACCAUGU GG	NMR	29	Yes
6DU4	GGUUGGCGUAGGCUACAGAGAAGCCAA CC	X-RAY	29	Yes
1AUD	GGCAGAGUCCUUCGGGACAUUGCACCU GCC	NMR	30	Yes
1EBR	GGUGGGCGCAGCUUCGGCUGACGGUAC ACC	NMR	30	Non
1EKZ	GGACAGCUGUCCCUUCGGGGACAGCUG UCC	NMR	30	Yes
1HVU	AGAUUCCGUUUUCAGUCGGGAAAAACU GAA	X-RAY	30	Yes
1HWQ	GGUGCGAAGGGCGUCGUCGCCCCGAGC GCC	NMR	30	Non
1KP7	GGCGCUCAAUGCUCUUCGGAGACGACC GCC	NMR	30	Non
1LDZ	GCGACCGAGCCAGCGAAAGUUGGGAGU CGC	NMR	30	Non
1NA2	GGGCGUUUUUCUCGCGUGACUUUCAGC CCC	NMR	30	Non
1RFR	GGCACUCUGGUAUCACGGUACCUUUGU GUC	NMR	30	Non
5Y58	GGACUUAUAGAUGGCUAAAAUCUGAGU CCA	X-RAY	30	Yes
6MCE	GGGCAGAUUGAGCCUGGGAGCUCUCUG CCC	NMR	30	Yes

1J07	AGUAGAAACAAGGCUUCGGCCUGCUUU UGCU	NMR	31	Non
1MFY	AGUAGAAACAAGGCUUCGGCCUGCUUU CGCU	NMR	31	Non
1YNC	GGAUUAUGAUACA AUUUGAUCAGUAU AUCC	NMR	31	Non
1YNG	GGAUUAUGAUUA AUUUGAUCAGUAU AUCC	NMR	31	Non
2LDT	GGGAGUACGGCCGCAAGGUUAAAACU CCCC	NMR	31	Non
5KMZ	AACCUUCACCAAUUAGGUUCAAUAAG UGGU	NMR	31	Non
5UZT	GGUGUUGACUGUUGAAUCUCAUGGCAA CACC	NMR	31	Non
6HYK	GGCACGGUGAUGACCUUCGGGUCUGAG UGCC	NMR	31	Non
1G70	GGUCUGGGCGCACUUCGGUGACGGUAC AGGCC	NMR	32	Yes
1KAJ	GGCGCAGUGGGCUAGCGCCACUCAAAA GCCCG	NMR	32	Non
1KPD	GGCGCAGUGGGCUAGCGCCACUCAAAG GCCCG	NMR	32	Non
1XHP	GAGCAGUUC CCCUGCAUAAGGAUGAAC CGUUC	NMR	32	Non
1Z31	GAGGUCGGCCCGACUUCGGUCACUGCC ACCUC	NMR	32	Non
2LBS	GGGAUACCAUGUUCAAGUGAACGUGGU AUCUC	NMR	32	Yes
2LI4	GGACCGAUAAAGGUAGAAAUGCCUUAUC GGUCC	NMR	32	Non
2LUP	GGGAUACCAUGUUCAGAAGAACGUGGU AUCUC	NMR	32	Yes
5A18	GACGAUAUCGAGCAUCAAGAGUGAAUA UCGUC	NMR	32	Non
1EXY	GGGCGCCGGUACGCAAGUACGACGGUA CGCUCC	NMR	33	Yes
2JXV	GGAGGUAGUAGGUCGAAAGACCAUUCU GCCUCC	NMR	33	Non
3ID5	GGGGCGCCUCUGAGCGUUCGCGCUGU GAUGAA	X-RAY	33	Yes
1ETF	GGUCUGGGCGCAGCGCAAGCUGACGGU ACAGGCC	NMR	34	Yes
1P5N	GGCAGAAAGCGUCUAGCCAUGGCGUUA GUAUGCC	NMR	34	Non
1R2P	GAGCCGUAUGCGAUGAAAGUCGCACGU ACGGUUC	NMR	34	Non
1R7W	GGAGGACAUCCCUCACGGGUGACCGUG GUCCUCC	NMR	34	Non
1R7Z	GGAGGACAUUCCUCACGGGUGACCGUG GUCCUCC	NMR	34	Non
1RNK	GGCGCAGUGGGCUAGCGCCACUCAAAA GGCCCAU	NMR	34	Non
1T28	GGUCAUCGUUGAGAAAACGAAACAGAC GGUGGCC	NMR	34	Non
2EUY	GGCCUUAAGGAAACAGUUCGCUGUGCCG AAAGGUC	NMR	34	Non

2F88	GAGCCGUGUGCGAUGAAAGUCGCAAGC ACGGUUC	NMR	34	Non
2JTP	GGAUGGGGAAAGAAGCCCCGCAAUUUC CCCAUCC	NMR	34	Non
2KPV	GGAGGUAGUAGGUCGAAAGACCGUUCU ACACUCC	NMR	34	Non
2L3C	GGUAGUAUAACAAUAUCCGUGUUGUUA UAGUACC	NMR	34	Yes
2RVO	GGCGCUUUGACACAAUCUACAUUGUAA AAGCGCC	NMR	34	Non
400G	CAUGUCAUGUCAUGAGUCCAUGGCAUG GCAUGGC	X-RAY	34	Yes
4X40	GGCCGCGGCAGGUUCGAAUCCUGCCGC GAUCGCC	X-RAY	34	Yes
6SDY	GGCAGAAGCUGCCUCUUCGGAGGCAGU UUCUGCC	NMR	34	Yes
1ULL	GGCUGGACUCGUACUUCGGUACUGGAG AAACAGCC	NMR	35	Yes
2DRB	GGCCCCGGGCGGUUCGAUUCGCCCCUG GGCCACCA	X-RAY	35	Yes
2L3E	GGCUUUUGCUCGCCGUGCUUCGGCACG GAAAAGCC	NMR	35	Non
2M57	GGAGCCGU AUGCGGUAGUUCGCACGU ACGGAUCU	NMR	35	Non
2PCV	GGACCCGCCACUGCAGAGAUGCAAUCC AGUGGUCC	NMR	35	Non
4C4W	GAGCAAGAGCCAUUGCACUCCGGUUUG AUGACCUC	X-RAY	35	Yes
5FJ4	GAGGGAGCGCCAUUGCACUCCGGUGCG AAGAACUC	X-RAY	35	Yes
6BHJ	UACCCCCCUUCGGUGCUUUGCACCGA AGGGGGGG	X-RAY	35	Yes
1N8X	GGACUCGGCUUGCUGGAGACGGCAAGA GGCGAGUCC	X-RAY	36	Non
2FDT	GGUUGUACGUCGCUUUGGAUAAAAGCG UCUGCGACC	NMR	36	Non
2HW8	GGGGUGAAGGAGGCUUCGGCCGCGAAA CUUCACCCC	X-RAY	36	Yes
2N6S	GGAAUUUAUGAGUACCUUCGGAUACUU AUAGAUUCC	NMR	36	Non
2TPK	GCUGACCAGCUAUGAGGUCAUACAUCG UCAUAGCAC	NMR	36	Non
4X4P	GGCCGCGGCAGGUUCGAGUCCUGCCGC GAUCGCCAC	X-RAY	36	Yes
5KQE	GGGUGUACUUAACGUUUGCUUCGGCAA ACUACAUC	NMR	36	Non
6SY6	GCGGAGAAUGUUAUGGCCUUCGGGCAG AGAAAACCG	X-RAY	36	Yes
2LHP	GGGUGAAAUGGAGGACUUCGGUCCUCA AAUUUCACCC	NMR	37	Non
2LUB	GGGUGAAACGGAGGACUUCGGUCCUCA AGUUUCACCC	NMR	37	Non
6DTD	UGUUGCAUCUGCCUUCUUUUUGAAAGG UAAAAACAAC	X-RAY	37	Yes
6U79	GGUUUCUAGCACGAAGAUCUCGAUGU CUAAGAAACC	NMR	37	Non

1B36	GGUGCAGAAACAGCGCUUCGGCGCGUA UAUUAACGCACC	NMR	38	Non
1M5L	GCGCAGGACUCGGCUUCUUCGGAAGGG ACGAGGGGCGC	NMR	38	Non
1TXS	GGGCUAGCACUCUGGUAAUACGGUACC UUUGUGCGCCC	NMR	38	Non
2A9L	GGGUGACUCCAGAGGUCGAGAGACCGG AGAUAUCACCC	NMR	38	Non
2KHY	GGGAUCACAAGUAGGACUUCGGUCCGA AUACAGAUCUC	NMR	38	Non
4PDB	GGGAUGCUCAGUGAUCCUUCGGGAUUA CAGGGCAUCCC	X-RAY	38	Yes
6D12	GGGCUGCAUGUGGCAGCUCGGGCUGCA UGUGGCAGCUC	X-RAY	38	Yes
2MXL	GGAUUUGCAGGCCUACCAGAAACGGAU GGGAGUGCAGAU	NMR	39	Non
2NBY	GGGCAGAAGGUACCCCAUUGUAUGGGA UCUGAUCUGCCC	NMR	39	Non
4KR7	GCCCGGAUAGUGUCCUUGGGAAACCAA GUCCGGGCACCA	X-RAY	39	Yes
4KR9	GCCCGGAUAGUGUCCUUGGGAAACCAA GUCCGGGCACCA	X-RAY	39	Yes
2HUA	GGCCUCCAGCGACGGCCUUCGGGACUA GCAAACGGAGGCC	NMR	40	Non
2NBZ	GGGCUCCGUGCACAUGCUIUACAUGUG UUUAGUCGAGCCC	NMR	40	Non
4PMI	GGGAGUAUAUGGGCGCACUUCGGUGAC GGUACAGGCUCU	X-RAY	40	Yes
1A51	GGCCGAUGGUAGUGUGGGGUCUCCCCA UGCGAGAGUAGGCC	NMR	41	Non
1ZC5	GGCGAUCUGGCCUCCUACAAGGGAAG GCCAGGGAAUUGCC	NMR	41	Non
4M6D	GCGGCUAAAGAGUGCAGAGUUACUUAG UUCACUGCAGACGC	X-RAY	41	Yes
5V17	GGAUCAACCCAGGUGUGGCACACCAG UCAUACCUUGAUCC	NMR	41	Non
5W1H	AAGAUAGCCCAAGAAAGAGGGCAAUA CCAGAUUAUAGCCUG	X-RAY	41	Yes
6W3M	GGUUUGUUUGAAUAGAGAGCAGAUCUC UGGAAAAAUGAACC	NMR	41	Non
1MNX	GGGUGACGAUACUGUAGGCGAGAGCCU GCGGAAAAAUAGCCC	NMR	42	Non
2L2J	GGUAAGGUGGGUGGAAUCCUUCGGGAU CCCACCUACCCUGCC	NMR	42	Non
2N6T	GGGUAGAGUAUAAUUAGUCUUCGGACU UCCUUUAUCUUAUCC	NMR	42	Non
5WLH	GAAGAUAGCCCAAGAAAGAGGGCAAUA ACCAGAUUAUAGCCUG	X-RAY	42	Yes
1CQ5	GGCGUUUACCAGGUCAGGUCCGGAAGG AAGCAGCCAAGGCGCC	NMR	43	Non
1CQL	GGCGUUUACCAGGUCAGGUCCGGAAGG AAGCAGCCAAGGCGCC	NMR	43	Non
2ADT	GGGAUAUGGAAGAACCGGGAAACUUG GUUCUUCUUAAGUCCU	NMR	43	Non
2FEY	GAGACUAUCGACAUUUGAUACACUAUU UAUCAUUGGAUGUCUC	NMR	43	Non

2N6X	GGUGAGUACGUAGAGUAUACUUCGGUA UACUUAUUAUACUACC	NMR	43	Non
1A60	GGGAGCUCAACUCUCCCCCUUUUCC GAGGGUCAUCGGAACCA	NMR	44	Non
1P6V	GAUUCGACGGGGACUUCGGUCCUCGGA CGCGGGUUCGAUUCCTGC	X-RAY	45	Yes
1Z2J	GGGAAGAUCUGGCCUUCACACAAGGGA AGGCCAGGGAAUCUCCCC	NMR	45	Non
1S03	GGACGAUGGCGAAACUGCAUGAGGCAA UUCAUGCAAGUCCUCGUCC	X-RAY	47	Yes
1YMO	GGGCUGUUUUUCUCGCUGACUUUCAGC CCCAAACAAAAAAGUCAGCA	NMR	47	Non
2MTJ	GGACCUCCCGUCCUUGGACGGUCGAGC GAAAGCUUGUGAUUGGUCCG	NMR	47	Non
2PXL	GCGGGUGUUUACCAGGUCAGGUCCGAA AGGAAGCAGCCAAGGCACUUC	X-RAY	47	Yes
5KH8	GGCGAUGGUGUUCGCCAUAAACGCUCU UCGGAGCUAAUGACACCUAC	NMR	47	Non
2K95	GGGCUGUUUUUCUCGCUGACUUUCAGC CCCAAACAAAAAUGUCAGCA	NMR	48	Non
2KE6	GGCUUGAUUGUAUUUUUAAAUAUUUC UUAAAAACUACAAUUAAGCC	NMR	48	Non
2KUR	GGCUUGAUUGUAUUUAUUAAAUAUUUC UUAAUAACUACAAUUAAGCC	NMR	48	Non
2KUU	GGCUUGAUUGUAUGUGUAAAUAUUUC UUACACACUACAAUUAAGCC	NMR	48	Non
2KUV	GGCUCGAUUGUAUUUUUAAAUAUUUC UUAAAAACUACAAUUCGAGCC	NMR	48	Non
2KUW	GGGGUUGGUGUAUUUUUAAAUAUUUC UUAAAAACUACAAUCAGCUCC	NMR	48	Non
2M8K	GGUUUCUUUUUAGUGAUUUUUCCAAAC CCCUUUGUGCAAAAUCAUUA	NMR	48	Non
2VPL	GGAGUGAAGGAGGCUCGCGAACUCGCG AAGCCGAGAAACUUCACUCCC	X-RAY	48	Yes
4C70	UGUUGGUUCUCCCACAACGCGGAAGCG UGUGCCGGGAUGUAGCUGGCA	X-RAY	48	Yes
1U63	GGGAGUGAAGGAGGCUCGCGAACUCGCG GAAGCCGAGAAACUUCACUCCC	X-RAY	49	Yes
2LU0	GGAAUAUGCUCUACGAAAGUGAAUCAG CUUCGGCUGAGAGCUAAGUCC	NMR	49	Non
2PXB	GGUGCUGUUUACCAGGUCAGGUCCGAA AGGAAGCAGCCAAGGCAGUGCC	X-RAY	49	Yes
2PXD	GGGGCUGUUUACCAGGUCAGGUCCGAA AGGAAGCAGCCAAGGCAGUCC	X-RAY	49	Yes
2PXE	GGGUCUGUUUACCAGGUCAGGUCCGAA AGGAAGCAGCCAAGGCAGGUCC	X-RAY	49	Yes
2PXF	GGGGGUGUUUACCAGGUCAGGUCCGAA AGGAAGCAGCCAAGGCACUCC	X-RAY	49	Yes
2PXK	GGGUGUGUUUACCAGGUCAGGUCCGAA AGGAAGCAGCCAAGGCACGUCC	X-RAY	49	Yes
2PXP	GGGGCUGUUUACCAGGUCAGGUCCGAA AGGAAGCAGCCAAGGCAGCUCC	X-RAY	49	Yes
2PXQ	GGGCCUGUUUACCAGGUCAGGUCCGAA AGGAAGCAGCCAAGGCAGGUCC	X-RAY	49	Yes
2PXT	GCUGCUGUUUACCAGGUCAGGUCCGAA AGGAAGCAGCCAAGGCAGCGCC	X-RAY	49	Yes

2PXU	GGUGCUGUUUACCAGGUCAGGUCCGAA AGGAAGCAGCCAAGGCAGCGCC	X-RAY	49	Yes
2PXV	GGUGGUGUUUACCAGGUCAGGUCCGAA AGGAAGCAGCCAAGGCACUGCC	X-RAY	49	Yes
6MXQ	GGCAGUAUAGUCCGAACUGCAACUUCG GUUCACCUUCUCUCUAACUGCC	NMR	49	Non
6IV9	CACUGGUGCAAUUUGCACUAGUCUAA AACUCCUCGAUUACAUACACAAA	X-RAY	50	Yes
6IV8	CACUGGUGCAAUUUGCACUAGUCUAA AACUCCUCGAUUACAUACACAAAG	X-RAY	51	Yes
2MHI	GGAAACGCCGCGGUCAGCUCGGCUGCU GCGAAGAGUUCGUCUCUGUUGUUCC	NMR	53	Non
2N4L	GGAAUAUUUUUGCUGUACUUCUUAUAG UGAAUAGAGUUAGGCAGGGAUUAUCC	NMR	53	Non
1P5M	GGCUGUGAGGAACUACUGUCUUCACGC CUUCGGGAGUGUCGUGCAGCCUCCAGC C	NMR	55	Non
2HGH	GGGCCAUACCUCUUGGGCCUGGUUAGU ACCUCUUCGGUGGGAUACCAGGUGCC C	NMR	55	Yes
2KZL	GGGAGUAAAGAUUGAGACAAGUAGGAC UUCGGUCCGAUACACUCAUGAACUCC C	NMR	55	Non
2LC8	GGUCAGGGUCAGGAGCCCCCCCCUGAA CCCAGGAUAACCCUCAAGUCGGGGGG CA	NMR	56	Non
6NOA	GGCAGUAUAGUCCGAACUGCAAUCUU AUUUUCUUUCACCUUCUCUCUAACUG CC	NMR	56	Non
5IEM	GGGAUCUGUCACCCCAUUGAUCGCCUU CGGGCUGAUCUGGCUGGCUAGGCGGGU CCC	NMR	57	Non
6MCF	GGGAUCUGUCACCCCAUUGAUCGCCGA GAGGCUGAUCUGGCUGGCUAGGCGGGU CCC	NMR	57	Yes
4M4O	GGGUUCAUCAGGGCUAAAGAGUGCAGA GUUACUUAUAGUUCACUGCAGACUUGACG AACCC	X-RAY	59	Yes
6DB8	GGAUUGCCCUUGAAAAGCCUGCGAAAC ACGCAGCUGGUGAAUGACAGCUAUGGC GCAUCC	X-RAY	60	Yes
1UN6	GCCGGCCACACCUACGGGGCCUGGUUA GUACCUGGGAAACCUGGGAAUACCAGG UGCCGGC	X-RAY	61	Yes
4U7U	AUAAACCGGGCUGCCUGUCGGUUGUAA UUGAUAAUGUUGAGAGUUCGCCGCGCC AGCGGGG	X-RAY	61	Yes
2N3Q	GCAGCAGGGAACUCACGCUUGCGUAGA GGCUAAGUGCUCGGCACAGCACAAAGC CCGUCGCG	NMR	62	Non
3EGZ	GAGGGAGAGGUGAAGAAUACGACCACC UAGGUACCAUUGCACUCCGGUACCUAA AACAUACCCUC	X-RAY	65	Yes
2NC1	GGGGCUGAAGGAUGCCCAGAGAGAUUCU GGGGCCUCGGGAGAUUCGAGGUAAAAA ACGUCUAGGCCCC	NMR	67	Non

5WT1	GCGGUAGCUCAGCCUGGGAGAGCACCG GACUGAAGAUCCGGGUGUCGGGGUUC AAAUCCCCCCCCGC	X-RAY	67	Yes
2MQT	GGGCGAGGGUCUCCUCUGAGUGAUUGA CUACCCGUCAGCGGGGUCUUUCAUUU GGGGCUCGUGCCC	NMR	68	Non
2N6W	GGAAUUUAUGAGUACGUAGAGUAUAAU UAGUCUUCGGACUCCUUAUACUUAUA UACUUAUAGAUUCC	NMR	68	Non
5HR6	CCCCUUCGUCUAGAGGCCAGGACACC GCCCUUUCACGGCGGUAACAGGGGUUC GAAUCCCCUAGGGG	X-RAY	68	Yes
6U8D	GGGUCUCGCGGAACCGGUGAGUACACC GGAAUCCAGGAAACUGGAUUUGGGCGU GCCCCGCGAGACC	X-RAY	68	Yes
3EPJ	CUCGU AUGGCGCAGUGGUAGCGCAGCA GAUUGCAAUCUGUUGGUCCUAGUUC GAUCCUGAGUGCGAG	X-RAY	69	Yes
1KXK	GUCUACCUAUCGGGCUAAGGAGCCGUA UGCGAUGAAAGUCGCACGUACGGUUCU AUGCCCGGGGAAAAC	X-RAY	70	Non
2DET	GUCCCCUUCGUCUAGAGGCCAGGACA CCGCCUUUCACGGCGGUAACAGGGGU UCGAAUCCCCUAGGGG	X-RAY	70	Yes
2N8V	GGAACAGCUGUACUGGGCAGUACAGC AGUCGUAUGGUAACACAUGCGGCGUUC CGAAAUACCAUGCCUG	NMR	70	Non
5HR7	CCCCUUCGUCUAGAGGCCAGGACACC GCCCUUUCACGGCGGUAACAGGGGUUC GAAUCCCCUAGGGGAC	X-RAY	70	Yes
5V6X	GGAAACCUGAUCAUGUAGAU CGAAUGG ACUCUAAAUCCGUUCAGCCGGGUUAGA UUCCCCGGGGUUUCCGC	X-RAY	70	Yes
2DU3	GCCAGGGUGGCAGAGGGGCUUUGCGGC GGACUGCAGAUCCGCUUUACCCCGGUU CGAAUCCGGGCCCUGGC	X-RAY	71	Yes
2DU5	GCCAGGGUGGCAGAGGGGCUUUGCGGC GGACUUCAGAUCCGCUUUACCCCGGUU CGAAUCCGGGCCCUGGC	X-RAY	71	Yes
2DU6	GCCAGGGUGGCAGAGGGGCUUUGCGGC GGACUCUAGAUCGCUUUACCCCGGUU CGAAUCCGGGCCCUGGC	X-RAY	71	Yes
2L3J	GGCAUUAAGGUGGGUGGAAUAGUAUAA CAAUAUGC UAAAUGUUGUUAUAGUAUC CCACCUACCCUGAUGCC	NMR	71	Yes
2MSO	GGCUCGUUGGUCUAGGGGUAUGAUUCU CGCUUAGGGUGCGAGAGGUCCCGGGU CAAUCCCGGACGAGCC	NMR	71	Yes
2ZZN	GCCGGGGUAGUCUAGGGGCUAGGCAGC GGACUGCAGAUCCGCCUUACGUGGGU CAAUCCCACCCCCGGC	X-RAY	71	Yes
4YVI	UGGGAGGUCGUCU AACGGUAGGACGGC GGACUCUGGAUCCGCUGGUGGAGGUUC GAGUCCUCCCCUCCAG	X-RAY	71	Yes
4YVJ	UGGGAGGUCGUCU AACGGUAGGACGGC GGACUCUUGAUCCGCUGGUGGAGGUUC GAGUCCUCCCCUCCAG	X-RAY	71	Yes

4YVK	UGGGAGGUCGUCUAACGGUAGGACGGC GGACUCUCGAUCCGCUGGUGGAGGUUC GAGUCCUCCCCUCCAG	X-RAY	71	Yes
5TF6	GGUCAUUUGAAACAAUACAGAGAUGA UCAGCAGUUCCCCUGCAUAAGGAUGAA CCGUUUUACAAAGAGAC	X-RAY	71	Yes
1DRZ	GGCCGGCAUGGUCCCAGCCUCCUCGCU GGCGCCGGCUGGGCAACACCAUUGCAC UCCGGUGGCGAAUUGGGAC	X-RAY	72	Yes
1EUQ	GGGGUAUCGCCAAGCGGUAAGGCACCG GAUUCUGAUUCCGGCAGCGAGGUUCGA AUCCUCGUACCCCAGCCA	X-RAY	72	Yes
2AKE	GACCUCGUGGCGCAAUGGUAGCGCGUC UGACUCCAGAU CAGAAGGUUGCGUGUU CGAAUCACGUCGGGGUCA	X-RAY	72	Yes
2MF0	UGUCGACGGAUAGACACAGCCAUCAAG GACGAUGGUCAGGACAUCGCAGGAAGC GAUUCAUCAGGACGAUGA	NMR	72	Yes
2ZNI	GGGGGUGGAUCGAAUAGAUACACCGG ACUCUAAAUUCGUGCAGGCGGGUGAAA CUCCCGUACUCCCCGCCA	X-RAY	72	Yes
4ZT0	GAUGAGACGCGUUUUAGAGCUAGAAAU AGCAAGUUAAAAUAAGGCUAGUCCGUU AUCAACUUGAAAAAGUGU	X-RAY	72	Yes
3WC1	GCCAUCAUAGUAUAGUGGUCAUUUAAA AUCGUUGUGGCCGAUUAGACCCAAGUU CGAUUCUUGGUGAUGGCAC	X-RAY	73	Yes
3WFQ	GGCCAGGUAGCUCAGUUGGUAGAGCAC UGGACUGAAAAUCCAGGUGUCGGCGGU UCGAUUCGCCCCUGGCCA	X-RAY	73	Yes
4X0A	GGCAGGUAGCUCAGUUGGUAGAGCACU GGACUGAAAAUCCAGGUGUCGGCGGUU CGAUUCGCCCCUGCCACC	X-RAY	73	Yes
5VW1	UUGUAAAAAGUUUUAGAGCUAGAAAU AGCAAGUUAAAAUAAGGCUAGUCCGUU AUCAACUUGAAAAAGUGUC	X-RAY	73	Yes
5WT3	GGGGCGGUAGCUCAGCCUGGGAGAGCA CCGGACUGAAGAUCCGGGUGUCGGGGG UUCAAAUCCCCCCCCGCC	X-RAY	73	Yes
1GTR	GGGGUAUCGCCAAGCGGUAAGGCACCG GAUUCUGAUUCCGGCAUCCGAGGUUC GAAUCCUCGUACCCCAGCCA	X-RAY	74	Yes
2DER	GUCCCCUUCGUCUAGAGGCCCAGGACA CCGCCUUUCACGGCGGUAACAGGGGU UCGAAUCCCCUAGGGGACGC	X-RAY	74	Yes
2ZM5	GCCCGGAUAGCUCAGUCGGUAGAGCAG GGGAUUGAAAAUCCCCGUGUCCUUGGU UCGAUUCGAGUCCGGGCAC	X-RAY	74	Yes
3AKZ	UGGGAGGUCGUCUAACGGUAGGACGGC GGACUCUGGAUCCGCUGGUGGAGGUUC GAGUCCUCCCCUCCAGCCA	X-RAY	74	Yes
3FOZ	GCCCGGAUAGCUCAGUCGGUAGAGCAG GGGAUUGAAAAUCCCCGUGUCCUUGGU UCGAUUCGAGUCCGGGCAC	X-RAY	74	Yes
3TUP	GCCGAGGUAGCUCAGUUGGUAGAGCAU GCGACUGAAAAUCGCAGUGUCGGCGGU UCGAUUCGUCUCCUCGGCAC	X-RAY	74	Yes

3WC2	GCGGAUUUAGCUCAGUUGGGAGAGCGC CAGACUGUGGAUCUGGAGGUCCUGUGU UCGAUCCACAGAAUUCGCAC	X-RAY	74	Yes
3WFS	GGCCAGGUAGCUCAGUUGGUAGAGCAC UGGACUGAAAAUCCAGGUGUCGGCGGU UCGAUUCGCCCCUGGCCAC	X-RAY	74	Yes
4YCO	GCGCGGAUAGCUCAGUCGGUAGAGCAG GGGAUUGAAAAUCCCCGUGUCCUUGGU UCGAUUCGAGUCCGCGCAC	X-RAY	74	Yes
4YYE	GUUAUAAUAGCUAAAUUGGUAGAGCAU UCGUUUUGUAAUCGAAAGGUUUGGGGU UCAAAUCCCUAAUAUAACAC	X-RAY	74	Yes
5D6G	GCCUAAGACAGCGGGGAGGUUGGCUUA GAAGCAGCCAUCUUAAGAGUGCGU AACAGCUCACCCGUCGAGGC	X-RAY	74	Yes
1FFY	GGGCUUGUAGCUCAGGUGGUUAGAGCG CACCCUGAUAAAGGGUGAGGUCGGUGG UUCAAGUCCACUCAGGCCAC	X-RAY	75	Yes
1N77	GGCCCCAUCGUCUAGCGGUUAGGACGC GGCCCUCUCAAGGCCGAAACGGGGGUU CGAUUCCCCUGGGGUCACCA	NMR	75	Yes
2DR2	GACCUCGUGGCGCAAUGGUAGCGCGUC UGACUCCAGAU CAGAAGGUUGCGUGUU CGAAUCACGUCGGGGUCACCA	X-RAY	75	Yes
2IHX	CUGCCCUCAUCCGUCUCGCUAAUUCGG GGAGCGGACGAUGACCCUAGUAGAGGG GGCUGCGGCUUAGGAGGGCAG	NMR	75	Yes
2ZUE	GGACCGGUAGCCUAGCCAGGACAGGGC GGCGGCCUCCUAAAGCCGAGGUCCGGG GUUCAAAUCCCCGCCGGUCCG	X-RAY	75	Yes
3WQY	GGGCUCGUAGCUCAGCGGGAGAGCGCC GCCUUUGCGAGGCGGAGGCCGCGGGUU CAAUCCCCGCCGAGUCCACCA	X-RAY	75	Yes
3WQZ	GGACUCGUAGCUCAGCGGGAGAGCGCC GCCUUUGCGAGGCGGAGGCCGCGGGUU CAAUCCCCGCCGAGUCCACCA	X-RAY	75	Yes
4TZV	GAGUAGUUCAGUGGUAGAACCACCUU UGCCAAGGUGGGGGUCGCGGGUUCGAA UCCCGUCUCGGGCGAAAGCCC	X-RAY	75	Yes
4WC2	GGCCAGGUAGCUCAGUUGGUAGAGCAC UGGACUGAAAAUCCAGGUGUCGGCGGU UCGAUUCGCCCCUGGCCACC	X-RAY	75	Yes
5X6B	GCCGGGUAGUCUAGGGGCUAGGCAGC GGACUGCAGAUCCGCCUACGUGGGUU CAAUCCCCACCCCGGCUCCA	X-RAY	75	Yes
1EIY	GCCGAGGUAGCUCAGUUGGUAGAGCAU GCGACUGAAAAUCGCAGUGUCCGCGGU UCGAUUCGCGCCUCGGCACCA	X-RAY	76	Yes
2K4C	GGGUGAUUAGCUCAGCUGGGAGAGCAC CUCCCUUACAAGGAGGGGGUCGGCGGU UCGAUCCCGUCAUCACCCACCA	NMR	76	Non
4WC3	GGCCAGGUAGCUCAGUUGGUAGAGCAC UGGACUGAAAAUCCAGGUGUCGGCGGU UCGAUUCGCCCCUGGCCACCA	X-RAY	76	Yes
4WJ3	UCCGCGAUAGCUCAGUCGGUAGAGCAA AUGACUGUAAUCAUUGGGUCCCUUGGU UCGAGUCCAGGUCGCGGAGCCA	X-RAY	76	Yes

1P5P	GGCUGUGAGGAACUACUGUCUUCACGC AGAAAGCGUCUAGCCAUGGCGUUAGUA UGAGUGUCGUGCAGCCUCCAGCC	NMR	77	Non
3A2K	GGACCUUUAGCUCAGUUGGUUAGAGCA GACGGUCUAUAACCGUCCGGCCGUAGG UUCGAGUCCUACAAGGUCCACCA	X-RAY	77	Yes
4X0B	GGGCCAGGUAGCUCAGUUGGUAGAGCA CUGGACUGAAAAUCCAGGUGUCGGCGG UUCGAUUCGCCCCUGGCCACC	X-RAY	77	Yes
5CCB	GGCCCCGAUAGCUCAGUCGGUAGAGCA UCAGACUUUUAUCUGAGGGUCCAGGG UUCAAGUCCUGUUCGGGCGCCA	X-RAY	77	Yes
3AMT	GGGCCCGUAGCUUAGCCAGGUCAGAGC GCCCGCUCUAUAACCGGGCGGUCGAGG GUUCGAAUCCUCCGGGCCACCA	X-RAY	78	Yes
3U4M	GGGAUGCGUAGGAUAGGUGGGAGCCUG UGAACCCCGCCUCCGGGUGGGGGGA GGCGCCGGUGAAAUACCACCCUCC	X-RAY	80	Yes
6B14	GACGCGACCGAAAUGGUGAAGGACGGG UCCAGUGCGAAACACGCACUGUUGAGU AGAGUGUGAGCUCCGUAACUGGUCGCG UC	X-RAY	83	Yes
6B3K	GACGCGACCGAAAUGGUGAAGGACGGG UCCAGUGCGAGACCCGCACUGUUGAGU AGAGUGUGAGCUCCGUAACUGGUCGCG UC	X-RAY	83	Yes
2ZZM	GCAGGGGUCGCCAAGCCUGGCCAAAGG CGCUGGGCCUAGGACCCAGUCCCGUAG GGGUUCCAGGGUUCAAAUCCUGCCCC UGC	X-RAY	84	Yes
3A3A	GCCCGGAUGAUCCUCAGUGGUCUGGGG UGCAGGCUUCAACCUGUAGCUGUCUA GCGACAGAGUGGUUCAAUCCACCUUU CGGGC	X-RAY	86	Non
3K0J	GCGACUCGGGGUGCCCUCAUUGCACU CCGGAGGUCGAGAAAUACCCGUUAUCAC CUGAUCUGGAUAAUGCCAGCGUAGGGA AGUCGC	X-RAY	87	Yes
1WZ2	GCGGGGGUUGCCGAGCCUGGUCAAAGG CGGGGACUCAAGAUCCCCUCCCGUAG GGGUUCCGGGGUUCGAAUCCCCGCCCC CGCACCA	X-RAY	88	Yes
5XBL	UGCUCUUGGCGUUUAGAGCUAGAAAU AGCAAGUUAAAAUAAGGCUAGUCCGUU AUCAACUUGAAAAAGUGGCACCGAGUC GGUGCUU	X-RAY	88	Yes
2N7M	GGCCUUAUGCACGGGAAAUACGCAUUA CAGUGAGGAUUCGUCCGAGAUUGUGUU UUUGCUGGUGUAAAUCAGCAGUUCCCC UGCAUAAGGCU	NMR	92	Non
3ADB	GGCCGCCGCCACCGGGUGGUCCCCGG GCCGGACUUCAGAUCCGGCGCGCCCCG AGUGGGGCGCGGGGUUCAAUCCCCGC GGCGGCCGCA	X-RAY	92	Yes
3W1K	GGGAGUAGAUAGGCGCUGGUGUGCCUC CUAGACUUCAAAUCUACGGUCUCGCUA	X-RAY	92	Yes

	UUUAAGCGAGAGGUGGGUUCGAUCCCA ACAUCUCCCG			
2V3C	GCGGUGGGGGAGCAUCUCCUGUAGGG GAGAUGUAACCCCUUUACCUGCCGAA CCCCGCCAGGCCCGGAAGGGAGCAACG GUAGGCAGGACGUCG	X-RAY	96	Yes
3KTW	AGAUAGUCGUGGGUCCCUUUCUGGAG GGAGAGGGAAUCCACGUUGACCGGGG GAACCGGCCAGGCCCGGAAGGGAGCAA CCGUGCCCGGCUAUC	X-RAY	96	Yes
1LNG	UCGGCGGUGGGGGAGCAUCUCCUGUAG GGGAGAUGUAACCCCUUUACCUGCCG AACCCCGCCAGGCCCGGAAGGGAGCAA CGGUAGGCAGGACGUC	X-RAY	97	Yes
6JXM	GCGGAAGUAGUUCAGUGGUAGAACACC CGACAGACGAAGCGCUAAAACGUGGGA UUCUGUCGGGGGUCGCGGGUUCGAGU CCCGUCUCCGCUCCA	X-RAY	97	Non
3W3S	GGGAGAGGUUGGCCGGCUGGUGCCGCC CCGGGACUUCAAUCCCGUGGGAGGUC CCGCAAGGGAGCUCGAGGGUUCGAU UCCCUCCUCUCCCGCC	X-RAY	98	Yes
1S9S	GGCGGUACUAGUUGAGAAACUAGCUCU GUAUCUGGCGGACCCGUGGUGGAACUG UGAAGUUCGGAACCCCGGCCCAACC CUGGGAGAGGUCCAGGGUU	NMR	101	Non
6MJ0	AAGUUCUGAUCUUUAAAAUCGUUAGC UCGCCAGUUAGCGAGGUCUGCGAAAGC AGAUAAUCGGGUGCAACUCCCGCCCUU UCUCCGAGGGUCAUCGGAAC	X-RAY	101	Non
2KRL	GGACGGUGGCAGCACUGUCUAGCUGCG GGCAUUAGACUGGAAAACUAGUGCUCU UUGGGUAACCACUAAAUCCCGAAAGG GUGGGCUGUGGUGACCCUCCG	NMR	102	Non
2XXA	GCAUUGCUGGUGCAGCGCAGCGCGGAC GCCCGAACCUGGUCAGAGCCGGAAGGC AGCAGCCAUAAGGGAUGCUUUGCGGGU GCCGUUGCCUUCGCGCAAUGC	X-RAY	102	Yes
7K1Z	GGCUCUGGUAACUAGAGAUCCCUAGCA CCCUUUUAGUCAGUGUGGAAAUCUCU AGCAGUGGCGCCCGAACAGGGACUUGA AAGCGAAAGUAAAGCCAGAGCC	NMR	103	Non
2NBX	GGGGCUGAAGGAUGCCAGAAGGUACC CCAUUGUAUGGGAUUCUGAUCUGGGGCC UCGGUGCACAUGC UUACAUGUGUUUA GUCGAGGUUAAAAACGUCUAGGCCCC	NMR	108	Non
2LKR	GGCAAUACAGAGAUGAUCAGCAGUUC CCUGCAUAAGGAUGAACCGUUUUACAA AGAGAUUUUCUUCGGGAUCUCUUUGCC UUUUGGCUUAGAUCAAGUGUAGUAUCU GUC	NMR	111	Non
4P3E	GACUAAGUUCGGCAUCAUAUUGGUGAC CUCCCGGAGCGGGGGACCACCAGGUU GCCUAAGGAGGGGUGAACCGGCCAGG UCGGAAACGGAGCAGGUCAAACUCC GUGCUGAUCAGUAGUU	X-RAY	124	Yes

3IVK	UCCAGUAGGAACACUAUACUACUGGAU AAUCAAGACAAAUCUGCCCCGAAGGGC UUGAGAACAUCGAAACACGAUGCAGAG GUGGCAGCCUCCGGUGGGUUAAAACCC AACGUUCUCAACAAUAGUGA	X-RAY	128	Yes
3NDB	GUCUCGUCCCCGUGGGGCUCGGCGGUGG GGGAGCAUCUCCUGUAGGGGAGAUGUA ACCCCUUUACCUGCCGAACCCCGCCA GGCCCGGAAGGGAGCAACGGUAGGCAG GACGUCGGCGUCACGGGGGUGCGGGA C	X-RAY	136	Yes
2N1Q	GGUGCCCGUCUGUUGUGUGACUCUGGU GAGAGCCAGAGGAGAUCUCUCGACGCA GGACUCGGCUUGCUGGAGACGGCAAGA GGCGAGGGGCGGCGACUGGUGAGUACG CCAAAAUUUUGACUAGCGGAGGCUAG AAGGAGAGAGAUGGGUGCCC	NMR	155	Non
2R8S	GGAAUUGCGGGAAAGGGGUCAACAGCC GUUCAGUACCAAGUCUCAGGGGAAACU UUGAGAUGGCCUUGCAAAGGGUAUGGU AAUAAGCUGACGGACAUGGUCCUAACA CGCAGCCAAGUCCUAAGUCAACAGAUC UUCUGUUGAUUUGGAUGCAGUUA	X-RAY	159	Yes
4P8Z	GGUUGGGUUGGGAAGUAUCAUGGCUAA UCACCAUGAUGCAAUCGGGUUGAACAC UUAAUUGGGUUAAAACGGUGGGGGACG AUCCCGUAACAUCGUCUUAACGGCGA CAGACUGCACGGCCCUGCCUCUAGGU GUGUCCAUGAACAGUCGUUCCGAAAG GAAGCAUCCGGUAUCCCAAGACAAUC	X-RAY	188	Non
1GRZ	GACCGUCAAAUUGCGGGAAAGGGGUCA ACAGCCGUUCAGUACCAAGUCUCAGGG GAAACUUUGAGAUGGCCUUGCAAAGGG UAUGGUAAUAAGCUGACGGACAUGGUC CUAACCACGCAGCCAAGUCCUAAGUCA ACAGGAGACUGUUGAUUUGGAUGCAGU UCACAGACUAAAUGUCGGUCGGGGAAG AUGUAUUCUUCUCAUAAGAUUAGUCG GACCUCUCCCGAAAGGGAGUUGGAGUA CUCG	X-RAY	247	Non
5IWA	AAAUUGGAGAGUUUGAUCCUGGCUCAG GGUGAACGCUGGCGGCGUGCCUAAGAC AUGCAAGUCGUGCGGGCCGCGGGGUUU UACUCCGUGGUCAGCGGCGGACGGGUG AGUAACGCGUGGGUGACCUACCCGGAA GAGGGGACAACCCGGGAAACUCGGG CUAAUCCCCCAUGUGGACCCGCCCUU GGGGUGUGUCCAAGGGCUUUGCCCGC UUCGGAUGGGCCCGCUGCCAUACAGC UAGUUGGUGGGGUAAUGGCCACCAAG GCGACGACGGGUAGCCGGUCUGAGAGG AUGGCCGGCCACAGGGGCACUGAGACA CGGGCCCCACUCCUACGGGAGGCAGCA GUUAGGAAUCUCCGCAAUGGGCGCAA GCCUGACGGAGCGACGCCGCUUGGAGG AAGAAGCCCUUCGGGGUGUAAACUCCU GAACCCGGGACGAAACCCCGACGAGG GGACUGACGGUACCGGGGUAAUAGCGC	X-RAY	1509	Yes

<p> CGGCCAACUCCGUGCCAGCAGCCGCGG  UAAUACGGAGGGCGCGAGCGUUACCCG  GAUUCACUGGGCGUAAAGGGCGUGUAG  GCGGCCUGGGCGUCCCAUGUGAAAGA  CCACGGCUCAACCGUGGGGGAGCGUGG  GAUACGCUCAGGCUAGACGGUGGGAGA  GGGUGGUGGAAUUCGGGAGUAGCGGU  GAAAU GCGCAGAUACCGGGAGGAACGC  CGAUGGCGAAGGCAGCCACCUGGUCCA  CCCUGACGCUGAGGCGCGAAAGCGUG  GGGAGCAAACCGGAUUAGAUACCGGG  UAGUCCACGCCCUAAACGAU GCGCGCU  AGGUCUCUGGGUCUCCUGGGGGCCGAA  GCUAACGCGUUAAGCGCGCCCGUGGG  GAGUACGGCCGCAAGGCUGAAACUCAA  AGGAAUUGACGGGGGCCCGCACAAGCG  GUGGAGCAUGUGGUUUAAUUCGAAGCA  ACGCGAAGAACCUUACCAGGCCUUGAC  AUGCUAGGGAACCCGGGUGAAAGCCUG  GGGUGCCCCGCGAGGGGAGCCCUAGCA  CAGGUGCUGCAUGGCCGUCGUCAGCUC  GUGCCGUGAGGUGUUGGGUUAAGUCC  GCAACGAGCGCAACCCCCGCGUUAAGU  UGCCAGCGGUUCGGCCGGGCACUCUAA  CGGGACUGCCCGCGAAAGCGGGAGGAA  GGAGGGGACGACGUCUGGUCAGCAUGG  CCCUUACGGCCUGGGCGACACACGUGC  UACAAUGCCCACUACAAAGCGAUGCCA  CCCGGCAACGGGGAGCUAAUUCGCAAAA  AGGUGGGCCCAGUUCGGAUUGGGGUCU  GCAACCCGACCCCAUGAAGCCGGAAUC  GCUAGUAAUCGCGGAUCAGCCAUGCCG  CGGUGAAUACGUUCCCGGGCCUUGUAC  ACACCGCCCGUCACGCCAUGGGAGCGG  GCUCUACCCGAAGUCGCCGGGAGCCUA  CGGGCAGGCGCCGAGGGUAGGGCCCGU  GACUGGGGCGAAGUCGUAACAAGGUAG  CUGUACCGGAAGGUGCGGCUGGAU </p>			
---	--	--	--

Afin de réduire l’empreinte carbone, les tableaux annexes volumineux associés à la prédiction de structure secondaire ont été déposés sur GitHub.

Annexe 3. Disponible sur Github<sup>7</sup>. Valeur  $Apta_D$  des prédictions du jeu de données ADN associées à RNAfold ou mfold avec différents modèles thermodynamiques pour chaque structure PDB. L’absence de prédiction est marquée d’un « / ».

Annexe 4. Disponible sur Github<sup>7</sup>. Valeur  $Apta_D$  des prédictions du jeu de données ARN associées à RNAfold ou mfold avec différents modèles thermodynamiques pour chaque structure PDB. L’absence de prédiction est marquée d’un « / ».

Annexe 5. Disponible sur Github<sup>7</sup>. Valeur  $Apta_D$  des prédictions du jeu de données ADN associées à CONTRAfold, CentroidFold, Linearfold-C, Linearfold-V, MC-Fold (défaut ou « pseudoknotted »), MXfold2, Ufold et SPOT-RNA pour chaque structure PDB. L’absence de prédiction est marquée d’un « / ».

Annexe 6. Disponible sur Github<sup>7</sup>. Valeur  $Apta_D$  des prédictions du jeu de données ADN associées à CONTRAfold, CentroidFold, Linearfold-C, Linearfold-V, MC-Fold (défaut ou « pseudoknotted »), MXfold2, Ufold et SPOT-RNA pour chaque structure PDB. L’absence de prédiction est marquée d’un « / ».

---

<sup>7</sup> [github.com/GEC-git/Annexes\\_Predictions](https://github.com/GEC-git/Annexes_Predictions)

Annexe 7. Mesure du tanimoto modifié obtenue pour les 27 structures G-quadruplexes à partir des prédictions de RNAfold ADN ou ARN incluant le paramètre « *gquad* »

<b>PDB</b>	<b>RNAfold ADN</b>	<b>RNAfold ARN</b>
<b>148D</b>	0,467	0,467
<b>1HAO</b>	0,467	0,467
<b>2N21</b>	1,000	1,000
<b>5NYS</b>	1,000	1,000
<b>1I34</b>	0,600	0,600
<b>2KF8</b>	0,818	0,818
<b>1OZ8</b>	0,667	0,667
<b>5VHE</b>	0,833	0,833
<b>2HY9</b>	1,000	1,000
<b>7CV4</b>	0,462	0,923
<b>6EVV</b>	0,615	0,615
<b>4I7Y</b>	0,593	0,593
<b>7CV3</b>	0,741	0,741
<b>2M8Z</b>	0,704	0,704
<b>6H1K</b>	0,714	0,393
<b>2M91</b>	0,733	0,733
<b>5CMX</b>	0,900	0,967
<b>2M93</b>	0,563	0,563
<b>6SUU</b>	0,813	0,813
<b>6T2G</b>	0,813	0,813
<b>2M90</b>	0,875	0,875
<b>7CLS</b>	0,576	0,545
<b>5MTA</b>	0,941	0,382
<b>2M92</b>	0,588	0,206
<b>6ZL9</b>	1,000	1,000
<b>6ZL2</b>	1,000	1,000
<b>6ZTE</b>	0,889	0,889



<b>2LO5 6FKE 3WPH 1LA8 1POU 5F55 2EXF 2JZW 6FK5 2M8Y 1UUT 1ACT 6FK4 1XUE 1EN1</b>																
	<i>RNAComposer</i>					<i>SimRNA</i>					<i>RNAAdenovo</i>					
Exp <sup>a</sup>	3,94	3,41	3,21	3,24	4,01	9,30	3,98	6,19	4,18	4,16	3,44	3,98	3,40	3,94	6,62	
Mfold <sup>b</sup>	4,14	3,62	2,82	/	/	7,06	/	/	/	/	/	/	/	/	14,02	6,87
Exp <sup>a</sup>	2,71	1,55	4,09	3,47	2,58	11,97	3,23	5,30	2,28	2,59	2,89	3,76	2,77	4,74	8,63	
Mfold <sup>b</sup>	3,48	1,55	3,06	/	/	11,40	/	/	/	/	/	/	/	13,03	8,51	
3,02	1,68	1,84	3,50	2,97	11,25	3,62	3,98	2,74	2,81	3,08	4,34	2,82	3,11	8,39		
4,11	1,38	2,83	3,53	3,10	12,52	2,50	5,21	1,64	3,79	3,12	3,76	2,27	4,04	8,60		
3,24	1,69	2,39	3,66	3,09	10,70	2,51	4,42	2,65	2,81	3,53	4,09	2,93	3,78	8,49		
3,12	1,59	4,49	3,28	2,91	11,99	2,46	4,31	2,54	2,79	3,22	3,53	2,43	3,89	8,65		
2,96	1,77	1,92	3,66	2,35	12,16	3,24	5,16	2,59	2,81	3,17	3,60	2,41	4,63	8,74		
4,07	1,37	1,89	3,38	2,89	10,60	3,56	4,86	2,54	2,77	3,45	4,06	2,47	3,57	9,25		
2,71	1,90	2,35	3,30	2,95	13,07	3,04	4,32	2,65	2,77	3,19	3,46	2,71	3,73	8,25		
3,53	1,59	3,92	3,33	2,88	11,23	3,82	4,94	2,57	3,27	3,76	4,00	2,80	3,91	8,86		
2,87	1,74	2,14	3,44	2,84	10,66	2,41	4,25	2,61	3,06	3,23	3,31	2,77	4,08	8,87		
3,20	1,79	2,46	3,51	2,98	13,17	2,52	4,86	2,53	3,15	3,45	3,67	2,19	4,04	8,57		
5,83	2,47	4,30	/	/	14,41	/	/	/	/	/	/	/	2,80	8,78		
3,20	1,43	13,52	/	/	16,39	/	/	/	/	/	/	/	4,51	10,32		
2,97	1,40	3,12	/	/	14,26	/	/	/	/	/	/	/	3,97	10,45		
3,33	1,57	3,60	/	/	14,25	/	/	/	/	/	/	/	4,26	9,71		
2,80	3,35	4,27	/	/	14,46	/	/	/	/	/	/	/	3,12	9,26		
3,41	2,51	4,15	/	/	13,72	/	/	/	/	/	/	/	4,03	9,18		
3,16	1,33	15,27	/	/	14,12	/	/	/	/	/	/	/	4,58	9,48		
2,77	1,41	2,67	/	/	14,34	/	/	/	/	/	/	/	3,79	9,79		
16,92	6,01	3,59	/	/	14,13	/	/	/	/	/	/	/	3,58	9,87		
11,10	1,33	3,40	/	/	14,31	/	/	/	/	/	/	/	3,83	9,11		

		<b>4KB0 4KB1 1ECU 3Q0A 4FF1 3C46 3Q23 3Q24 2A60 2L5K 3DSD 2VHG 1OSB 1ZM5 2CDM</b>														
		<b>RNAComposer</b>														
	Exp <sup>a</sup>	3,93	3,79	2,81	6,59	6,42	6,53	6,44	6,42	4,70	7,37	7,63	4,12	7,56	7,86	10,60
	Mfold <sup>b</sup>	3,05	3,00	4,14	/	/	7,87	7,73	7,23	/	12,70	/	/	9,02	9,03	8,01
		<b>SimRNA</b>														
	Exp <sup>a</sup>	3,21	3,00	3,23	12,34	12,56	12,31	12,06	14,22	4,23	5,92	7,96	3,51	8,15	8,53	10,19
	Mfold <sup>b</sup>	3,51	3,52	2,56	/	/	11,94	12,16	14,03	/	5,38	/	/	7,99	8,42	9,64
		<b>RNAAdenovo</b>														
	Exp <sup>a</sup>	3,38	3,64	2,24	9,71	9,91	12,70	12,41	8,84	4,11	4,16	7,58	6,12	7,80	12,94	10,14
	Mfold <sup>b</sup>	3,50	3,17	2,22	9,04	8,93	8,43	12,19	10,39	3,67	4,53	7,65	5,51	8,98	12,35	10,32
	Exp <sup>a</sup>	4,16	3,43	2,22	8,63	12,14	12,50	12,34	9,29	4,30	4,07	6,40	6,39	9,21	8,54	9,11
	Mfold <sup>b</sup>	3,59	3,73	2,51	8,92	12,09	11,38	11,69	9,92	3,96	4,34	8,10	5,75	12,60	9,07	12,92
	Exp <sup>a</sup>	3,86	3,87	2,32	9,13	9,91	8,39	7,86	9,38	3,92	3,95	7,57	5,77	8,43	8,01	11,16
	Mfold <sup>b</sup>	3,77	3,44	2,27	9,48	9,74	8,20	8,01	9,78	4,37	4,79	7,88	6,67	10,95	10,60	9,67
	Exp <sup>a</sup>	3,70	3,70	2,63	9,59	9,02	12,15	8,20	10,16	4,18	4,49	8,54	5,07	10,94	8,62	10,12
	Mfold <sup>b</sup>	3,56	3,59	2,26	9,13	9,53	7,41	8,10	11,02	4,09	3,74	8,21	5,92	11,64	9,54	13,41
	Exp <sup>a</sup>	3,95	3,11	2,42	9,98	9,02	8,77	8,13	13,94	4,38	4,27	8,08	6,07	8,59	8,77	9,27
	Mfold <sup>b</sup>	3,83	3,76	2,51	9,31	8,72	8,39	8,29	10,21	4,59	4,21	8,42	5,59	9,60	11,14	9,65
	Exp <sup>a</sup>	3,59	3,24	3,00	/	/	11,77	7,22	9,65	/	4,58	/	/	11,37	9,05	9,89
	Mfold <sup>b</sup>	4,35	3,15	2,36	/	/	12,15	7,88	10,00	/	4,37	/	/	10,67	8,85	10,06
	Exp <sup>a</sup>	3,17	3,25	2,91	/	/	12,33	8,28	9,64	/	4,51	/	/	12,08	14,18	10,33
	Mfold <sup>b</sup>	2,99	3,11	2,19	/	/	11,52	8,15	9,70	/	4,87	/	/	11,67	11,44	9,00
	Exp <sup>a</sup>	3,31	3,15	2,41	/	/	12,52	8,20	9,61	/	5,03	/	/	8,39	9,93	9,97
	Mfold <sup>b</sup>	3,10	3,12	2,63	/	/	7,89	7,94	9,93	/	4,33	/	/	9,47	11,68	9,77
	Exp <sup>a</sup>	5,97	3,19	3,10	/	/	9,42	9,01	10,01	/	4,37	/	/	8,69	12,07	10,03
	Mfold <sup>b</sup>	2,95	3,10	2,47	/	/	7,32	8,03	9,89	/	4,30	/	/	9,68	12,01	10,38
	Exp <sup>a</sup>	3,64	3,15	2,76	/	/	8,12	8,06	8,79	/	4,72	/	/	10,34	11,01	9,51
	Mfold <sup>b</sup>	2,91	3,29	2,68	/	/	12,42	7,70	8,85	/	4,07	/	/	10,97	8,81	10,13

		<b>2VIC 5N2Q 1NGU 1NGO 1JVE 3ZH2 4HT4 1YTB 1B4Y 4ER8 6SEI 5HRU 4F41 4F43 5HTO</b>														
		<i>RNAComposer</i>														
Exp <sup>a</sup>	Mfold <sup>b</sup>	5,91	13,55	7,77	5,24	5,76	9,71	13,00	4,22	15,67	7,98	6,05	5,78	7,48	7,71	5,30
		<i>SimRNA</i>														
Exp <sup>a</sup>	Mfold <sup>b</sup>	4,82	14,58	8,88	6,08	4,40	11,69	12,43	5,31	3,93	7,48	5,80	6,13	6,51	6,42	4,39
		<i>RNAAdenovo</i>														
Exp <sup>a</sup>	Mfold <sup>b</sup>	3,85	/	/	/	/	14,33	/	/	/	6,65	/	13,04	6,31	5,93	5,11
Exp <sup>a</sup>	Mfold <sup>b</sup>	5,09	14,13	9,17	5,83	5,63	7,82	12,21	4,05	8,78	8,33	6,11	6,46	6,45	6,66	5,04
Exp <sup>a</sup>	Mfold <sup>b</sup>	4,00	12,32	9,41	5,72	4,86	8,72	12,75	4,23	8,25	9,18	5,20	5,02	6,99	6,18	7,57
Exp <sup>a</sup>	Mfold <sup>b</sup>	4,09	13,62	9,59	5,90	4,69	6,63	13,04	4,28	11,52	9,21	5,16	8,33	6,38	6,61	7,28
Exp <sup>a</sup>	Mfold <sup>b</sup>	4,19	13,44	9,15	5,70	4,68	10,27	12,45	4,65	10,11	10,13	5,56	5,56	6,38	6,59	6,51
Exp <sup>a</sup>	Mfold <sup>b</sup>	4,89	13,58	9,16	5,77	4,77	7,70	13,16	4,41	10,35	8,96	5,74	9,37	6,38	6,60	6,08
Exp <sup>a</sup>	Mfold <sup>b</sup>	3,87	10,69	10,09	5,72	4,80	9,42	14,22	4,28	8,75	9,15	5,00	8,21	6,38	6,58	7,43
Exp <sup>a</sup>	Mfold <sup>b</sup>	3,91	12,76	9,36	6,13	4,84	7,82	13,96	4,70	10,20	9,57	5,39	6,08	6,38	6,59	6,72
Exp <sup>a</sup>	Mfold <sup>b</sup>	3,83	13,35	9,42	5,77	4,82	8,60	14,38	4,31	11,29	8,33	5,86	5,74	6,38	6,58	7,64
Exp <sup>a</sup>	Mfold <sup>b</sup>	4,01	10,56	9,53	5,65	4,81	7,56	13,14	4,90	9,48	7,90	5,02	7,09	6,38	6,58	9,58
Exp <sup>a</sup>	Mfold <sup>b</sup>	4,41	10,74	10,08	5,82	4,67	8,04	14,04	4,77	11,66	8,41	5,36	7,18	6,39	6,58	6,11
Exp <sup>a</sup>	Mfold <sup>b</sup>	3,91	/	/	/	/	14,90	/	/	/	9,53	/	12,70	6,40	6,02	7,83
Exp <sup>a</sup>	Mfold <sup>b</sup>	3,30	/	/	/	/	14,46	/	/	/	9,72	/	13,01	6,43	6,57	10,05
Exp <sup>a</sup>	Mfold <sup>b</sup>	4,09	/	/	/	/	14,81	/	/	/	10,11	/	13,08	6,41	6,02	9,36
Exp <sup>a</sup>	Mfold <sup>b</sup>	3,57	/	/	/	/	14,99	/	/	/	10,03	/	13,12	6,48	6,23	7,76
Exp <sup>a</sup>	Mfold <sup>b</sup>	3,67	/	/	/	/	15,24	/	/	/	7,57	/	11,55	6,37	6,16	10,40
Exp <sup>a</sup>	Mfold <sup>b</sup>	3,07	/	/	/	/	15,61	/	/	/	8,82	/	12,84	6,48	6,15	10,03
Exp <sup>a</sup>	Mfold <sup>b</sup>	3,71	/	/	/	/	15,07	/	/	/	8,81	/	12,66	6,44	5,98	7,96
Exp <sup>a</sup>	Mfold <sup>b</sup>	3,18	/	/	/	/	14,76	/	/	/	10,99	/	12,61	6,45	6,92	12,34
Exp <sup>a</sup>	Mfold <sup>b</sup>	4,96	/	/	/	/	15,31	/	/	/	10,14	/	11,64	6,61	6,23	10,71
Exp <sup>a</sup>	Mfold <sup>b</sup>	4,35	/	/	/	/	14,57	/	/	/	10,19	/	12,84	6,48	6,55	11,07

<b>2VJU 1EZN 1SNJ 6U82 3HXO 2N8A 3THW</b>										
<i>RNAComposer</i>			<i>SimRNA</i>			<i>RNAde novo</i>				
Exp <sup>a</sup>	Mfold <sup>b</sup>		Exp <sup>a</sup>	Mfold <sup>b</sup>		Exp <sup>a</sup>	Mfold <sup>b</sup>			
8,72	12,20	10,16	5,74	9,40	11,78	11,44				
11,90	12,29	10,00	/	18,64	7,19	/				
6,06	11,07	7,29	7,10	7,04	13,05	7,43				
9,95	10,72	7,10	/	23,62	13,91	/				
5,60	12,90	15,22	6,76	12,27	15,13	10,32				
7,34	12,96	13,06	6,09	9,29	15,88	12,24				
5,32	12,05	12,60	6,56	10,37	11,70	11,74				
6,33	12,98	9,85	6,44	9,83	10,18	10,82				
6,45	12,52	11,81	6,82	10,12	10,57	11,49				
6,75	13,03	9,13	7,31	9,66	14,64	10,19				
7,01	13,02	10,77	6,19	11,10	16,09	10,24				
6,52	14,14	12,30	5,81	10,85	15,65	9,40				
6,99	12,95	13,70	6,43	7,05	10,20	9,55				
8,01	12,85	12,42	6,22	10,64	10,66	11,32				
6,20	12,61	11,60	/	16,63	14,20	/				
10,01	13,24	13,07	/	15,04	13,39	/				
10,61	12,72	13,18	/	19,61	15,92	/				
9,93	12,15	14,45	/	18,26	15,97	/				
8,70	12,82	13,04	/	19,72	15,97	/				
10,52	11,58	10,05	/	16,66	9,90	/				
10,13	12,05	12,35	/	18,14	16,36	/				
10,47	13,35	13,48	/	17,33	15,01	/				
9,68	14,23	11,49	/	18,83	14,60	/				
11,27	9,67	13,66	/	21,75	16,34	/				

Annexe 9. Résultats de TTclust pour les dynamiques sur structures expérimentales en TIP3P et système neutralisé non concentré en NaCl des 5 structures étudiées. Les clusters sont affichés dans l'ordre de leur représentativité dans l'aMD ainsi que la RMSD de la structure représentative face à la structure expérimentale.

Structure	aMD n°		Cluster 1	Cluster 2	Cluster 3	Cluster 4
1EZN	1	Etats occupés (%) <sup>a</sup>	45,18	27,69	17,09	10,04
		RMSD (Å)	10,80	5,40	11,15	10,97
	2	Etats occupés (%) <sup>a</sup>	33,15	30,08	26,55	10,23
		RMSD (Å)	12,46	10,56	4,98	10,94
1NGO	1	Etats occupés (%) <sup>a</sup>	51,05	42,91	6,05	
		RMSD (Å)	3,07	2,76	5,06	
	2	Etats occupés (%) <sup>a</sup>	96,86	3,12		
		RMSD (Å)	3,72	5,15		
3HXO	1	Etats occupés (%) <sup>a</sup>	50,35	26,48	15,46	7,72
		RMSD (Å)	5,49	4,58	6,85	10,83
	2	Etats occupés (%) <sup>a</sup>	38,54	35,32	26,15	
		RMSD (Å)	3,23	6,08	4,02	
3THW	1	Etats occupés (%) <sup>a</sup>	61,80	17,84	11,41	8,96
		RMSD (Å)	6,63	4,79	10,84	8,32
	2	Etats occupés (%) <sup>a</sup>	38,00	37,15	24,86	
		RMSD (Å)	5,39	7,53	3,92	
5HTO	1	Etats occupés (%) <sup>a</sup>	49,14	43,35	7,52	
		RMSD (Å)	1,69	2,08	2,31	
	2	Etats occupés (%) <sup>a</sup>	45,11	26,62	21,64	6,64
		RMSD (Å)	2,50	2,62	2,52	2,11

<sup>a</sup> Pourcentage de l'aMD occupé par une conformation issue d'un cluster.

Annexe 10. Résultats de TTclust pour les dynamiques sur structures expérimentales en TIP4P-Ew et système neutralisé non concentré en NaCl des 5 structures étudiées. Les clusters sont affichés dans l'ordre de leur représentativité dans l'aMD ainsi que la RMSD de la structure représentative face à la structure expérimentale.

Structure	aMD n°		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
1EZN	1	Etats occupés (%) <sup>a</sup>	34,05	30,39	13,15	12,15	10,27
		RMSD (Å)	6,67	4,87	8,66	10,66	6,52
	2	Etats occupés (%) <sup>a</sup>	35,53	29,73	23,95	10,80	
		RMSD (Å)	5,29	5,64	10,80	6,60	
1NGO	1	Etats occupés (%) <sup>a</sup>	31,87	31,80	12,71	12,11	11,52
		RMSD (Å)	2,79	3,52	4,05	4,03	3,16
	2	Etats occupés (%) <sup>a</sup>	44,70	40,20	15,11		
		RMSD (Å)	3,03	3,96	3,03		
3HXO	1	Etats occupés (%) <sup>a</sup>	40,30	31,25	17,65	10,82	
		RMSD (Å)	3,04	2,36	3,88	3,43	
	2	Etats occupés (%) <sup>a</sup>	40,82	34,54	24,65		
		RMSD (Å)	2,89	3,08	4,55		
3THW	1	Etats occupés (%) <sup>a</sup>	30,84	24,49	22,76	11,40	10,53
		RMSD (Å)	7,10	6,88	5,66	8,76	10,22
	2	Etats occupés (%) <sup>a</sup>	52,16	30,99	8,53	8,32	
		RMSD (Å)	4,25	5,04	5,51	7,65	
5HTO	1	Etats occupés (%) <sup>a</sup>	49,71	26,67	14,06	9,57	
		RMSD (Å)	1,38	2,14	2,24	2,15	
	2	Etats occupés (%) <sup>a</sup>	83,65	11,32	5,04		
		RMSD (Å)	2,01	2,64	5,82		

<sup>a</sup> Pourcentage de l'aMD occupé par une conformation issue d'un cluster.

Annexe 11. Résultats de TTclust pour les dynamiques sur modèles RNAdenovo produits à partir de la structure secondaire expérimentale en TIP3P et système neutralisé non concentré en NaCl des 5 structures étudiées. Les clusters sont affichés dans l'ordre de leur représentativité dans l'aMD ainsi que la RMSD de la structure représentative face à la structure expérimentale.

Structure	aMD n°		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
1EZN	1	Etats occupés (%) <sup>a</sup>	36,44	28,46	26,33	8,78	
		RMSD (Å)	13,19	9,03	10,93	11,35	
	2	Etats occupés (%) <sup>a</sup>	34,05	25,52	23,01	17,43	
		RMSD (Å)	3,98	12,38	10,67	5,54	
1NGO	1	Etats occupés (%) <sup>a</sup>	43,65	21,59	18,26	9,12	7,38
		RMSD (Å)	3,77	3,37	3,44	4,29	3,82
	2	Etats occupés (%) <sup>a</sup>	77,26	22,74			
		RMSD (Å)	4,56	4,44			
3HXO	1	Etats occupés (%) <sup>a</sup>	40,84	34,81	13,24	11,12	
		RMSD (Å)	6,06	11,11	10,58	3,84	
	2	Etats occupés (%) <sup>a</sup>	26,37	22,71	18,98	18,01	13,93
		RMSD (Å)	3,53	5,48	4,53	6,82	3,17
3THW	1	Etats occupés (%) <sup>a</sup>	42,54	21,84	19,26	16,38	
		RMSD (Å)	11,60	12,02	10,43	10,70	
	2	Etats occupés (%) <sup>a</sup>	35,74	25,88	22,17	16,21	
		RMSD (Å)	10,50	12,78	10,49	10,33	
5HTO	1	Etats occupés (%) <sup>a</sup>	62,57	31,48	5,96		
		RMSD (Å)	7,33	10,93	4,51		
	2	Etats occupés (%) <sup>a</sup>	43,06	34,35	12,42	10,18	
		RMSD (Å)	6,27	7,15	9,35	7,53	

<sup>a</sup> Pourcentage de l'aMD occupé par une conformation issue d'un cluster.

Annexe 12. Résultats de TTclust pour les dynamiques sur modèles RNAdenovo produits à partir de la structure secondaire expérimentale en TIP4P-Ew et système neutralisé non concentré en NaCl des 5 structures étudiées. Les clusters sont affichés dans l'ordre de leur représentativité dans l'aMD ainsi que la RMSD de la structure représentative face à la structure expérimentale.

Structure	aMD n°		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
1EZN	1	Etats occupés (%) <sup>a</sup>	30,94	28,60	27,80	12,66	
		RMSD (Å)	12,22	12,78	11,74	6,25	
	2	Etats occupés (%) <sup>a</sup>	34,61	25,00	21,91	18,48	
		RMSD (Å)	5,83	5,66	5,61	10,94	
1NGO	1	Etats occupés (%) <sup>a</sup>	30,48	27,85	20,63	21,05	
		RMSD (Å)	3,26	4,32	3,98	3,71	
	2	Etats occupés (%) <sup>a</sup>	63,10	22,66	14,25		
		RMSD (Å)	4,06	3,52	4,57		
3HXO	1	Etats occupés (%) <sup>a</sup>	54,33	17,99	14,04	7,54	6,12
		RMSD (Å)	11,41	5,51	6,37	11,87	12,48
	2	Etats occupés (%) <sup>a</sup>	40,76	25,84	24,52	8,89	
		RMSD (Å)	4,48	2,74	8,78	8,65	
3THW	1	Etats occupés (%) <sup>a</sup>	41,47	37,23	13,66	7,65	
		RMSD (Å)	11,27	9,17	8,99	7,82	
	2	Etats occupés (%) <sup>a</sup>	34,17	24,66	21,35	19,83	
		RMSD (Å)	11,47	11,99	12,68	10,08	
5HTO	1	Etats occupés (%) <sup>a</sup>	48,83	20,21	17,25	13,72	
		RMSD (Å)	5,29	4,14	5,16	3,39	
	2	Etats occupés (%) <sup>a</sup>	44,80	24,68	17,41	7,17	5,96
		RMSD (Å)	4,27	3,66	3,72	3,18	4,40

<sup>a</sup> Pourcentage de l'aMD occupé par une conformation issue d'un cluster.

Annexe 13. Résultats de TTclust pour les dynamiques sur structures expérimentales en TIP3P et système concentré à 0,1 M de NaCl des 5 structures étudiées. Les clusters sont affichés dans l'ordre de leur représentativité dans l'aMD ainsi que la RMSD de la structure représentative face à la structure expérimentale.

Structure	aMD n°		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
1EZN	1	Etats occupés (%) <sup>a</sup>	44,77	38,01	17,23			
		RMSD (Å)	12,57	10,98	9,85			
	2	Etats occupés (%) <sup>a</sup>	45,66	38,80	15,54			
		RMSD (Å)	7,36	5,56	6,91			
1NGO	1	Etats occupés (%) <sup>a</sup>	65,08	19,60	15,33			
		RMSD (Å)	3,82	3,21	4,56			
	2	Etats occupés (%) <sup>a</sup>	38,36	27,66	25,13	8,86		
		RMSD (Å)	3,19	3,00	4,10	4,22		
3HXO	1	Etats occupés (%) <sup>a</sup>	51,92	49,08				
		RMSD (Å)	6,76	3,46				
	2	Etats occupés (%) <sup>a</sup>	53,29	43,31	3,41			
		RMSD (Å)	5,10	6,60	7,59			
3THW	1	Etats occupés (%) <sup>a</sup>	42,64	25,09	15,36	9,16	6,93	0,84
		RMSD (Å)	6,33	6,29	8,13	4,74	10,44	13,68
	2	Etats occupés (%) <sup>a</sup>	28,22	26,06	18,71	14,41	12,61	
		RMSD (Å)	9,37	9,09	4,72	3,91	11,82	
5HTO	1	Etats occupés (%) <sup>a</sup>	53,80	18,86	9,82	9,79	7,73	
		RMSD (Å)	1,89	2,06	3,24	2,59	2,68	
	2	Etats occupés (%) <sup>a</sup>	44,09	37,17	13,54	5,22		
		RMSD (Å)	2,00	1,98	3,33	2,13		

<sup>a</sup> Pourcentage de l'aMD occupé par une conformation issue d'un cluster.

Annexe 14. Résultats de TTclust pour les dynamiques sur structures expérimentales en TIP4P-Ew et système concentré à 0,1 M de NaCl des 5 structures étudiées. Les clusters sont affichés dans l'ordre de leur représentativité dans l'aMD ainsi que la RMSD de la structure représentative face à la structure expérimentale.

Structure	aMD n°		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
1EZN	1	Etats occupés (%) <sup>a</sup>	51,55	34,11	14,35		
		RMSD (Å)	5,41	5,93	10,94		
	2	Etats occupés (%) <sup>a</sup>	57,03	36,25	6,72		
		RMSD (Å)	13,05	12,89	10,05		
1NGO	1	Etats occupés (%) <sup>a</sup>	65,80	17,44	16,77		
		RMSD (Å)	3,26	4,79	4,36		
	2	Etats occupés (%) <sup>a</sup>	40,14	31,84	23,57	4,46	
		RMSD (Å)	3,41	4,72	3,58	5,82	
3HXO	1	Etats occupés (%) <sup>a</sup>	61,72	38,29			
		RMSD (Å)	3,80	2,70			
	2	Etats occupés (%) <sup>a</sup>	32,39	24,02	20,06	14,19	9,34
		RMSD (Å)	3,84	5,36	4,73	3,35	4,47
3THW	1	Etats occupés (%) <sup>a</sup>	38,81	30,15	19,63	11,43	
		RMSD (Å)	4,09	5,82	3,97	8,84	
	2	Etats occupés (%) <sup>a</sup>	33,38	30,55	19,68	16,41	
		RMSD (Å)	5,30	4,41	10,15	3,47	
5HTO	1	Etats occupés (%) <sup>a</sup>	35,02	34,44	30,41		
		RMSD (Å)	2,83	3,37	2,93		
	2	Etats occupés (%) <sup>a</sup>	50,48	49,53			
		RMSD (Å)	2,86	1,65			

<sup>a</sup> Pourcentage de l'aMD occupé par une conformation issue d'un cluster.

Annexe 15. Résultats de TTclust pour les dynamiques sur modèles RNAdenovo produits à partir de la structure secondaire expérimentale en TIP3P et système concentré à 0,1 M de NaCl des 5 structures étudiées. Les clusters sont affichés dans l'ordre de leur représentativité dans l'aMD ainsi que la RMSD de la structure représentative face à la structure expérimentale.

Structure	aMD n°		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
1EZN	1	Etats occupés (%) <sup>a</sup>	29,95	25,88	25,53	14,21	4,45
		RMSD (Å)	11,28	10,66	12,79	5,16	9,17
	2	Etats occupés (%) <sup>a</sup>	44,44	34,87	12,74	7,97	
		RMSD (Å)	5,66	12,49	5,48	7,51	
1NGO	1	Etats occupés (%) <sup>a</sup>	34,68	27,01	20,57	17,75	
		RMSD (Å)	3,62	4,88	4,86	5,42	
	2	Etats occupés (%) <sup>a</sup>	41,25	27,82	19,98	10,96	
		RMSD (Å)	3,62	5,03	4,25	4,31	
3HXO	1	Etats occupés (%) <sup>a</sup>	45,43	19,07	16,23	16,05	3,24
		RMSD (Å)	4,02	6,85	6,25	4,35	6,55
	2	Etats occupés (%) <sup>a</sup>	39,73	27,30	18,29	14,69	
		RMSD (Å)	7,99	7,74	4,39	3,97	
3THW	1	Etats occupés (%) <sup>a</sup>	40,26	30,29	18,34	11,11	
		RMSD (Å)	12,81	9,33	13,94	14,12	
	2	Etats occupés (%) <sup>a</sup>	51,47	37,85	10,69		
		RMSD (Å)	12,99	9,14	4,59		
5HTO	1	Etats occupés (%) <sup>a</sup>	33,62	26,18	20,41	19,80	
		RMSD (Å)	9,57	7,83	8,88	8,02	
	2	Etats occupés (%) <sup>a</sup>	43,16	26,02	19,74	11,08	
		RMSD (Å)	8,34	9,69	9,29	5,22	

<sup>a</sup> Pourcentage de l'aMD occupé par une conformation issue d'un cluster.

Annexe 16. Résultats de TTclust pour les dynamiques sur structures expérimentales en TIP4P-Ew et système concentré à 0,1 M de NaCl des 5 structures étudiées. Les clusters sont affichés dans l'ordre de leur représentativité dans l'aMD ainsi que la RMSD de la structure représentative face à la structure expérimentale

Structure	aMD n°		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
1EZN	1	Etats occupés (%) <sup>a</sup>	58,85	26,69	14,47			
		RMSD (Å)	5,84	7,26	9,22			
	2	Etats occupés (%) <sup>a</sup>	46,32	25,75	19,19	8,75		
		RMSD (Å)	12,56	13,36	11,19	10,08		
1NGO	1	Etats occupés (%) <sup>a</sup>	52,29	33,73	13,99			
		RMSD (Å)	4,08	3,78	4,92			
	2	Etats occupés (%) <sup>a</sup>	73,44	19,05	7,51			
		RMSD (Å)	3,77	4,23	3,81			
3HXO	1	Etats occupés (%) <sup>a</sup>	43,53	29,73	17,08	9,67		
		RMSD (Å)	7,51	6,13	6,85	10,06		
	2	Etats occupés (%) <sup>a</sup>	41,44	27,88	27,63	3,06		
		RMSD (Å)	10,28	11,91	11,11	10,17		
3THW	1	Etats occupés (%) <sup>a</sup>	35,54	33,72	14,98	9,69	6,08	
		RMSD (Å)	8,94	7,60	7,18	3,99	7,35	
	2	Etats occupés (%) <sup>a</sup>	35,34	28,70	17,45	12,18	6,35	
		RMSD (Å)	11,17	11,88	11,67	8,58	8,25	
5HTO	1	Etats occupés (%) <sup>a</sup>	32,16	20,02	17,50	11,34	11,30	7,69
		RMSD (Å)	13,91	6,27	14,98	13,83	14,73	17,93
	2	Etats occupés (%) <sup>a</sup>	33,40	25,13	20,24	11,97	9,27	
		RMSD (Å)	14,07	14,01	14,02	14,54	15,26	

<sup>a</sup> Pourcentage de l'aMD occupé par une conformation issue d'un cluster.

Annexe 17. Résultats de TTclust pour les dynamiques sur structures expérimentales en TIP3P et système concentré à 0,1 M de NaCl des 5 structures étudiées. Les clusters sont affichés dans l'ordre de leur représentativité dans la GaMD ainsi que la RMSD de la structure représentative face à la structure expérimentale.

Structure	GaMD n°		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
1EZN	1	Etats occupés (%) <sup>a</sup>	33,96	29,12	29,11	7,81	
		RMSD (Å)	6,30	5,36	11,76	9,01	
	2	Etats occupés (%) <sup>a</sup>	27,12	26,15	19,66	17,22	9,86
		RMSD (Å)	6,25	5,93	5,65	11,13	8,36
1NGO	1	Etats occupés (%) <sup>a</sup>	38,87	32,85	28,29		
		RMSD (Å)	3,21	5,51	4,39		
	2	Etats occupés (%) <sup>a</sup>	59,25	25,07	15,69		
		RMSD (Å)	3,99	3,46	5,06		
3HXO	1	Etats occupés (%) <sup>a</sup>	49,79	24,02	19,87	6,32	
		RMSD (Å)	7,26	7,81	5,76	5,16	
	2	Etats occupés (%) <sup>a</sup>	40,41	27,12	21,16	11,31	
		RMSD (Å)	2,53	2,88	3,64	3,33	
3THW	1	Etats occupés (%) <sup>a</sup>	40,03	37,96	22,01		
		RMSD (Å)	5,81	8,64	2,98		
	2	Etats occupés (%) <sup>a</sup>	55,89	16,20	14,41	13,51	
		RMSD (Å)	3,70	6,49	10,91	13,18	
5HTO	1	Etats occupés (%) <sup>a</sup>	69,70	30,30			
		RMSD (Å)	1,70	1,89			
	2	Etats occupés (%) <sup>a</sup>	64,32	21,92	13,76		
		RMSD (Å)	1,86	2,16	2,11		

<sup>a</sup> Pourcentage de la GaMD occupé par une conformation issue d'un cluster.

Annexe 18. Résultats de TTclust pour les dynamiques sur structures expérimentales en TIP4P-Ew et système concentré à 0,1 M de NaCl des 5 structures étudiées. Les clusters sont affichés dans l'ordre de leur représentativité dans la GaMD ainsi que la RMSD de la structure représentative face à la structure expérimentale

Structure	GaMD n°		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
1EZN	1	Etats occupés (%) <sup>a</sup>	47,98	43,74	8,29			
		RMSD (Å)	10,69	5,16	10,12			
	2	Etats occupés (%) <sup>a</sup>	32,25	30,50	29,22	8,03		
		RMSD (Å)	5,04	4,97	6,38	9,46		
1NGO	1	Etats occupés (%) <sup>a</sup>	44,98	39,50	15,53			
		RMSD (Å)	3,71	4,20	3,20			
	2	Etats occupés (%) <sup>a</sup>	30,78	29,19	27,26	12,77		
		RMSD (Å)	2,69	3,32	2,43	2,19		
3HXO	1	Etats occupés (%) <sup>a</sup>	29,23	27,09	25,73	17,96		
		RMSD (Å)	2,27	3,25	3,27	2,24		
	2	Etats occupés (%) <sup>a</sup>	50,68	28,98	13,46	6,88		
		RMSD (Å)	5,03	4,04	3,63	4,80		
3THW	1	Etats occupés (%) <sup>a</sup>	35,22	22,43	19,48	14,56	8,32	
		RMSD (Å)	3,43	3,34	4,29	5,58	8,26	
	2	Etats occupés (%) <sup>a</sup>	39,02	25,84	22,06	5,30	4,64	3,14
		RMSD (Å)	2,55	3,52	3,66	6,14	5,05	7,68
5HTO	1	Etats occupés (%) <sup>a</sup>	43,36	27,99	21,08	7,58		
		RMSD (Å)	1,86	2,09	1,92	2,93		
	2	Etats occupés (%) <sup>a</sup>	41,30	28,35	17,16	13,19		
		RMSD (Å)	1,76	1,87	2,46	2,23		

<sup>a</sup> Pourcentage de la GaMD occupé par une conformation issue d'un cluster.

Annexe 19. Résultats de TTclust pour les dynamiques sur modèles RNAdenovo produits à partir de la structure secondaire expérimentale en TIP3P et système concentré à 0,1 M de NaCl des 5 structures étudiées. Les clusters sont affichés dans l'ordre de leur représentativité dans la GaMD ainsi que la RMSD de la structure représentative face à la structure expérimentale

Structure	GaMD n°		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
1EZN	1	Etats occupés (%) <sup>a</sup>	38,56	36,37	25,07	/	/
		RMSD (Å)	11,64	10,99	9,92	/	/
	2	Etats occupés (%) <sup>a</sup>	31,52	31,04	23,02	14,42	/
		RMSD (Å)	11,70	5,87	6,17	8,15	/
1NGO	1	Etats occupés (%) <sup>a</sup>	54,00	42,04	3,96	/	/
		RMSD (Å)	4,37	3,80	4,74	/	/
	2	Etats occupés (%) <sup>a</sup>	58,25	22,87	18,89	/	/
		RMSD (Å)	3,81	4,62	3,19	/	/
3HXO	1	Etats occupés (%) <sup>a</sup>	56,56	16,03	14,11	13,31	/
		RMSD (Å)	4,48	3,83	5,22	10,59	/
	2	Etats occupés (%) <sup>a</sup>	32,22	30,79	22,64	14,35	/
		RMSD (Å)	6,05	8,35	6,00	5,65	/
3THW	1	Etats occupés (%) <sup>a</sup>	39,64	21,60	21,27	17,50	/
		RMSD (Å)	12,68	3,13	5,09	10,61	/
	2	Etats occupés (%) <sup>a</sup>	30,66	23,69	21,09	14,02	10,54
		RMSD (Å)	12,34	10,11	11,89	10,50	10,01
5HTO	1	Etats occupés (%) <sup>a</sup>	32,39	29,55	25,00	8,00	5,07
		RMSD (Å)	4,77	3,99	5,90	7,06	8,50
	2	Etats occupés (%) <sup>a</sup>	34,96	25,02	21,54	18,49	/
		RMSD (Å)	6,35	6,03	6,76	7,53	/

<sup>a</sup> Pourcentage de la GaMD occupé par une conformation issue d'un cluster.

Annexe 20. Résultats de TTclust pour les dynamiques sur modèles RNAdenovo produits à partir de la structure secondaire obtenue avec mfold en TIP3P et système concentré à 0,1 M de NaCl des 3 structures étudiées. Les clusters sont affichés dans l'ordre de leur représentativité dans la GaMD ainsi que la RMSD de la structure représentative face à la structure expérimentale

Structure	GaMD n°		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
1EZN	1	Etats occupés (%) <sup>a</sup>	28,50	25,80	23,40	22,31	/
		RMSD (Å)	10,46	4,78	10,95	5,64	/
	2	Etats occupés (%) <sup>a</sup>	50,90	32,89	16,22	/	/
		RMSD (Å)	6,16	4,86	6,00	/	/
3HXO	1	Etats occupés (%) <sup>a</sup>	93,12	6,88	/	/	/
		RMSD (Å)	19,87	21,39	/	/	/
	2	Etats occupés (%) <sup>a</sup>	30,20	12,77	17,79	15,75	23,49
		RMSD (Å)	19,50	19,30	18,73	19,22	20,58
5HTO	1	Etats occupés (%) <sup>a</sup>	60,54	39,46	/	/	/
		RMSD (Å)	4,21	5,38	/	/	/
	2	Etats occupés (%) <sup>a</sup>	59,13	25,64	15,24	/	/
		RMSD (Å)	9,08	7,33	7,19	/	/

<sup>a</sup> Pourcentage de la GaMD occupé par une conformation issue d'un cluster.

Annexe 21. Résultats de TTclust pour les dynamiques sur modèles RNAdenovo produits à partir de la structure secondaire expérimentale en TIP4P-Ew et système concentré à 0,1 M de NaCl des 5 structures étudiées. Les clusters sont affichés dans l'ordre de leur représentativité dans la GaMD ainsi que la RMSD de la structure représentative face à la structure expérimentale

Structure	GaMD n°		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
1EZN	1	Etats occupés (%) <sup>a</sup>	37,64	33,92	28,45		
		RMSD (Å)	11,37	5,67	10,01		
	2	Etats occupés (%) <sup>a</sup>	26,61	25,59	23,88	12,50	11,42
		RMSD (Å)	10,91	10,93	10,92	10,72	11,16
1NGO	1	Etats occupés (%) <sup>a</sup>	54,75	27,66	11,24	6,35	
		RMSD (Å)	3,47	3,91	3,54	4,56	
	2	Etats occupés (%) <sup>a</sup>	71,93	16,54	11,53		
		RMSD (Å)	3,64	4,29	3,32		
3HXO	1	Etats occupés (%) <sup>a</sup>	31,07	29,28	19,25	10,24	10,17
		RMSD (Å)	5,96	7,27	6,05	7,56	9,93
	2	Etats occupés (%) <sup>a</sup>	38,35	37,40	24,25		
		RMSD (Å)	5,84	5,36	9,74		
3THW	1	Etats occupés (%) <sup>a</sup>	33,83	32,51	25,97	5,71	2,00
		RMSD (Å)	3,98	4,18	5,27	6,66	10,89
	2	Etats occupés (%) <sup>a</sup>	36,40	32,21	28,77	2,62	
		RMSD (Å)	8,15	7,00	10,77	6,29	
5HTO	1	Etats occupés (%) <sup>a</sup>	28,26	20,88	20,84	15,34	14,80
		RMSD (Å)	6,01	3,53	7,21	4,93	7,99
	2	Etats occupés (%) <sup>a</sup>	30,82	26,90	26,70	9,61	5,97
		RMSD (Å)	4,05	4,47	4,81	4,37	3,87

<sup>a</sup> Pourcentage de la GaMD occupé par une conformation issue d'un cluster.

Annexe 22. Résultats de TTclust pour les dynamiques sur modèles RNAdenovo produits à partir de la structure secondaire obtenue avec mfold en TIP4P-Ew et système concentré à 0,1 M de NaCl des 3 structures étudiées. Les clusters sont affichés dans l'ordre de leur représentativité dans la GaMD ainsi que la RMSD de la structure représentative face à la structure expérimentale

Structure	GaMD n°		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
1EZN	1	Etats occupés (%) <sup>a</sup>	47,48	30,89	17,59	4,04	
		RMSD (Å)	6,19	11,76	9,80	5,36	
	2	Etats occupés (%) <sup>a</sup>	53,25	37,29	9,46		
		RMSD (Å)	13,16	12,72	10,93		
3HXO	1	Etats occupés (%) <sup>a</sup>	66,53	24,14	9,34		
		RMSD (Å)	20,76	22,01	18,58		
	2	Etats occupés (%) <sup>a</sup>	29,62	27,65	24,00	18,74	
		RMSD (Å)	19,41	20,59	22,61	17,53	
5HTO	1	Etats occupés (%) <sup>a</sup>	30,72	30,21	25,51	8,93	4,64
		RMSD (Å)	10,02	8,48	9,22	9,45	9,02
	2	Etats occupés (%) <sup>a</sup>	50,79	30,36	18,85		
		RMSD (Å)	7,99	8,11	7,63		

<sup>a</sup> Pourcentage de la GaMD occupé par une conformation issue d'un cluster.