Modélisation, analyse et classification de motifs structuraux d'ARN à partir de leur contexte, par des méthodes d'algorithmique de graphes

Coline Gianfrotta

université PARIS-SACLAY

Modélisation, analyse et classification de motifs structuraux d'ARN à partir de leur contexte, par des méthodes d'algorithmique de graphes

*M*odeling, analysis and classification of RNA structural motifs from their context, using graph algorithmic methods

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 580 Sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat : Informatique
Graduate School : Informatique et sciences du numérique
Référent : Université de Versailles-Saint-Quentin-en-Yvelines

Thèse préparée dans les unités de recherche **DAVID (Université Paris-Saclay, UVSQ)**, et **LISN (Université Paris-Saclay, CNRS)** sous la direction de **Dominique BARTH**, Professeur des universités, le co-encadrement de **Alain DENISE**, Professeur

Thèse soutenue à Versailles le 10 octobre 2022, par

Coline GIANFROTTA

Composition du jury

Membres du jury avec voix délibérative

Sarah COHEN-BOULAKIA Professeure des universités, LISN, Université Paris Saclay	Présidente
Guillaume FERTIN Professeur, LS2N, Université de Nantes	Rapporteur & Examinateur
Christine GASPIN Directrice de recherche, INRAE-MIAT, Castanet-Tolosan	Rapportrice & Examinatrice
Alessandra CARBONE Professeure, LGQB, Sorbonne Université	Examinatrice

THESE DE DOCTORAT

NNT : 2022UPASG056

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Modélisation, analyse et classification de motifs structuraux d'ARN à partir de leur contexte, par des méthodes d'algorithmique de graphes.

Mots clefs : Bioinformatique, ARN, structures 3D, algorithmique de graphes, clustering

Résumé : Dans cette thèse, nous étudions le contexte structural de motifs structuraux d'ARN dans le but de progresser vers leur prédiction. En effet, certains motifs d'ARN, sous-structures apparaissant de façon récurrente dans les structures d'ARN, restent difficiles à prédire, en raison de la présence d'interactions non canoniques dans ces motifs, et en raison de la distance sur la séquence primaire séparant les différentes parties de ces motifs. Nous modélisons ainsi par des graphes le contexte structural topologique de ces motifs, et comparons les contextes des différentes occurrences en utilisant plusieurs algorithmes de graphes. Nous classifions ensuite les occurrences de motif selon leurs similarités de contexte topologique et selon leurs similarités de contexte 3D, à l'aide d'un algorithme de clustering recouvrant.

Dans un premier temps, nous montrons sur un jeu de données de trois motifs structuraux que les similarités observées entre les contextes topologiques sont cohérentes avec les similarités entre les contextes 3D. Cela indique que le contexte topologique peut être suffisant pour déterminer le contexte 3D pour ces trois motifs.

Dans un deuxième temps, nous étudions plusieurs classifications d'occurrences du motif A-minor, selon des similarités de contexte 3D. Nous y observons que des similarités de contexte 3D existent entre occurrences non homologues, ce qui pourrait être le signe d'un phénomène de convergence évolutive. De plus, nous observons que certaines parties du contexte 3D semblent mieux conservées que d'autres entre occurrences non homologues.

Dans un troisième temps, nous étudions la capacité de prédiction du contexte topologique commun à des occurrences de motif A-minor, partageant des contextes 3D similaires, ainsi que la capacité de prédiction d'un signal de séquence sur ces mêmes occurrences. Pour cela, nous étudions la fréquence d'apparition de cette topologie et de ces séquences dans des structures d'ARN en l'absence de motifs A-minor. Nous en concluons que la topologie et la séquence associées représentent un bon signal pour la majorité des classes d'occurrences non homologues étudiées.

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Title : Modeling, analysis and classification of RNA structural motifs from their context, using graph algorithmic methods.

Keywords : bioinformatics, RNA, 3D structures, graph algorithmic, clustering

Abstract : In this thesis, we study the structural context of RNA structural motifs in order to make progress in their prediction. Indeed, some RNA motifs, which are substructures appearing recurrently in RNA structures, remain difficult to predict, because of the presence of non-canonical interactions in these motifs, and because of the distance on the primary sequence between the different parts of these motifs. We therefore model the topological structural context of these motifs by graphs, and compare the contexts of the different occurrences using several graph algorithms. We then classify the motif occurrences according to their topological context similarities and according to their 3D context similarities, using an overlapping clustering algorithm.

First, we show on a dataset of three structural motifs that the observed similarities between the topological contexts are consistent with the similarities between the 3D contexts. This indicates that the topological context may be sufficient to determine the 3D context for these three motifs.

In a second step, we study several classifications of occurrences of the A-minor motif, according to 3D context similarities. We observe that 3D context similarities exist between non-homologous occurrences, which could be a sign of an evolutionary convergence phenomenon. Moreover, we observe that some parts of the 3D context seem to be better conserved than others between non-homologous occurrences.

In a third step, we study the predictive ability of the common topological context of A-minor motif occurrences, sharing similar 3D contexts, as well as the predictive ability of a sequence signal on these same occurrences. To this end, we study the occurrence of this topology and sequence in RNA structures in the absence of A-minor motifs. We conclude that the topology and the sequence represent a good signal for the majority of the studied classes.

Modélisation, analyse et classification de motifs structuraux d'ARN à partir de leur contexte, par des méthodes d'algorithmique de graphes

par Coline Gianfrotta

Direction de la thèse

Dominique BARTH Professeur, DAVID, UVSQ, Université Paris Saclay	Directeur de thèse
Alain DENISE Professeur, LISN & I2BC, Université Paris Saclay	Co-encadrant
Vladimir REINHARZ Professeur, Department of Computer Science, Université du Québec à Montréal	Invité

Remerciements

Une thèse n'est bien sûr pas seulement issue du travail de celui ou celle qui l'écrit. De nombreuses personnes y contribuent scientifiquement et humainement, et ces personnes méritent alors d'être mentionnées et remerciées.

Je remercie ainsi tout d'abord mes encadrants de thèse, Dominique Barth et Alain Denise, pour m'avoir guidée tout au long de cette thèse. J'ai gagné aussi bien en connaissances scientifiques qu'en expérience grâce à leurs conseils avisés.

Je remercie également les membres de mon jury de thèse, Christine Gaspin et Guillaume Fertin pour avoir accepté d'être mes rapporteurs, et Alessandra Carbone et Sarah Cohen-Boulakia pour avoir accepté de faire partie de mon jury. Leurs commentaires au travers de leurs rapports ou pendant la soutenance ont été très enrichissants.

Je remercie Vladimir Reinharz, professeur à l'université de Montréal, qui a grandement contribué aux travaux que je décris ici par son regard neuf et critique et ses idées nouvelles, et à qui je dois la plupart des données utilisées en entrée de mes méthodes.

Je remercie aussi les membres des deux équipes de recherche qui m'ont accueillie, l'équipe ALMOST du laboratoire DAVID de l'UVSQ, et l'équipe de Bioinfo du LISN de l'Université Paris-Saclay. Je ne me suis sans doute pas autant intégrée aux équipes que ce que j'aurai dû, à cause de mon caractère réservé, mais je savais pouvoir compter sur l'aide généreuse de tout le monde, que ce soit d'un point de vue pratique ou scientifique. Je remercie en particulier les doctorants (Stéphanie Chevalier, Pierre Andrieu, Wei Chen, Ylène Aboulfath, Loric Duhazé, Stefi Nouleho, Maël Guiraud…) avec qui j'ai pu échanger des expériences, des conseils ou des moments de doute ou de saturation. Je remercie également les permanents et non permanents avec qui j'ai travaillé pour les enseignements à Versailles (Sandrine Vial, Franck Quessette, Yann Strozecki, Pierre Coucheney, Thierry Mautor…). J'ai beaucoup appris d'eux et des enseignements que j'ai assurés, et je suis ainsi heureuse d'avoir eu l'occasion de le faire.

Je remercie aussi mes co-bureaux de stage avant la thèse, Bintou Fofana et Nassim Haddad, pour avoir partagé avec moi ces 5-6 mois passés sous les combles, qui m'ont mis le pied à l'étrier pour commencer la thèse.

Enfin, je remercie ma famille, mon père, ma mère et mes soeurs, pour leur soutien inébranlable dans tout ce que je fais, et leur support pendant la rédaction de cette thèse. Petit merci particulier à ma mère pour sa relecture attentive de ce manuscrit, malgré les difficultés que cela pose pour un non-initié. Caresses à ma chienne chérie.

Table des matières

Liste des figures

Liste des tables

Introduction

Les ARN sont des molécules biologiques, possédant une multitude de fonctions. Nos connaissances sur ces molécules sont en pleine expansion ; de nouveaux ARN sont ainsi régulièrement découverts [52], et la fonction de certains d'entre eux n'est pas encore élucidée. Des hypothèses considèrent même l'ARN comme impliqué dans l'origine de la vie [57], étant donné que la diversité de ses fonctions dépasse celle de l'ADN et des protéines, autres acteurs de l'expression de l'information génétique.

De nombreuses familles d'ARN sont ainsi impliquées dans l'expression de l'information génétique, telles que les ARN messagers, les ARN de transfert ou les ARN ribosomiques. Les ARN messagers sont les vecteurs de la traduction de l'information stockée dans l'ADN en protéines. On dit souvent que ce sont des ARN *codants*. Toutes les autres familles d'ARN sont appelées *non codants* [73], car ils ne sont pas traduits en protéines. La majorité des ARN produits sont non codants et impliqués dans des fonctions de régulation. On compte par exemple les riboswitches [45], les ARN interférents [33], ou les ARN guides [32]. Certains ARN, comme les ARN ribosomiques, ont quant à eux une fonction catalytique.

Une molécule d'ARN est produite, à partir d'une portion d'ADN, sous la forme d'une chaîne linéaire de *nucléotides*, appelée *séquence primaire*, et se replie ensuite dans l'espace, par l'intermédiaire d'interactions entre nucléotides, pour se stabiliser et assurer sa fonction. Elle adopte alors une *structure tridimensionnelle* (3D).

La fonction d'une molécule d'ARN lui est ainsi conférée par sa structure tridimensionnelle (3D). Pour connaître la fonction d'une molécule d'ARN nouvellement découverte, ou pour reproduire la fonction d'un ARN en biologie de synthèse [28, 21], il est ainsi crucial de connaître sa structure tridimensionnelle.

Pour déterminer la structure tridimensionnelle d'une molécule d'ARN, il existe des méthodes telles que la cristallographie aux rayons X [39] ou la résonance magnétique nucléaire (RMN) [9], permettant de déterminer la position de chaque atome constituant une structure tridimensionnelle de molécules d'ARN, à partir d'un échantillon en solution. Cependant ces méthodes présentent des désavantages, parmi lesquels leur coût en temps et en matériel.

C'est ainsi qu'a émergé le problème de prédiction de structures d'ARN,

visant à être capable, par des méthodes computationnelles, de déterminer la structure tridimensionnelle d'une molécule d'ARN, à partir de sa séquence primaire. La séquence primaire d'une molécule d'ARN est en effet bien plus facile à obtenir que sa structure 3D, par les méthodes de séquençage à haut débit de ces vingt dernières années [76]. Etant donné que les molécules d'ARN sont constituées, à l'instar de l'ADN, d'un alphabet (4 types de nucléotides) bien plus restreint que celui des protéines, le problème de prédiction de structures d'ARN était au départ pensé plus facile à résoudre que celui visant à prédire les structures de protéines [112]. Cependant, il s'est avéré plus complexe qu'il n'y paraissait. Cette complexité vient notamment de la diversité des interactions apparaissant au sein d'une structure d'ARN, et des modèles thermodynamiques utilisés pour représenter un repliement de manière réaliste. Il a d'ailleurs été démontré que des modèles thermodynamiques réalistes rendent NP-difficile le problème de repliement d'une molécule d'ARN [104].

Etudié depuis plus de 40 ans, le problème de prédiction de structures d'ARN a ainsi fait l'objet de nombreux travaux. Le repliement de l'ARN étant considéré comme hiérarchique [15, 112], les premiers travaux se sont basés sur des modélisations de structures intermédiaires, telles que les *structures secondaires*. Une structure secondaire d'ARN est constituée d'un ensemble d'interactions, dites *canoniques*, apparaissant entre des paires de nucléotides, et formant des empilements de paires de nucléotides. La possibilité de considérer cette structure secondaire comme un arbre a permis l'utilisation d'algorithmes de programmation dynamique pour sa prédiction, en ayant recours à des modèles thermodynamiques pour trouver la (ou les) structure(s) la (ou les) plus stable(s) en énergie [81, 129]. Plus récemment, des approches stochastiques ont également été utilisées [29, 68].

Cependant, la prédiction de la structure secondaire ne suffit pas à élucider la fonction d'une molécule d'ARN [83]. Les autres interactions, non prises en compte dans la structure secondaire, jouent un rôle important dans le repliement de la molécule, et donc dans sa structure tridimensionnelle fonctionnelle. Elles sont plus difficiles à prédire, car leur contribution énergétique à la structure globale n'est pas facile à déterminer. De plus, la prise en compte d'interactions *longue distance*, c'est-à-dire reliant des nucléotides qui sont éloignés les uns des autres sur la séquence primaire, augmente la complexité des problèmes à résoudre.

Plusieurs types de méthodes sont actuellement étudiées pour tenter de surmonter ces difficultés. Certaines d'entre elles utilisent la dynamique moléculaire [109, 103, 50, 84, 11], visant à simuler le repliement d'une molécule d'ARN, et ainsi prédire la structure la plus probable. Cependant, ces méthodes ne sont applicables que pour de petites molécules. D'autres méthodes utilisent des approches algorithmiques comme la théorie des jeux [13] ou des méthodes probabilistes (Monte Carlo notamment [54, 55]) pour tenter de prédire la structure tridimensionnelle à gros grain, c'est-à-dire, la forme globale de la molécule, sans connaître la position exacte des atomes et les interactions de façon précise. Les structures obtenues par ces méthodes sont cependant

très imprécises. Enfin, des méthodes utilisent le fait que les structures d'ARN sont constituées d'un ensemble de sous-structures récurrentes, appelées *motifs structuraux* ou *modules* [26, 83, 24, 126, 123, 96]. Prédire la présence de ces motifs peut alors permettre de guider la prédiction de la structure 3D globale. Certains de ces motifs, liant des nucléotides éloignés sur la séquence primaire, restent cependant difficiles à prédire. C'est le cas du *motif A-minor* par exemple [66]. Ainsi, de nombreuses difficultés n'ont pas encore été surmontées dans la prédiction des structures 3D d'ARN.

Dans cette thèse, nous nous intéressons en particulier à ce dernier type de motifs structuraux, apparaissant de manière récurrente dans les structures d'ARN, et reliant des régions éloignées de la séquence primaire (que nous appellerons alors *à longue distance*). Nous étudions le *contexte structural* de motifs d'ARN à longue distance, à deux échelles : en considérant, d'une part, l'ensemble des interactions canoniques et non canoniques apparaissant autour du motif (nous définirons par la suite exactement quelles interactions sont considérées), et d'autre part, la sous-structure 3D de ce même contexte, c'est-à-dire la position des atomes des nucléotides de ce contexte.

Le but de cette thèse est de déterminer dans quelle mesure le contexte structural formé des interactions canoniques et non canoniques d'un motif structural peut déterminer la structure 3D [41]. Si l'importance de ce contexte structural pour la structure 3D est considérable, cela signifie que le contexte structural peut être un bon critère pour la prédiction de la structure 3D des motifs étudiés.

Pour répondre à cette question, nous modélisons par des graphes le contexte structural formé des interactions canoniques et non canoniques. Ce contexte structural sera alors qualifié de *topologique* pour le différencier de la sous-structure 3D qui sera qualifié de *contexte 3D*.

Nous recherchons ensuite des similarités entre les contextes topologiques des occurrences de motif d'ARN d'un côté, et entre les contextes 3D de ces mêmes occurrences de l'autre côté, pour déterminer si notre modélisation capture bien les similarités de contexte 3D. Les similarités de contextes topologiques sont calculées dans des sous-graphes communs maximisant le nombre d'arêtes (*Maximum Common Edge Subgraph* MCES) [90], tandis que les similarités de contexte 3D sont calculées à l'aide d'une mesure d'alignements de structures 3D qu'est la RMSD (*Root Mean Square Deviation*) [16]. Ces mesures peuvent alors donner lieu à des classifications d'occurrences de motif d'ARN selon des topologies de contexte similaires ou selon des contextes 3D similaires.

Parmi les motifs que nous étudions, l'un a particulièrement retenu notre intérêt car c'est un motif très répandu dans les structures d'ARN [63] et sa prédiction reste impossible par les méthodes actuelles, comme évoqué précédemment. Il s'agit du motif *A-minor*. Ce motif n'a fait l'objet que de peu de classifications exhaustives selon des informations structurales [101]. Etant donné que le contexte 3D contient davantage d'informations que la topologie seule, nous étudions ainsi les similarités et différences structurales dans les

contextes 3D des occurrences de ce motif, dans le but d'en déduire des caractéristiques structurales communes, qui pourraient apporter des connaissances nouvelles sur ce motif et sa formation. De plus, nous souhaitons déterminer dans quelle mesure la présence d'une occurrence de motif A-minor peut être prédite, en examinant uniquement des informations sur son contexte topologique et la séquence associée. Pour cela, nous utilisons les caractéristiques communes de contexte topologique et de séquence, associées à des ensembles d'occurrences de motif A-minor possédant des contextes 3D similaires, pour tenter d'inférer la présence d'un motif A-minor.

Ce document est composé de 5 chapitres. Le premier chapitre posera le problème de prédiction des motifs structuraux d'ARN à longue distance, après avoir fait une présentation succincte des connaissances actuelles sur les molécules d'ARN, et des méthodes de prédiction existantes. Le deuxième chapitre décrira la modélisation sous forme de graphes que nous utilisons pour représenter le contexte structural de motifs d'ARN, les algorithmes permettant de comparer ces contextes, ainsi que la définition de contexte 3D que nous considérons. Dans le troisième chapitre, nous étudierons le cas de trois motifs d'ARN à longue distance, dont le motif A-minor, pour tenter de répondre à notre question initiale. Pour savoir s'il existe un lien entre le contexte structural topologique et le contexte 3D pour ces trois motifs, nous comparerons ainsi les topologies de contexte et les contextes 3D précédemment définis. Nous étudierons davantage cette corrélation pour le motif A-minor, dans les quatrième et cinquième chapitres. Dans le quatrième chapitre, nous présenterons plusieurs classifications d'occurrences de motif A-minor selon leur contexte 3D, dans le but d'étudier les similarités et les différences structurales entre occurrences de motif A-minor. Enfin, le cinquième chapitre étudiera la capacité de prédiction de la topologie de contexte et de la séquence pour des occurrences de motif A-minor possédant des contextes 3D similaires.

Chapitre 1

L'ARN et la prédiction de structures et de motifs structuraux

1.1 Introduction

Comme évoqué dans l'introduction, les molécules d'ARN possèdent de nombreuses fonctions essentielles à la survie et au fonctionnement d'une cellule. Ces fonctions leur sont conférées par la structure tridimensionnelle qu'elles adoptent dans l'espace.

Le problème de prédiction de structures d'ARN consiste ainsi à tenter de déterminer la structure tridimensionnelle qu'adopte une molécule d'ARN à partir de sa séquence en nucléotides. C'est un problème étudié depuis plus de 40 ans, mais de nombreuses difficultés restent encore à surmonter.

Dans ce chapitre, nous allons présenter les principales fonctions des molécules d'ARN dans la cellule dans les sections 1.2 et 1.3. Puis, nous présenterons la structure chimique d'une molécule d'ARN dans la section 1.4. Dans la section 1.5, nous aborderons les étapes du repliement d'une molécule d'ARN dans l'espace, et les modélisations à l'aide de graphes qui ont été utilisées pour les représenter. Dans la section 1.6, nous évoquerons les méthodes non computationnelles permettant de déterminer une structure tridimensionnelle d'ARN. Enfin, dans la section 1.7, nous présenterons les méthodes actuelles de prédiction des structures secondaires et 3D d'ARN, ainsi que les difficultés rencontrées dans ces prédictions.

1.2 La cellule et l'expression de l'information génétique

La cellule est le plus petit constituant fondamental de tout organisme vivant. Une cellule est constituée d'une membrane plasmique entourant un *cytoplasme* dans lequel se trouvent les molécules nécessaires au fonctionnement de la cellule. Au cours de l'évolution, les organismes se sont séparés en trois règnes du Vivant. Deux règnes sont constitués d'organismes

unicellulaires *procaryotes* (les Bactéries et les Archées), et le troisième est constitué d'organismes unicellulaires ou multicellulaires *eucaryotes*, comme les animaux et les plantes. Les cellules eucaryotes possèdent des organites délimités par des membranes internes, contrairement aux cellules procaryotes. En particulier, une cellule eucaryote possède un noyau dans lequel se trouve le matériel génétique alors que le matériel génétique d'une cellule procaryote se trouve dans le cytoplasme.

D'après la théorie fondamentale de la biologie moléculaire, l'expression de l'information génétique dans une cellule se déroule en plusieurs étapes. Une molécule d'Acide RiboNucléique dit messager (ARNm) est créée à partir d'un gène de l'Acide DésoxyriboNucléique (ADN), par une étape de *transcription* (Figure 1.1a). Puis cet ARN messager peut subir des modifications et est ensuite pris en charge par le ribosome, qui est un complexe de protéines et d'ARN ribosomiques (ARNr), pour permettre la synthèse d'une protéine, au cours de la *traduction* (Figure 1.1b). La protéine est synthétisée par ajout d'une suite d'acides aminés, codés par lecture des triplets de nucléotides de l'ARNm (voir section 1.4), appelés codons, grâce aux ARN de transfert (ARNt).

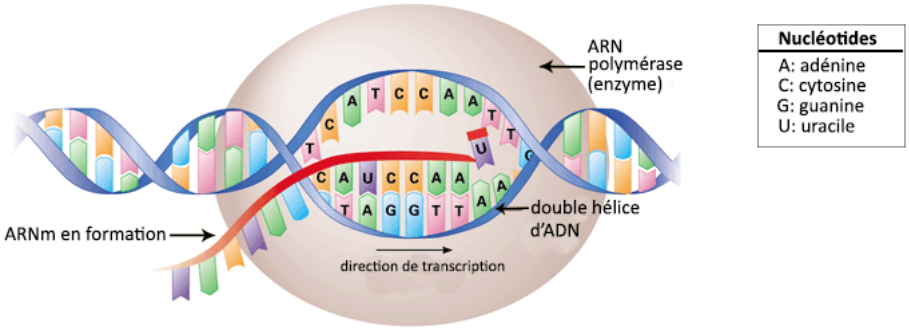
Les protéines produites peuvent alors assurer une multitude de fonctions, telles que, par exemple, la catalyse de réactions chimiques, le transport d'autres molécules, ou le maintien de la structure de certains constituants de la cellule.

Il a ainsi été longtemps pensé que l'ARN était uniquement un vecteur de l'information génétique, intermédiaire entre l'ADN, stockant l'information, et les protéines, exprimant cette information en une fonction. Cependant, les molécules d'ARN ont de nombreuses autres fonctions, ce que nous allons aborder dans les paragraphes suivants.

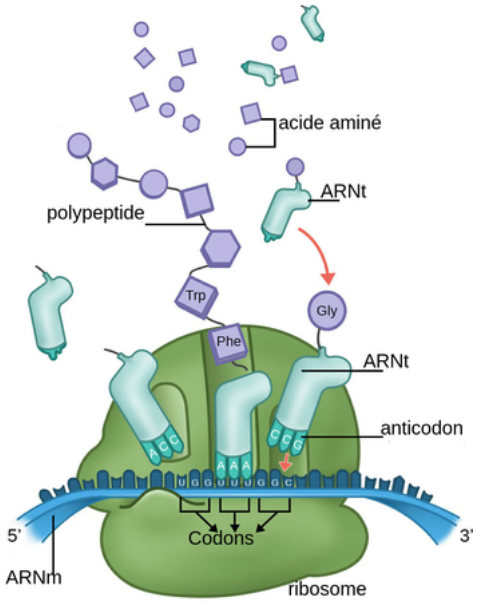
1.3 Les ARN, des molécules aux multiples fonctions

Les molécules d'ARN possèdent de nombreuses fonctions essentielles à un organisme vivant. De nombreux ARN agissent au cours du processus de synthèse des protéines. L'ARNm, en plus d'être vecteur de l'information génétique, peut avoir une action sur la régulation de sa traduction [77], en adoptant une structure tridimensionnelle particulière (voir section 1.5). L'ARNm subit des modifications avant la traduction, ce qu'on appelle épissage de l'ARNm. Au cours de cet épissage, des portions de l'ARNm, appelées introns, sont supprimées. Les autres portions, appelées exons, ne sont pas toujours toutes conservées, ce qui permet d'obtenir plusieurs protéines différentes à partir d'un même transcrit. Les introns possèdent sans doute d'autres fonctions, non encore élucidées. Ils semblent être par exemple impliqués dans la protection du génome, en empêchant la formation de complexes ADN/ARN, ce qui est délétère pour la cellule [12].

D'autres petits ARN non codants ont également été découverts, qui représentent en réalité la majorité des transcrits à partir de l'ADN. Ces petites molécules sont très diverses et très nombreuses. En effet, on compte 4069



(a) transcription d'une portion d'ADN en ARNm
(issue de https ://www.utsouthwestern.edu/labs/bioinformatics-lab/analysis/rna-seq/)



(b) traduction d'un ARNm en protéine
(image issue de https ://theory.labster.com/rna-translation-fr/)

Figure 1.1 – Processus de transcription et de traduction

familles distinctes d'ARN dans la base de données RFAM [44, 52] (en décembre 2021), qui recense et organise en familles fonctionnelles les molécules d'ARN connues à ce jour. Par l'étude des génomes et transcriptomes d'organismes vivants, de nouveaux ARN non codants sont régulièrement découverts [51]. Parmi ces ARN non codants, certains ont un rôle de régulation de l'expression génétique. C'est le cas des *ARN interférents* [33], qui se fixent par des interactions Watson-Crick (voir section 1.5) à un ARN messager pour empêcher le ribosome de s'y fixer à son tour, et ainsi inhiber la traduction de cet ARNm. On peut également évoquer l'implication de petits *ARN guides* dans le système CRISPR/Cas9 [47]. Ce système, présent chez certains procaryotes, est un système de protection de la cellule contre l'ADN étranger. Une molécule d'ARN guide se fixe, par l'intermédiaire d'interactions Watson-Crick, sur une portion d'ADN étranger. La nucléase Cas9 reconnaît cet ARN et coupe spécifiquement la molécule d'ADN étranger en des sites particuliers. Ce système a été montré

comme pouvant être utilisé en édition génétique, dans des approches thérapeutiques, notamment dans [22], ce qui a révolutionné le domaine de la biologie moléculaire et valu un prix Nobel aux auteurs de ces travaux.

On peut également évoquer les *riboswitchs* [45]. Ce sont de petites molécules se trouvant le plus souvent dans la partie non traduite, en amont d'un ARNm. Un riboswitch comporte deux parties. L'une de ces parties, appelée aptamère, peut lier un ligand, ce qui modifie la structure tridimensionnelle du riboswitch et peut ainsi bloquer ou activer la traduction de l'ARNm.

Un grand nombre de molécules d'ARN sont des *ribozymes*, c'est-à-dire qu'ils ont une activité catalytique. C'est le cas de l'ARN ribosomique, par exemple, qui catalyse la formation des liaisons entre acides aminés pour former les protéines[80]. La ribonucléase P est également un ribozyme permettant la maturation des ARNt pour les rendre fonctionnels.

Notons également que l'ARN peut être génome, comme c'est le cas de certains ARN de virus.

1.4 La structure chimique des molécules d'ARN

L'ARN est constitué d'une chaîne linéaire d'éléments, appelés *nucléotides*.

Un nucléotide est constitué de trois ensembles de groupements chimiques : une base azotée, un sucre (le ribose), et un groupement tri-phosphate. Tous les nucléotides contiennent un ribose et un groupement phosphate (souvent appelés squelette ribose-phosphate), mais le type de base azotée peut différer et donner ainsi plusieurs types de nucléotides différents. Il existe 4 types principaux de bases azotées dans l'ARN, et nous ne considérerons que celles-ci dans cette thèse : l'adénine (A), la guanine (G), la cytosine (C) et l'uracile (U) (Figure 1.2). L'adénine et la guanine sont des *purines*, et la cytosine et l'uracile des *pyrimidines*.

Dans cette chaîne, souvent dénommée *séquence primaire*, les nucléotides sont reliés entre eux par des liaisons phosphodiester. Ce sont des liaisons covalentes fortes, difficiles à briser. Une liaison phosphodiester a lieu entre le ribose d'un nucléotide et le groupement phosphate d'un autre nucléotide. Par convention, la séquence primaire est orientée de son extrémité 5'-phosphate (noté 5', car le carbone 5' du ribose est lié au groupement phosphate) à son extrémité 3'-OH (noté 3' car le carbone 3' du ribose est lié à l'atome d'oxygène du groupement OH du ribose) (Figure 1.2)

Une molécule d'ARN peut contenir un nombre très variable de nucléotides, allant de quelques dizaines pour les petits ARN à 2000 ou 3000 nucléotides pour les ARN ribosomiques, et pouvant même atteindre plusieurs dizaines de milliers de nucléotides pour les ARN de virus [115].

Dans le but d'assurer sa fonction, une molécule d'ARN se replie dans l'espace pour former une structure tridimensionnelle. Nous allons à présent aborder les interactions permettant le repliement de l'ARN, ainsi que les modélisations par des graphes permettant de représenter les différents niveaux de structures de l'ARN.

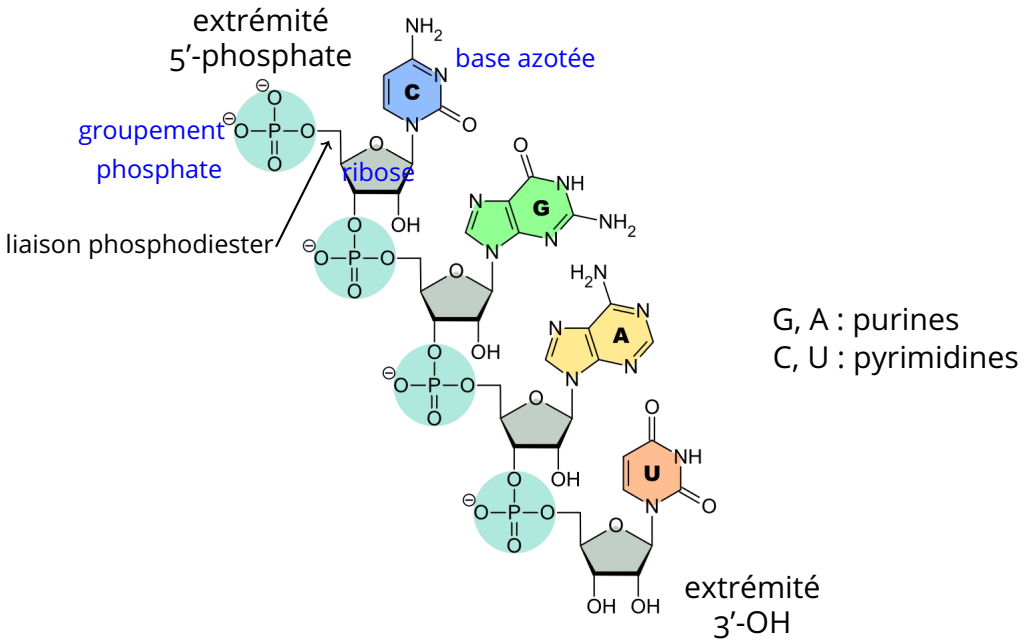


Figure 1.2 – Séquence primaire de 4 nucléotides CGAU (orientée de l'extrémité 5'-phosphate à l'extrémité 3'-OH). Les composants de chaque nucléotide sont indiqués.

1.5 Les structures d'ARN et leurs représentations

Synthétisée sous la forme d'une chaîne linéaire de nucléotides, une molécule d'ARN se replie sur elle-même pour se stabiliser. Des appariements entre nucléotides se forment ainsi, par l'intermédiaire de liaisons hydrogène, le plus souvent entre deux bases azotées. Ces liaisons hydrogène apparaissent entre un atome d'hydrogène d'un nucléotide et un atome d'oxygène ou un atome d'azote d'un autre nucléotide, car ce sont des atomes électronégatifs. Ces liaisons sont des liaisons dipôle-dipôle, de plus faible énergie que les interactions phosphodiester covalentes. Un même nucléotide peut former ce type d'interactions avec plusieurs autres nucléotides, et presque toutes les paires de types de nucléotides (A,C,G,U) peuvent interagir. Les interactions les plus stabilisantes sont appelées *canoniques* [62], tandis que les autres sont appelées *non canoniques*. D'autres types interactions apparaissent également comme les interactions électrostatiques d'empilements entre paires de bases (*stacking*).

Les interactions canoniques et non canoniques ont été caractérisées et classifiées comme nous allons l'aborder dans le paragraphe suivant.

1.5.1 Les interactions canoniques et non canoniques

Toutes les interactions formées de liaisons hydrogène entre paires de nucléotides ont été décomposées en familles, dans la nomenclature de Leontis–Westhof [62].

Chaque nucléotide possède trois côtés pouvant interagir avec un autre nucléotide : le côté Watson-Crick, le côté Hoogsteen et le côté Sugar (voir Figure 1.3). L'interaction entre deux nucléotides peut se faire selon deux orientations, en fonction de l'orientation de la liaison entre la base azotée et le ribose pour chaque nucléotide, par rapport aux liaisons hydrogène. L'orientation est *cis* si la liaison ribose-base azotée est dans le même sens pour les deux nucléotides par rapport aux liaisons hydrogène, et sinon l'orientation est *trans*.

La nomenclature Leontis-Westhof [62] définit ainsi 12 familles, en fonction du côté de chaque nucléotide impliqué dans l'interaction, et en fonction de l'orientation *cis* ou *trans* de l'interaction. A chaque famille est associé un symbole, comme décrit dans la Table 1.1.

Les interactions canoniques font partie de l'une de ces familles : la famille *cis* Watson-Crick/ Watson-Crick, car le côté Watson-Crick de chaque nucléotide est impliqué dans l'interaction, selon l'orientation *cis*. Elles apparaissent entre des paires de nucléotides de types particuliers : trois liaisons hydrogène peuvent apparaître entre la base azotée d'un nucléotide à cytosine (C) et la base azotée d'un nucléotide à guanine (G), et deux liaisons hydrogène peuvent se former entre la base azotée d'un nucléotide à adénine (A) et la base azotée d'un nucléotide à uracile (U) (Figure 1.4). Moins classiquement mais fréquemment, deux liaisons hydrogène peuvent également se former entre un G et un U. Ce type de liaisons est appelé liaison "Wobble", et sera considéré comme canonique dans cette thèse.

A l'instar de la double hélice d'ADN, ces interactions canoniques induisent la formation de double hélice dans l'ARN, constituées d'une suite de paires de nucléotides empilées les unes sur les autres (voir Figure 1.4). La contribution énergétique de ces empilements a été étudiée et caractérisée dans des modèles thermodynamiques [36, 114].

Toutes les autres interactions impliquant des liaisons hydrogène sont appelées non canoniques. Il n'existe pas de modèle thermodynamique permettant de les décrire.

Orientation	Côtés de l'interaction	Symbole
Cis	Watson–Crick / Watson–Crick (cWW)	-●-
Trans	Watson–Crick / Watson–Crick (tWW)	-○-
Cis	Watson–Crick / Hoogsteen (cWH)	●-■
Trans	Watson–Crick / Hoogsteen (tWH)	○-□
Cis	Watson–Crick / Sugar Edge (cWS)	●- ▶
Trans	Watson–Crick / Sugar Edge (tWS)	○-▷
Cis	Hoogsteen / Hoogsteen (cHH)	-■-
Trans	Hoogsteen / Hoogsteen (tHH)	-□-
Cis	Hoogsteen / Sugar Edge (cHS)	■- ▶
Trans	Hoogsteen / Sugar Edge (tHS)	□- ▷
Cis	Sugar Edge / Sugar Edge (cSS)	-▶-
Trans	Sugar Edge / Sugar Edge (tSS)	-▷-

Table 1.1 – Description et symboles de la nomenclature Leontis–Westhof [62]

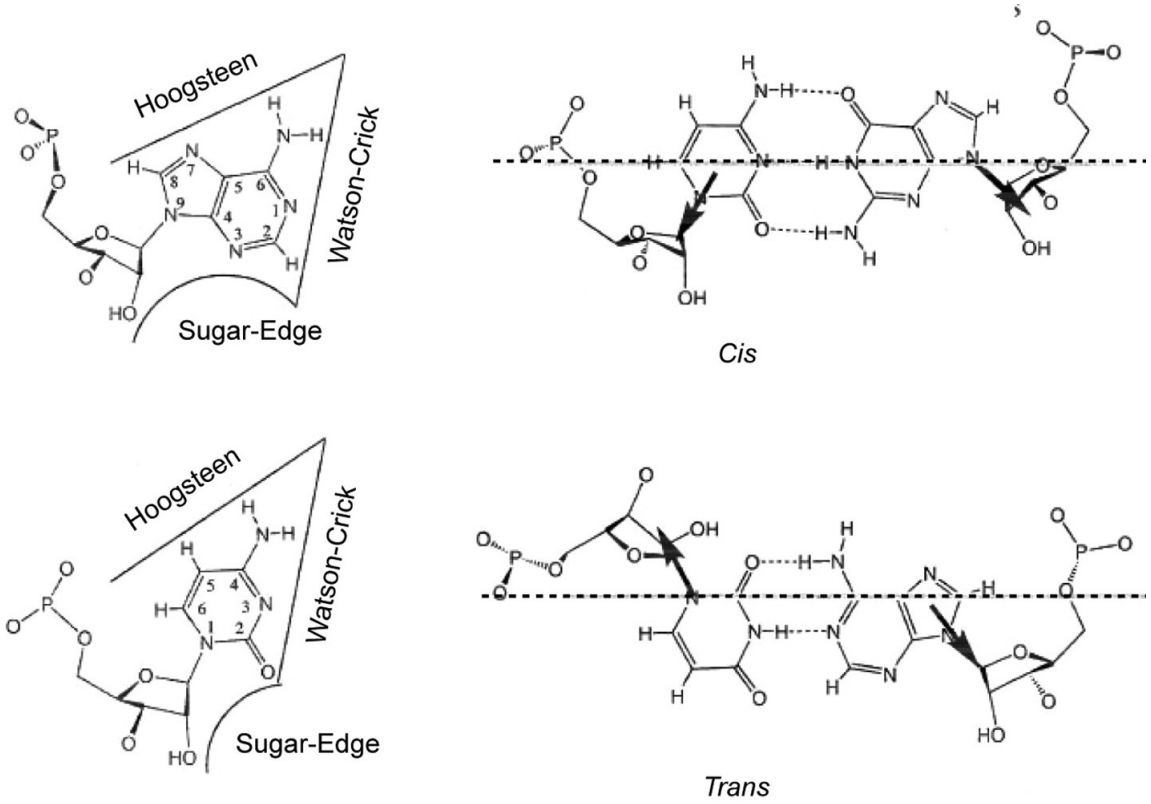


Figure 1.3 – Les trois côtés d'un nucléotide (à gauche), avec en haut une purine, et en bas une pyrimidine, et les deux orientations possibles d'une interaction (à droite). Figure issue de [62]

1.5.2 Le repliement hiérarchique de l'ARN

Les molécules d'ARN ont tendance à se replier de manière à devenir les plus stables possibles en énergie. Les interactions qui se forment permettent ainsi de stabiliser la molécule et casser ces interactions nécessite ensuite une énergie importante. Cependant, le repliement adopté par une molécule donnée n'est pas unique. Plusieurs conformations différentes peuvent même coexister [111]. Ce repliement peut dépendre de nombreux facteurs, tels que l'environnement dans lequel se trouve la molécule [112], et les interactions avec d'autres constituants de la cellule.

Le repliement d'une molécule d'ARN est depuis longtemps considéré comme hiérarchique [112]. Des interactions canoniques semblent apparaître en premier, formant des empilements solides de paires de nucléotides. Cet ensemble d'interactions est généralement appelé *structure secondaire*. Puis d'autres interactions apparaissent : notamment des interactions canoniques formant des *pseudonœuds* dans la structure secondaire, et des interactions non canoniques, de plus faible énergie. L'ensemble des interactions canoniques et non canoniques forme une structure secondaire parfois qualifiée d'*enrichie*. Finalement, l'ensemble de toutes les interactions formées induit la structure tertiaire ou tridimensionnelle de la molécule, qui est sa structure native et lui confère une fonction donnée. La Protein Data Bank

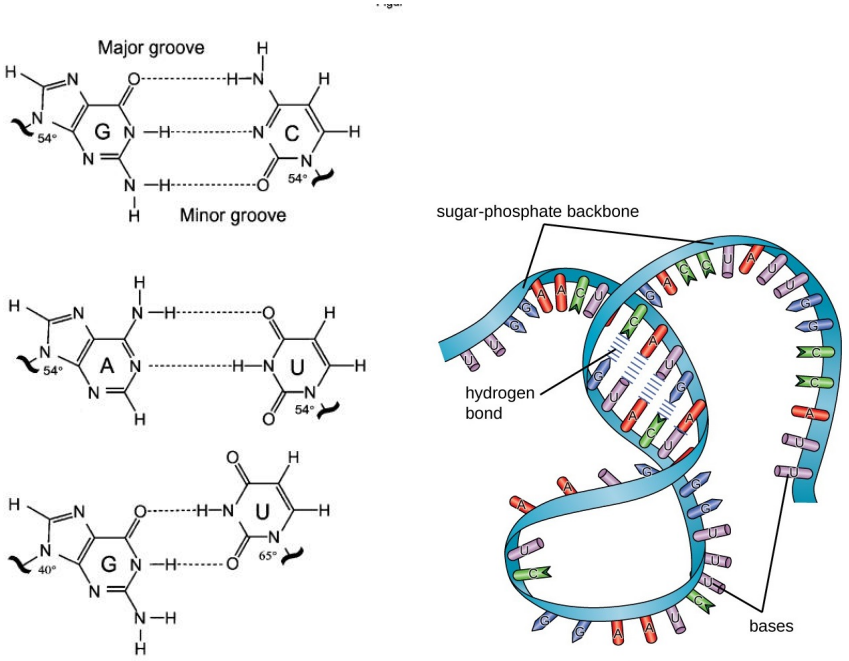


Figure 1.4 – Interactions canoniques formées des bases azotées C-G, A-U et G-U (à gauche, image issue de [116], *minor* et *major groove* indiquent le petit et le grand sillon de l'hélice formée) et formation d'une hélice dans une séquence primaire (à droite, image issue de https ://microbenotes.com/rna-properties-structure-types-and-functions/).

(PDB [8]) est une base de données très fournie, stockant notamment les structures 3D d'ARN et de protéines obtenues expérimentalement (voir section 1.6). La plupart des études faites sur les structures d'ARN utilisent cette base.

Ces différents niveaux de structures ont alors été modélisés par des graphes, dans le but d'étudier et d'utiliser ces niveaux intermédiaires pour tenter de prédire la structure native. Dans la suite de cette thèse, nous appellerons *topologie* un graphe permettant de représenter un ensemble d'interactions au sein d'une structure d'ARN.

Nous allons à présent décrire plus en détails ces différents niveaux de structures, et aborder les modélisations les plus classiquement utilisées.

1.5.3 Les différents niveaux représentant la topologie d'une structure d'ARN

La structure secondaire sans pseudonœud

Une structure secondaire est ainsi composée du sous-ensemble le plus grand possible, de toutes les interactions canoniques de la molécule, qui ne forment pas de croisement lorsque la structure est dessinée sur un demi-plan supérieur, avec les nucléotides de la séquence alignés de l'extrémité 5' à l'extrémité 3' (séquence arc-annotée, voir Figure 1.5b pour un exemple). On peut également dessiner cette structure comme un graphe planaire extérieur

et non orienté dans lequel les sommets correspondent aux nucléotides et les arêtes aux interactions covalentes et canoniques (voir Figure 1.5a pour un exemple). De nombreuses autres représentations équivalentes, que nous ne détaillerons pas ici, existent également (expression bien parenthésée, graphe de corde, etc).

Les interactions canoniques d'une structure secondaire forment des empilements de paires de nucléotides appelés *hélices*, et les nucléotides non appariés forment des *boucles* [4, 107].

Il existe différents types de boucles en fonction du nombre d'hélices voisines (voir Figure 1.5) :

- une *boucle terminale* est une suite de nucléotides non appariés, fermée par une seule paire de bases. Une boucle terminale est donc adjacente à une seule hélice.

- une *boucle interne* est constituée de deux suites de nucléotides non appariés fermées par une paire de bases à chaque extrémité, connectant ainsi deux hélices.

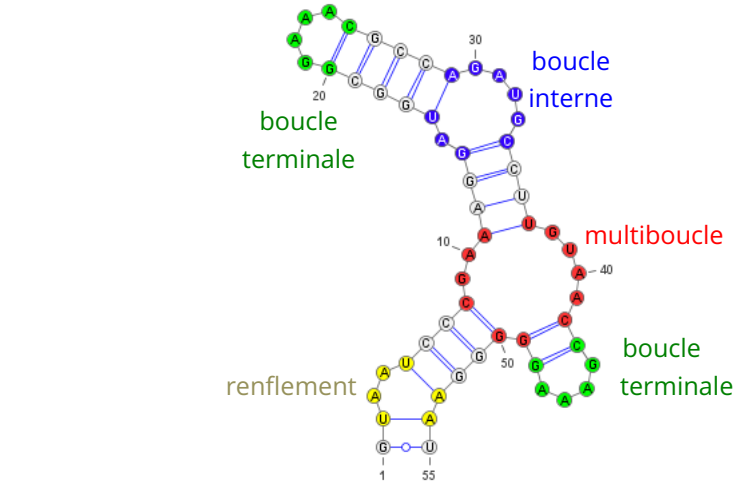
- un *renflement* (*bulge* en anglais) est une boucle interne particulière, dans laquelle une seule suite de nucléotides non appariés est fermée par deux paires de bases à chaque extrémité, et connecte deux hélices

- une *multiboucle* ou *jonction* est un ensemble de nucléotides non appariés qui connecte trois hélices ou plus.

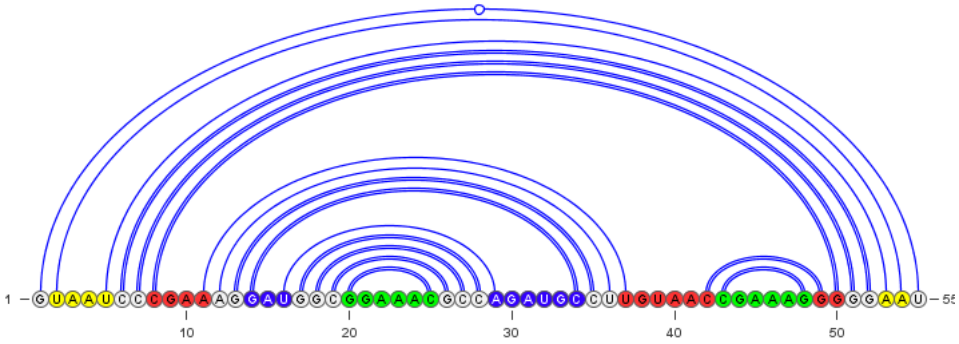
Chaque boucle et chaque hélice d'une structure secondaire sont appelées *élément de structure secondaire* (SSE).

La propriété de non-croisement d'une structure secondaire rend possible de la représenter par un arbre. C'est ce qui a permis d'utiliser des algorithmes de programmation dynamique pour prédire ces structures [81, 129, 68], comme nous le verrons plus en détail par la suite. De plus, la représentation arborescente des structures secondaires est utilisée pour comparer et aligner des structures secondaires d'ARN dans le but d'en déduire des classes fonctionnelles, et dans un but de prédiction également [59, 102, 46].

Bien que l'échelle du nucléotide soit souvent utilisée dans les méthodes de prédiction utilisant la programmation dynamique [81, 68], l'échelle d'un élément de structure secondaire est également considérée. Dans ce cas, un arbre représentant une structure secondaire possède des sommets correspondant aux boucles (terminales, internes, renflements, multiboucles) et une arête est présente entre deux sommets si les deux boucles sont connectées dans la structure secondaire par une hélice d'au moins une paire de bases [59]. Cela permet de s'affranchir des différences dans la taille des SSE ou dans les types de nucléotides impliqués. En effet, ces variations importent moins pour comparer des molécules d'une même classe fonctionnelle que les variations entre arrangements globaux de SSE [43]. D'autres modélisations utilisent de même des arbres dans lesquels les sommets représentent des éléments de structure secondaire pondérés par leur taille [35, 105].



(a) graphe planaire extérieur



(b) séquence arc-annotée

Figure 1.5 – Deux représentations d'une structure secondaire fictive, obtenues avec le logiciel VARNA [25]. En (a) sont indiqués les différents types de SSE dans un graphe planaire, et en (b) est présentée la séquence arc-annotée. C'est une structure secondaire sans pseudonœud puisqu'il n'y a pas de croisement d'arêtes dans cette deuxième représentation.

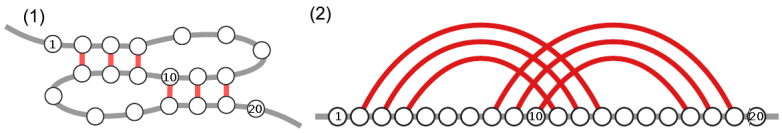
La structure secondaire enrichie

Nous appellerons *structure secondaire enrichie*, une structure secondaire comme décrite précédemment, à qui sont ajoutées les autres interactions canoniques et non canoniques apparaissant dans la structure.

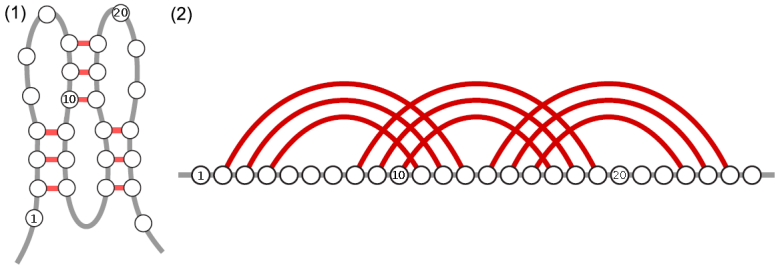
Ces interactions sont souvent organisées en ensembles d'interactions comme nous allons le décrire dans les deux paragraphes suivants.

Les pseudonœuds

Les nucléotides d'une boucle (interne ou terminale) de structure secondaire peuvent former des appariements canoniques avec des nucléotides d'une autre portion d'ARN (de la même molécule ou d'une autre molécule). On appelle ces ensembles d'interactions des pseudonœuds. Ces pseudonœuds possèdent des



(a) pseudonœud de type H, entre une boucle terminale et une autre boucle, située en dehors de l'hélice qui la délimite



(b) pseudonœud de type kissing-hairpin, entre deux boucles terminales

Figure 1.6 – Exemples de deux types de pseudonœud, sous forme planaire (1) et par la séquence arc-annotée (2).

rôles importants dans le repliement et la fonction de la molécule [77]. Ils sont par exemple impliqués dans le décalage du cadre de lecture dans la traduction[42].

Les pseudonœuds sont le résultat d'au moins un croisement entre deux hélices, que l'on peut notamment observer, dans la représentation de séquence arc-annotée, par un croisement d'arêtes (voir Figure 1.6).

Une structure secondaire avec pseudonœud ne peut être représentée par un arbre. D'autres modélisations ont alors été définies, comme par exemple, le *graphe dual* [37, 85] dans lequel les sommets correspondent aux hélices de la structure secondaire et une arête est présente entre deux sommets si les deux hélices sont adjacentes à une même boucle.

Les interactions non canoniques organisées en motifs structuraux

Il a longtemps été pensé que les interactions canoniques de la structure secondaire avec ou sans pseudœud étaient les plus importantes pour le repliement de la molécule [112], et que les considérer suffisait à caractériser et prédire ce repliement. Cependant, depuis une vingtaine d'années, le rôle des autres interactions, en particulier non canoniques, se révèle de plus en plus primordial dans le repliement correct de la molécule, ainsi que dans sa fonction [83]. C'est la raison pour laquelle ces interactions sont de plus en plus étudiées et considérées dans les modèles.

Les interactions non canoniques sont souvent organisées en ensembles d'interactions qui apparaissent conjointement de manière récurrente dans les structures d'ARN, parfois également avec des interactions canoniques. On les appelle alors *motifs structuraux* ou *modules*.

Ici, nous appellerons *motifs locaux* les ensembles d'interactions non

canoniques apparaissant au sein d'un élément de structure secondaire, et *motifs à longue distance* les ensembles d'interactions non canoniques apparaissant entre plusieurs éléments de structure secondaire distincts. Le terme motif fait souvent davantage référence à une séquence plutôt qu'à une structure récurrente, c'est pourquoi nous ajoutons les suffixes "locaux" ou "à longue distance" pour éviter la confusion.

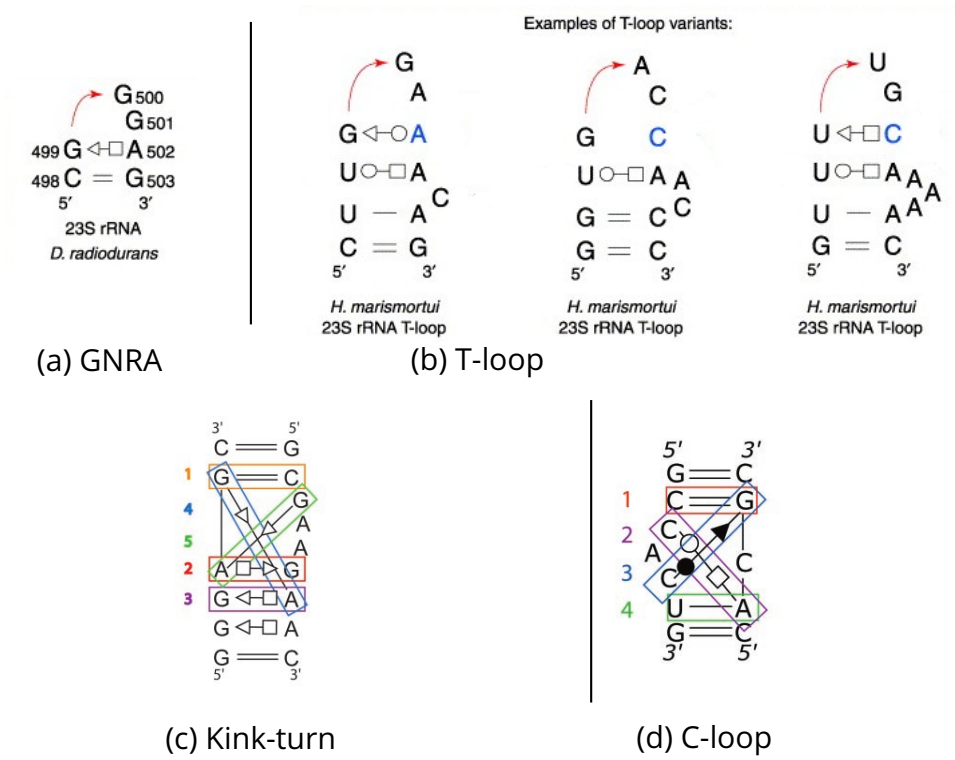
Certains de ces motifs structuraux ont d'abord été observés et étudiés, par l'expertise humaine [6, 56, 75]. Ils ont été caractérisés par des similarités structurales et/ou des séquences consensus.

On distingue ainsi les motifs locaux des motifs à longue distance :

- les motifs locaux apparaissent au sein d'une boucle de structure secondaire. Ces motifs sont souvent caractérisés par une signature de séquence, un ensemble d'interactions non canoniques et une géométrie particulière. Parmi ces motifs, on peut évoquer par exemple :
  - les motifs se trouvant au sein de boucles terminales, comme par exemple les T-loop, ou les boucles GNRA à 4 nucléotides [63], dont le nom vient de leur séquence consensus (G, A ou C ou G ou U (N), G ou A (R pour purine) et A) (voir exemples en Figure 1.7)
  - les motifs se trouvant au sein de boucles internes, comme par exemple les kink-turns [56], provoquant l'apparition de coudes dans les structures 3D, ou bien les *sarcin-ricin loop* ou les C-loops (voir exemples en Figure 1.7)
  - les motifs se trouvant au sein d'une multiboucle (ou jonction), comme par exemple les *k-junctions* [118], basées sur l'architecture d'un kink-turn.
- les motifs à longue distance relient plusieurs éléments de structure secondaire plus ou moins éloignés sur la séquence primaire. C'est le cas du motif A-minor [66], sur lequel porte une partie de cette thèse. Un motif A-minor de type I/II est un assemblage de deux nucléotides consécutifs (souvent des A) interagissant avec deux paires canoniques consécutives, par le petit sillon de l'hélice (*minor groove*). C'est cette particularité qui lui a donné son nom.

  Il existe en réalité 4 versions d'interactions de type A-minor, en fonction de la position des nucléotides les uns par rapport aux autres, mais seules deux de ces versions forment le motif A-minor (type I/II) auquel nous allons nous intéresser (Figure 1.8). Il s'agit donc de l'assemblage des interactions de type I et des interactions de type II. Le type I est formé d'un nucléotide x relié aux deux extrémités d'une paire de bases canonique ; ce nucléotide est relié à l'une extrémité par une interaction de type *cis*-Sugar-Sugar (cSS), et à l'autre extrémité par une interaction de type *trans*-Sugar-Sugar (tSS). Le type II est formé uniquement d'un nucléotide y relié, par une interaction cSS, à un autre nucléotide, lui-même impliqué dans une paire de bases canonique. Il est à noter que

les deux nucléotides consécutifs x et y (numérotés 5 et 6 sur la Figure 1.8Ac) sont souvent des nucléotides à adénine, mais il existe des variantes. Les occurrences de motifs A-minor que l'on considérera dans cette thèse ne possèdent pas tous des nucléotides à adénine en ces positions.



(a) GNRA (b) T-loop

(c) Kink-turn (d) C-loop

Figure 1.7 – Exemples de motifs locaux, avec leurs séquences et leurs interactions non canoniques. Les motifs GNRA et T-loop présentés sont trouvés dans les molécules indiquées, et les motifs kink-turn et C-loop sont des motifs typiques trouvés dans plusieurs molécules. Les images sont issues de [63] pour GNRA et T-loop et de [65] pour kink-turn et C-loop. Pour les GNRA et T-loop, les flèches rouges indiquent l'orientation de la séquence primaire de l'extrémité 5' vers l'extrémité 3'. Les interactions encadrées pour les motifs kink-turn et C-loop sont les interactions principales.

Des méthodes automatiques ont été développées pour détecter des motifs structuraux connus dans les structures d'ARN et pour en découvrir de nouveaux, en utilisant différentes informations.

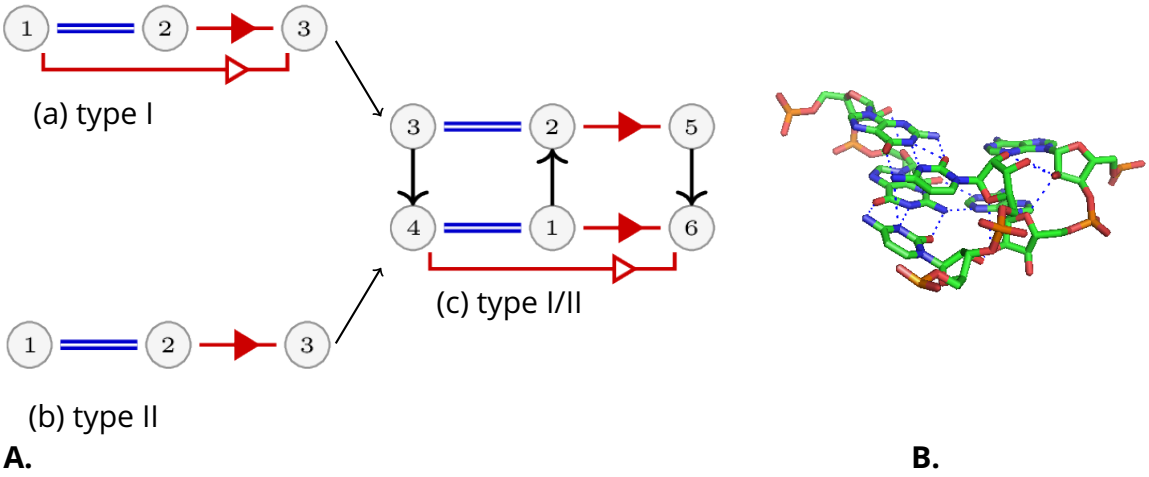


Figure 1.8 – Interactions du motif A-minor de type I/II (en A., images issues de [92]) et exemple de structure tridimensionnelle d'un motif A-minor de type I/II (en B), orientée dans le même sens que les interactions en Ac. En A, les numéros 1,2,3,4,5,6 désignent des nucléotides dont l'identité peut varier entre A,C,G et U.

De nombreuses méthodes ont cherché à détecter et identifier les motifs locaux récurrents apparaissant au sein de boucles de structure secondaire. Pour cela, différentes caractéristiques sont utilisées :

- Certaines méthodes utilisent uniquement les informations de séquence et de topologie, c'est-à-dire la position et le type des paires de bases canoniques ou non canoniques, en les représentant par des graphes. Par exemple, Pasquali *et al.* [85] modélisent des structures secondaires avec pseudonœud par des graphes duaux et recherchent des isomorphismes de graphes pour détecter des motifs de pseudoneuds. Djelloul et Denise[31] (RNA3DMotifs) utilisent une représentation classique de graphe non orienté pour modéliser un élément de structure secondaire et les interactions non canoniques qu'il contient : un sommet correspond à un nucléotide et est étiqueté par son type (A,C,G,U), et une arête correspond à une interaction covalente, canonique ou non canonique. Les auteurs recherchent alors les plus grands sous-graphes communs entre les éléments de structure secondaire ainsi représentés, pour détecter des motifs locaux récurrents, en prenant en compte uniquement les familles d'interactions non canoniques. Djelloul et Denise recherchent des isomorphismes exacts. Un peu plus tard, Zhong *et al.* [17] (RNAMotifScan) ont de même établi une méthode permettant de rechercher des occurrences de sous-structures dans une structure, par une approche d'alignement de structures. Pour être alignées, deux interactions non canoniques doivent être de même famille et les paires de bases doivent être isostériques [61]. Des paires de bases isostériques occupent exactement le même volume, et peuvent donc être substituées l'une à l'autre sans perturber la structure 3D.

- D'autres méthodes recherchent des motifs locaux récurrents caractérisés par une forme 3D commune (distances et valeurs d'angles particuliers entre atomes des nucléotides, ou positions des atomes dans l'espace). Par exemple, la base de données RNA 3D Motif Atlas [87] contient tous les motifs de boucles terminales et de boucles internes, groupés par géométrie similaire. Pour cela, les auteurs extraient toutes les boucles des structures de la PDB, alignent ces boucles 3D avec le programme FR3D [98], calculent des valeurs de géométrie et classifient les boucles en utilisant une recherche de clique maximum. Une autre base de données, RNA Bricks [23], classifie les éléments de structure secondaire possédant des structures 3D similaires en conservant les structures dans leur environnement avec d'autres constituants. De la même façon, RNAMotifContrast [48] classifie les boucles selon leurs similarités de structures 3D, en les comparant selon la RMSD [16] (RMSD introduite en introduction, page 17).

Ces méthodes ont ainsi permis de cataloguer toutes les occurrences des motifs déjà connus (GNRA, C-loop, kink-turn, etc), et d'en détecter automatiquement de nouveaux. Certaines d'entre elles étant régulièrement mises à jour (RNA 3D Motif Atlas, RNA Bricks 2), elles permettent de répertorier les motifs trouvés dans les structures 3D nouvellement obtenues.

Cependant, ces méthodes ne recherchent pas des motifs apparaissant entre différents éléments de structure secondaire, comme par exemple le motif A-minor. C'est en effet un problème plus complexe à résoudre du point de vue algorithmique, puisqu'il faut pouvoir considérer des ensembles d'interactions reliant des nucléotides éloignés sur la séquence primaire.

Récemment (en 2018), des études se sont intéressées à ce problème de détection de motifs à longue distance. La principale méthode permettant de détecter des motifs à longue distance, à ma connaissance, est celle utilisée dans la base de données CaRNAval [92] (ayant également une version de 2021[106]). Y sont regroupés tous les réseaux d'interactions récurrents (RIN), c'est-à-dire les ensembles d'interactions canoniques et non canoniques connectant au moins deux SSE, et pouvant être trouvés en au moins deux occurrences exactes dans les structures d'ARN. Les auteurs utilisent une représentation de graphes d'interactions, puis recherchent le sous-graphe commun maximum à deux graphes d'interactions.

Nous avons ainsi présenté un éventail des modélisations utilisées pour représenter et analyser les différents niveaux de structures secondaires des molécules d'ARN. Nous allons à présent évoquer différentes méthodes non computationnelles permettant de déterminer les structures tridimensionnelles d'ARN, avant de nous intéresser aux méthodes computationnelles tentant de prédire les structures secondaires et tridimensionnelles d'ARN.

1.6 Les méthodes de détermination des structures tridimensionnelles d'ARN

Des méthodes exploitant les propriétés physico-chimiques des molécules biologiques permettent de déterminer la position de chaque atome constituant la structure tridimensionnelle (ou tertiaire) d'une molécule d'ARN.

Parmi ces méthodes, celle qui reste la plus utilisée est la cristallographie aux rayons X [39]. Un cristal est obtenu à partir d'un système biologique constitué de molécules d'ARN au sein de leur environnement. La formation d'un cristal dépend des conditions physico-chimiques du milieu (pH, quantité d'ions, etc). Cette étape peut être longue à réaliser. Puis, le cristal est soumis à un faisceau de rayons X, qui va alors diffracter ces rayons. La figure de diffraction obtenue peut alors permettre de déterminer la position des atomes du cristal. Cette méthode peut donner des résultats très précis, mais elle est coûteuse en temps et en matériel, et ne garantit pas que la molécule se trouve exactement sous cette forme dans la cellule lorsqu'elle est en solution et non cristallisée.

La Résonance Magnétique Nucléaire (RMN) est parfois également utilisée[9]. Elle exploite les propriétés magnétiques des atomes des molécules biologiques. Les molécules en solution sont soumises à des champs magnétiques importants, ce qui donne un spectre RMN, à partir duquel peuvent être déduites les contraintes géométriques des atomes, et ainsi leur position dans l'espace. Notons que la RMN est surtout utilisée pour de petites molécules (quelques centaines d'atomes).

La cryo-microscopie électronique (cryoEM), utilisée depuis peu pour déterminer les structures 3D de protéines à l'échelle atomique, semble prometteuse pour les structures d'ARN [70]. Dans cette technique, les molécules sont illuminées par un faisceau d'électrons et peuvent être observées au microscope électronique, après cryofixation, c'est-à-dire après que les molécules sont figées par congélation.

A partir de ces structures tridimensionnelles dont le nombre augmente constamment, il est possible de déterminer les interactions canoniques et non canoniques constituant la structure, par des méthodes comme FR3D [98] ou DSSR [69]. La méthode FR3D est à la fois une méthode d'annotation d'interactions et de recherche de motifs dans les structures tertiaires d'ARN. Les interactions canoniques et non canoniques ainsi que les empilements sont calculés, à partir de la géométrie des atomes des nucléotides dans la structure 3D. La méthode DSSR détecte les empilements de paires de bases et détermine la famille d'interactions.

Cependant, comme évoqué précédemment, ces méthodes de détermination des structures 3D d'ARN peuvent être difficiles à réaliser, et sont longues et coûteuses. Ainsi, de nombreuses études ont cherché des alternatives, en essayant de prédire les interactions constituant les structures tridimensionnelles à partir de la séquence, ou de la structure secondaire.

1.7 La prédiction des structures d'ARN

La prédiction des structures d'ARN consiste à tenter de déterminer les interactions apparaissant au sein des structures, à partir des informations de séquence, et parfois d'informations supplémentaires, comme la structure secondaire, pour en déduire la forme globale en 3D et parfois la position des atomes correspondante.

Nous allons décrire dans cette section 1.7, les principales familles de méthodes existant actuellement pour prédire les structures secondaires et tertiaires d'ARN.

1.7.1 La prédiction des structures secondaires avec ou sans pseudonœud

La structure secondaire sans pseudonœud, de par ses propriétés analogues à celles d'un arbre, s'est révélée être une étape de modélisation très utile, car elle permet l'utilisation d'approches algorithmiques sur les arbres, et fournit des informations essentielles sur la structure locale de base des molécules d'ARN.

La structure secondaire la plus probablement observée à l'équilibre thermodynamique est la structure secondaire la plus stable en énergie libre[112]. Ainsi, l'une des approches de résolution du problème de prédiction de la structure secondaire vise à rechercher la structure secondaire d'énergie libre minimum (*minimum free energy*, MFE).

Les premières méthodes étudiées historiquement [81, 129] utilisent alors la programmation dynamique pour résoudre ce problème, partant du principe que la structure la plus stable en énergie est constituée des sous-structures les plus stables en énergie.

Différents modèles d'énergie ont alors été étudiés pour représenter au mieux l'énergie associée au repliement d'une structure secondaire. Comme les paires de nucléotides stabilisent la molécule, dans le premier algorithme historique de Nussinov et Jacobson [81] était maximisé le nombre de paires de bases. Des modèles d'énergie de plus en plus précis ont ensuite été utilisés, comme le modèle du plus proche voisin, utilisé dans l'algorithme de Zuker et Stiegler [129, 128], très populaire pour la prédiction de la structure secondaire d'énergie libre minimum. Ce modèle calcule la somme des contributions énergétiques de chaque empilement de paires de bases et de chaque boucle, plutôt que celui des paires de bases prises individuellement. Des paramètres expérimentaux avaient été auparavant calculés pour chacun de ces éléments, et de nouveaux paramètres seront calculés par la suite [114]. Ces algorithmes permettent d'obtenir une structure secondaire minimisant l'énergie libre avec une complexité en temps dans $O(n^3)$, avec n le nombre de nucléotides composant la séquence primaire en entrée.

Des méthodes se basant sur le même principe de programmation dynamique permettent la prédiction de certains pseudonœuds (voir par exemple [93, 91]). Ces algorithmes ont une complexité en temps plus élevée pour prendre en compte les cas d'appariements formant des pseudonœuds.

Notons que le problème de prédiction des pseudonœuds de tout type a été démontré comme étant NP-complet [3].

De plus, le repliement d'une molécule d'ARN est aujourd'hui considéré comme un phénomène stochastique, ce qui signifie que la molécule se replie en un ensemble de structures, chacune ayant une probabilité d'apparition particulière. Des modèles probabilistes ont ainsi été développés pour représenter ce phénomène, se basant sur l'équilibre de Boltzmann. Les premières méthodes permettaient alors de calculer une structure centroïde entre toutes les structures probables [29]. L'un des logiciels les plus utilisés aujourd'hui, RNAfold [68], calcule une fonction de partition, avec l'algorithme de McCaskill [74], pour calculer les probabilités d'appariements des nucléotides de la séquence, et ainsi trouver la structure maximisant la précision attendue (MEA). Cette méthode a une complexité asymptotique en $O(n^3)$, avec n le nombre de nucléotides de la séquence. Plus récemment, des méthodes utilisant des réseaux de neurones ont également été testées [99, 94].

Les molécules d'ARN peuvent également se replier selon différentes structures secondaires en fonction de leur environnement, constitué d'autres acteurs tels que les protéines et d'autres ARN avec qui elles peuvent interagir. C'est la raison pour laquelle des structures secondaires sous-optimales en énergie ont été également recherchées, par des méthodes de programmation dynamique [127, 121], et par des méthodes stochastiques [30, 68].

Des approches comparatives ont également permis des prédictions de structures secondaires d'ARN de bonne qualité, se basant sur le fait que des molécules d'ARN ayant une fonction commune partagent aussi une structure commune. Ces molécules *homologues*, c'est-à-dire dérivées d'un ancêtre commun, ont conservé une structure commune au cours de l'évolution. Certaines méthodes [72, 113] visent alors à effectuer un repliement et un alignement simultanés d'une famille de séquences homologues, selon l'algorithme de Sankoff [95], pour en déduire une structure secondaire consensus. Cependant, ces méthodes nécessitent que des informations d'homologie soient disponibles, ce qui n'est pas toujours le cas.

D'autres méthodes enfin utilisent la programmation linéaire en nombres entiers pour représenter et résoudre le problème de structure secondaire d'énergie libre minimum [100, 60].

Les résultats de ces différentes méthodes peuvent alors être utilisés en entrée des méthodes de prédiction de structures tridimensionnelles.

1.7.2 La prédiction des structures tridimensionnelles d'ARN

La prédiction d'une structure tertiaire d'ARN a fait l'objet de nombreuses études, pouvant être décomposées en plusieurs catégories :

- les méthodes utilisant la dynamique moléculaire. Dans ce cas, on cherche à simuler l'évolution temporelle d'un système moléculaire. Ces méthodes prennent en entrée la séquence primaire, une structure secondaire parfois (NAST [50]) et également des contraintes supplémentaires. Des

structures candidates sont ensuite produites par simulation de Monte Carlo (SimRNA [11]) ou de dynamique moléculaire (YUP [109]), et discriminées à l'aide d'une fonction d'énergie (souvent champ de force ou inverse de Boltzmann). Puis, la structure de plus faible énergie, ou le centroïde parmi les structures de plus faible énergie, est choisi. Certaines méthodes prennent en entrée une représentation atomique tandis que d'autres considèrent une représentation à gros grain, où plusieurs atomes représentent un nucléotide (iFoldRNA [103], HiRE-RNA [84]). Ces méthodes, longues à l'exécution, permettent d'obtenir des structures tridimensionnelles de bonne résolution, mais ne sont applicables que pour des molécules de quelques nucléotides, ou de quelques dizaines de nucléotides tout au plus.

- les méthodes utilisant une représentation à gros grain pour tenter de prédire la structure tridimensionnelle dans sa globalité, souvent à partir d'une structure secondaire connue. Par exemple, RNAJAG/RAGTOP [55] représente les multiboucles par des graphes à gros grain, calcule ensuite des potentiels statistiques à partir de mesures de torsion notamment, et peut ainsi prédire le positionnement des multiboucles en 3D par un algorithme de recuit simulé. La méthode ERNWIN [54] représente les hélices comme des cylindres, et calcule des potentiels basés sur les forces physiques et sur des structures déjà connues, pour déterminer la structure la plus probable par la méthode probabiliste de Monte-Carlo.

  D'autres méthodes, comme GARN [14, 13] utilisent des algorithmes de théorie des jeux. Dans la méthode GARN, la modélisation se fait par un graphe dans lequel un sommet correspond à une hélice de 5 paires de bases ou moins, ou bien à une boucle. Les multiboucles sont quant à elles représentées par deux sommets. Les arêtes du graphe relient des éléments adjacents dans la structure secondaire. Les sommets du graphe sont alors des joueurs qui peuvent se déplacer sur une grille 3D triangulaire, en fonction de stratégies définies par des algorithmes de minimisation de regret.

  Ces méthodes ne permettent cependant que d'obtenir des formes 3D peu précises.

- les méthodes utilisant des fragments d'ARN ou des motifs structuraux connus pour guider la prédiction. Ces méthodes pourront par exemple utiliser les motifs stockés dans les bases de données évoquées plus haut. Nous allons brièvement présenter quelques-unes de ces méthodes :

  - La méthode FARNA [26] utilise des fragments de chaîne de 1 à 3 nucléotides issus de la structure IFFK (dans la PDB). Elle modélise un nucléotide par une sphère située au centre de la base azotée, et par une méthode de Monte-Carlo, essaie de reconstituer la structure 3D à partir de ces fragments en prenant en compte les torsions, collisions et potentiels d'appariement.

- La méthode MC-Sym/MC-fold [83] cherche également à assembler des fragments pour obtenir la structure 3D globale. La méthode MC-Sym prédit la structure secondaire à partir de la séquence puis la méthode MC-fold recherche, par une méthode de Monte-Carlo également, l'assemblage de fragments ayant la plus faible contribution énergétique. Cette méthode peut également être couplée avec RNA-MoIP [124], utilisant la programmation en nombres entiers, pour améliorer une structure secondaire prédite en y insérant des motifs structuraux incluant des interactions non canoniques (ceux de RNA3DMotifs [31]), à partir de similarités de séquence.
- la méthode JAR3D [126] calcule les distributions de probabilités des séquences des motifs de boucle de RNA 3D Motif Atlas pour déterminer à partir d'une séquence, quel type de motifs de boucle sera le plus probable.
- de manière plus étendue, les méthodes RMDetect [24], puis BayesPairing [96] calculent un réseau Bayésien à partir d'un motif local quelconque. Ce réseau est un graphe orienté dans lequel les sommets correspondent aux nucléotides et chaque arc à une interaction entre deux nucléotides (canoniques, non canoniques ou empilement), et chaque sommet est associé à une table de probabilité, dont les valeurs pour chaque type de nucléotides (A,C,G,U) sont apprises à partir de séquences formant le motif structural local en question. A partir de ces réseaux Bayésiens et d'une séquence d'ARN en entrée, ces deux méthodes identifient des sous-séquences candidates dans la séquence pouvant se replier selon les motifs structuraux étudiés. Une nouvelle version de BayesPairing [97] améliore la complexité algorithmique de la méthode initiale en utilisant une décomposition arborescente du réseau Bayésien.
- Parmi ces méthodes, on peut également nommer RNA Composer [88], qui prend en entrée une séquence et une structure secondaire, découpe la structure secondaire en fragments de boucles et d'hélices, et recherche ces fragments dans une base de données pour les assembler en 3D. Cette méthode s'inspire des méthodes de traduction automatique en linguistique. La méthode VFold-3D [19], quant à elle, utilise des modèles complexes d'énergie pour prédire une structure secondaire à partir d'une séquence, puis pour construire des fragments 3D à partir de cette structure secondaire avant de les assembler. Enfin, la méthode 3dRNA [119] prend en entrée une structure secondaire et la découpe également en éléments de structure secondaire (boucles et hélices). La structure secondaire est ensuite représentée par un arbre dans lequel un noeud correspond à un élément de structure secondaire. Chaque noeud est associé à une structure 3D possible recherchée dans les bases de données existantes, puis un parcours de l'arbre

est effectué pour assembler les éléments.

Ces trois ensembles de méthodes tentent ainsi de résoudre le problème de prédiction des structures 3D d'ARN de différentes manières. Certaines de ces méthodes permettent de prédire la position précise des atomes dans l'espace pour de très petites molécules, et d'autres permettent de prédire une forme 3D très générale et très imprécise pour de plus grosses molécules. D'autres encore permettent de prédire certaines interactions non canoniques organisées en motifs structuraux locaux. Par contre, les interactions non canoniques à longue distance sur la séquence primaire restent difficiles à prédire, et pour les molécules possédant un grand nombre de nucléotides, comme les ARN ribosomiques par exemple, les structures 3D ne sont pas prédictibles.

1.7.3 La prédiction du motif A-minor

Le motif A-minor en particulier, reste difficile à prédire par les méthodes évoquées précédemment. Cela est dû au fait qu'il implique des interactions non canoniques, qu'il relie souvent des nucléotides éloignés sur la séquence primaire et qu'il possède peu d'information de séquence. Une étude de 2021[101] a étudié la classification et la prédiction de ce motif en particulier, en prenant en entrée des informations sur les éléments de structure secondaire constituant le motif et des informations de séquence. Cette méthode de prédiction se base sur une méthode de machine learning, l'algorithme de RandomForest [67]. Par cette méthode, les auteurs parviennent à prédire certains motifs A-minor particuliers, apparaissant entre des SSE qui sont proches sur la séquence, et apparaissant conjointement avec un pseudoœeud. Les autres motifs A-minor restent imprédictibles avec leur méthode.

L'une des questions pouvant se poser vis-à-vis des motifs longue distance, et en particulier du motif A-minor, concerne la hiérarchie de leur formation. Les motifs à longue distance peuvent se former de manière opportuniste, car le repliement global de la molécule a rapproché dans l'espace les SSE considérés : le motif se forme alors pour stabiliser la molécule. Ou , au contraire, les motifs à longue distance se forment en premier, et cela contraint le repliement de la structure dans sa globalité. On peut supposer que les deux cas se produisent. La formation opportuniste de certaines occurrences de motifs pourrait expliquer la difficulté rencontrée pour les prédire.

1.8 Conclusion

Les structures d'ARN ont fait l'objet de nombreuses modélisations et études, qui ont permis des avancées dans leur prédiction.

Depuis plus de 40 ans, de nombreux travaux se sont intéressés au problème de prédiction de structures 3D d'ARN. La structure secondaire, premier niveau de structure de la molécule d'ARN, est aujourd'hui prédictible à

partir d'une séquence primaire, par de nombreuses méthodes utilisant dans la plupart des cas des modèles thermodynamiques d'énergie, pour trouver la (ou les) structure(s) le(s) plus stable(s) en énergie.

La prédiction de la structure 3D présente davantage de difficultés en raison des interactions supplémentaires qu'elle contient. Des méthodes utilisent des notions de dynamique moléculaire pour y parvenir, qui nécessitent alors d'importants temps de simulation, ce qui exclut la possibilité de les utiliser sur des molécules de plus de quelques dizaines de nucléotides. D'autres méthodes utilisent des modélisations de graphes à gros grain pour prédire la structure 3D dans leur globalité, sans nécessairement connaître les interactions exactes permettant l'obtention de cette structure. Cependant, les résultats de ces méthodes sont trop imprécis pour permettre leur utilisation par des biologistes. D'autres encore utilisent les motifs structuraux, qui sont des sous-structures apparaissant de manière récurrente, pour guider la prédiction globale. Un certain nombre de motifs structuraux locaux, c'est-à-dire apparaissant au sein d'un élément de structure secondaire, sont aujourd'hui prédictibles. En revanche, les motifs apparaissant entre plusieurs SSE sont plus difficiles à identifier et à prédire. Des motifs à longue distance récemment découverts n'ont pas encore été étudiés, et d'autres déjà étudiés comme le motif A-minor ne sont pas prédictibles par les méthodes actuelles.

C'est la raison pour laquelle nous avons décidé de nous intéresser aux motifs à longue distance. Nous avons souhaité les étudier et les classifier pour apporter des informations sur leur contexte structural, dans le but de déterminer s'il est possible d'utiliser ces informations par la suite dans une approche prédictive.

Chapitre 2

Modèle de graphes et algorithmes de similarité

2.1 Introduction

Dans ce chapitre, nous allons définir formellement l'ensemble des modèles et algorithmes, permettant de représenter et comparer les contextes structuraux d'un motif à longue distance d'ARN, et qui seront utilisés dans les chapitres suivants.

Le modèle de graphes que nous construisons tout d'abord vise à capturer l'environnement local autour d'un motif d'ARN en s'intéressant uniquement aux interactions canoniques et non canoniques qui apparaissent dans cet environnement, sans avoir besoin de connaître la position dans l'espace de chaque atome le constituant.

Ainsi, nous présenterons plusieurs graphes. Tout d'abord, un graphe d'ARN, permettant de représenter les interactions canoniques et non canoniques au sein d'une structure d'ARN, constituant donc la topologie de la structure d'ARN comme évoqué dans le chapitre précédent. Puis, nous présenterons des graphes d'ARN particuliers, permettant de représenter les motifs à longue distance apparaissant dans les structures d'ARN. A partir de ces graphes, nous définirons le graphe permettant de représenter le contexte structural d'un motif structural, que l'on appellera *k-extension*.

Notre modèle vise également à pouvoir comparer ces environnements non pas à l'identique, mais en autorisant de légères différences dans le nombre de nucléotides ou d'interactions impliqués. En effet, il arrive qu'une différence concernant un seul nucléotide dans une boucle ou une hélice de structure secondaire ne change pas drastiquement la structure tridimensionnelle. Dans ce but, nous présenterons un graphe dérivé d'une k-extension, appelé *k-extension contractée*, permettant d'introduire cette flexibilité.

Après avoir défini, dans la section 2.2, le modèle de graphe que nous utilisons, nous présenterons des méthodes permettant de comparer les graphes entre eux dans la section 2.3. Nous verrons qu'il est possible de se ramener à une recherche de sous-graphe commun maximisant le nombre

d'arêtes (Maximum Common Edge Subgraph), souvent abrégé MCES [90], et nous nous baserons donc sur une méthode exacte de la littérature permettant de résoudre ce problème dans des graphes représentant de petites molécules. Nous présenterons également une méthode heuristique, dans la section 2.4, développée spécifiquement pour ce modèle, permettant de diminuer le temps d'exécution.

A partir de la définition d'un sous-graphe commun maximum à deux k-extensions, nous définirons également, dans la section 2.5, la notion de sous-graphe commun maximum à un sous-ensemble de k-extensions, permettant de caractériser ce sous-ensemble.

Toutes les notations sur les graphes utilisées dans ce chapitre sont issues de [7].

2.2 Modélisation du contexte structural topologique de motifs à longue distance

2.2.1 Définitions préalables

Cette partie permet d'introduire quelques définitions préalables de théorie des graphes, pour représenter les interactions au sein d'une structure secondaire enrichie d'ARN, et en particulier au sein des motifs structuraux. Nous allons tout d'abord définir un graphe d'ARN, permettant de représenter tout ou partie d'une structure secondaire enrichie, de manière assez classique, comme par exemple dans [31, 92], avec les sommets correspondant aux nucléotides et les arcs aux interactions entre nucléotides, mais avec un certain nombre de particularités, comme par exemple des types associés aux sommets différents des types de nucléotides (A,C,G,U). Nous définirons également des sous-graphes représentant uniquement la séquence primaire d'une molécule ou uniquement les interactions canoniques et non canoniques, qui nous aideront à définir le contexte structural d'un motif.

Graphe d'ARN

Définition 2.2.1 *Un **graphe d'ARN** est un graphe connexe orienté $G = (V, A_P, A_H)$, avec A_P et A_H deux ensembles d'arcs.*

Ce graphe représente tout ou partie d'une structure secondaire enrichie d'ARN. Les sommets correspondent aux nucléotides, les arcs de A_P aux liaisons covalentes de la séquence primaire et les arcs de A_H aux interactions canoniques et non canoniques entre nucléotides non consécutifs sur la séquence (qui sont de type hydrogène). Un exemple est présenté en Figure 2.1.

Un graphe d'ARN possède les caractéristiques suivantes :

- L'ensemble A_P des arcs induit un ou plusieurs chemin(s) deux à deux sommets-disjoints. Chaque chemin forme une séquence primaire de molécule d'ARN orientée de l'extrémité 5' à l'extrémité 3'. L'ensemble A_P

peut par exemple induire plusieurs chemins dans le cas d'un motif structural, comme nous le verrons dans la définition suivante (Définition 2.2.2).

- Pour chaque arc $(x, y) \in A_H$, on définit un type $\boldsymbol{t((x, y))}$, avec $t((x, y))$ élément de$\{ CAN, cWW, tWW, cSS, tSS, cSH, tSH, cHS, tHS, cHH, tHH, cSW, tSW, cWS, tWS, cHW, tHW, cWH, tWH \}$.

  Ce type indique la famille à laquelle appartient l'interaction, selon la nomenclature de Leontis–Westhof [62] (voir Table 1.1). En particulier, les interactions canoniques sont annotées par un type particulier que l'on notera CAN.

  Lorsqu'un arc $(x, y) \in A_H$ existe dans G, l'arc (y, x) existe également. Cependant, le type de (x, y) peut être différent du type de (y, x) car les interactions non canoniques ne sont pas toutes symétriques. Par exemple, dans une interaction de type *cis* Watson-Crick/Sugar (cWS), la face Watson-Crick d'un nucléotide et la face Sugar de l'autre nucléotide sont impliquées. Dans ce cas, $t((x, y))$ vaudra cWS et $t((y, x))$ vaudra cSW, si le sommet x correspond au nucléotide dont la face Watson-Crick est impliquée dans l'interaction, et le sommet y au nucléotide dont la face Sugar est impliquée dans l'interaction.

- Pour chaque sommet $x \in V$, on définit également un type $\boldsymbol{\rho(x)}$ en fonction de ses prédécesseurs et successeurs. Ce type aura de l'importance pour la recherche d'isomorphisme de graphes (voir section 2.3), qui tiendra compte des types d'arcs et de sommets.

  - $\rho(x) = 0$ si x n'est extrémité d'aucun arc de A_H. Il peut en revanche posséder un arc entrant et/ou un arc sortant de A_P.
  - $\rho(x) = C$ si x possède un arc sortant $(x, y) \in A_H$ tel que $t((x, y)) = CAN$, et aucun arc sortant $(x, y') \in A_H$ tel que $t((x, y')) \neq CAN$.
  - $\rho(x) = N$ si x possède au moins un arc sortant $(x, y) \in A_H$ tel que $t((x, y)) \neq CAN$ et aucun arc sortant $(x, y') \in A_H$ tel que $t((x, y')) = CAN$.
  - $\rho(x) = M$ si x possède un arc sortant $(x, y) \in A_H$ tel que $t((x, y)) = CAN$ et au moins un arc sortant $(x, y') \in A_H$ tel que $t((x, y')) \neq CAN$.

- Pour chaque sommet $x \in V$ tel que $\rho(x) = C$, on appelle **voisin canonique** de x, le sommet $y \in V$ tel que $(x, y) \in A_H$ et $t((x, y)) = CAN$. Notons que si $t((x, y)) = CAN$, alors l'arc (y, x) existe et est de type $t((y, x)) = CAN$ également. Par définition, le sommet y existe et est unique car un nucléotide ne peut pas être impliqué dans plus d'une interaction canonique.

Dans ce graphe d'ARN, le type de nucléotides (A, C, G, U) n'est pas pris en compte, car nous souhaitons nous intéresser au contexte structural

uniquement. De plus, nous pouvons noter qu'un graphe d'ARN est un graphe simple car deux mêmes nucléotides ne peuvent être reliés par deux interactions canoniques et/ou non canoniques différentes.

Motif et occurrence de motif

Nous représentons les motifs structuraux des molécules d'ARN par des graphes d'ARN particuliers, que nous appellerons **motifs**. Par exemple, le motif A-minor de type I/II sera représenté comme suit (Figure 2.1) :

Définition 2.2.2 *Le **motif A-minor** est un graphe d'ARN, noté $F_A = (V_A, A_{P,A}, A_{H,A})$, tel que :*

- $V_A = \{1, 2, 3, 4, 5, 6\}$ *avec les types ρ suivants :*
  $\rho(1) = \rho(3) = N$ *et* $\rho(2) = \rho(4) = \rho(5) = M$ *et* $\rho(6) = C$
- $A_{H,A} = \{(1,2), (2,1), (1,5), (5,1), (2,5), (5,2), (3,4), (4,3), (4,6), (6,4)\}$ *avec :*
  - $t((1,2)) = t((2,1)) = cSS$
  - $t((1,5)) = t((5,1)) = tSS$
  - $t((2,5)) = t((5,2)) = CAN$
  - $t((3,4)) = t((4,3)) = cSS$
  - $t((4,6)) = t((6,4)) = CAN$
- $A_{P,A} = \{(3,1), (2,4), (6,5)\}$

Nous définissons également une occurrence d'un motif F comme un sous-graphe partiel particulier d'un graphe d'ARN.

Définition 2.2.3 *Etant donné un graphe G d'ARN, une **occurrence de motif F** est un sous-graphe partiel de G, noté $O = (V^{\mathscr{O}}, A_P^{\mathscr{O}}, A_H^{\mathscr{O}})$, isomorphe à un motif F, en respectant les types d'arcs et de sommets.*

Nous noterons les sommets de $V^{\mathscr{O}}$ de la même façon que les sommets du motif F, par facilité d'écriture. Par exemple, pour le motif A-minor, les sommets de $V^{\mathscr{O}}$ seront notés 1,2,3,4,5,6 (exemple en Figure 2.2a).

Graphe de séquence

Définition 2.2.4 *Etant donné un graphe G d'ARN et une occurrence de motif O dans G, le **graphe de séquence** $G^{\mathscr{O}, A_P -} = (V, A_P \setminus A_P^{\mathscr{O}})$ est le sous-graphe couvrant de G ne comportant aucun arc de A_H et tous les arcs de A_P sauf ceux appartenant à l'occurrence du motif O.*

Ce graphe de séquence $G^{\mathscr{O}, A_P -}$ consiste donc en un ensemble de composantes (faiblement) connexes, chacune étant un chemin (Figure 2.2b). Chaque chemin a pour extrémité initiale et/ou finale un sommet de l'occurrence de motif O. Ce graphe nous permettra de définir le contexte structural d'une occurrence de motif (voir section 2.2.2).

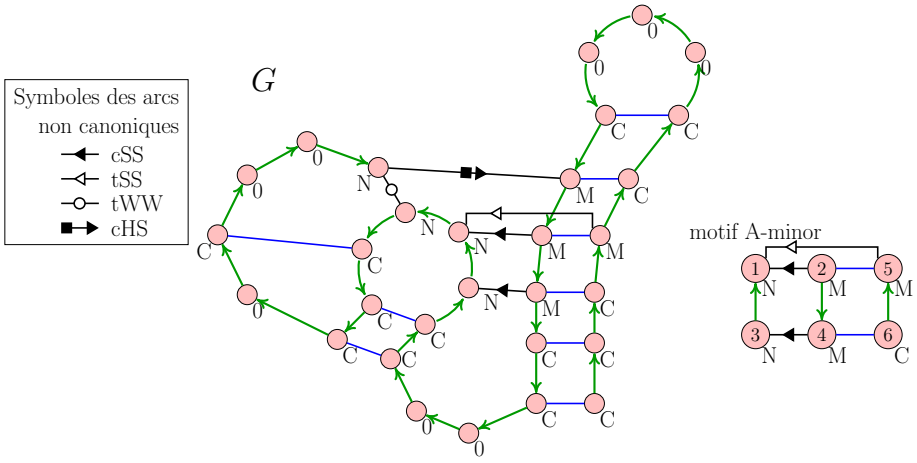


Figure 2.1 – Exemples de graphes d'ARN : un graphe d'ARN G fictif et le motif A-minor. Les arcs de A_P sont représentés en vert. Deux arcs symétriques de A_H sont représentés par un trait plein bleu ou noir. Les arcs de type canonique sont bleus et les arcs de type non canonique sont noirs, avec les symboles de la nomenclature de Leontis–Westhof [62], indiqués en légende également. Chaque sommet est annoté par son type ρ.

Graphe d'interactions

Définition 2.2.5 *Etant donné un graphe G d'ARN, le **graphe d'interactions** $G^{A_H} = (V, A_H)$ est le sous-graphe couvrant de G comportant tous les arcs de A_H mais aucun arc de A_P.*

Ce graphe est un graphe orienté symétrique, qui, dans la plupart des cas, n'est pas connexe. Contrairement au graphe de séquence, il ne dépend pas d'une occurrence de motif. Comme nous le verrons dans la partie 2.3, c'est dans ce type de graphe que nous pourrons rechercher une sous-structure commune. Un exemple de graphe d'interactions est présenté en Figure 2.3.

2.2.2 Définition d'une k-extension

Nous nous intéressons donc au contexte structural topologique d'un motif d'ARN. A partir des définitions précédentes, nous définissons ce contexte structural comme un sous-graphe d'un graphe d'ARN induit par les sommets d'une occurrence de motif, ainsi que par les sommets se trouvant à une distance strictement inférieure à k d'un sommet de l'occurrence de motif. Nous appellerons ce graphe la *k-extension* d'une occurrence de motif. Notons qu'un sous-ensemble de sommets de l'occurrence de motif est choisi pour construire la k-extension.

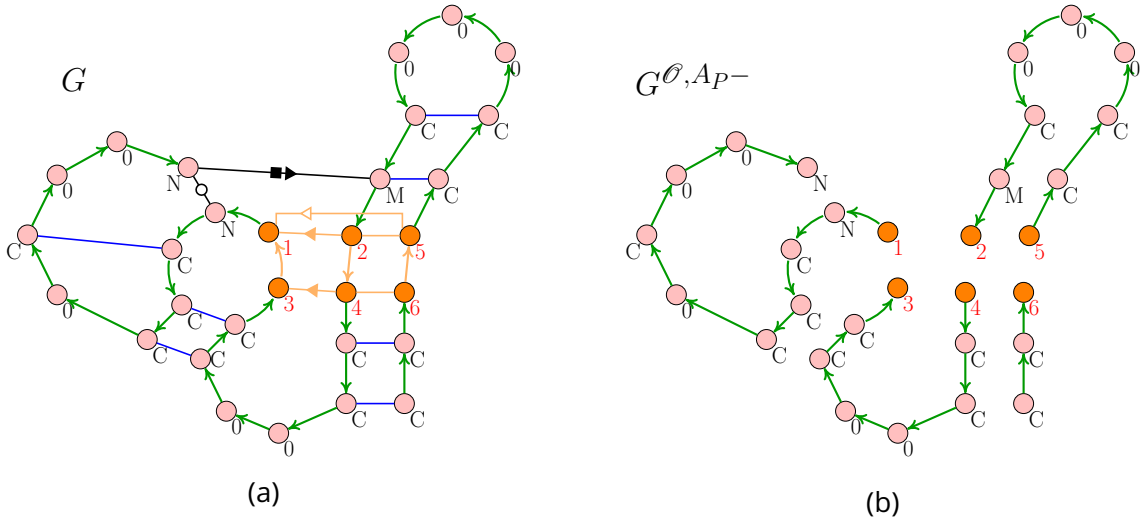


Figure 2.2 – En (a), le graphe d'ARN G de la Figure 2.1 avec une occurrence de motif A-minor en orange. En (b), le graphe de séquence $G^{\mathscr{O},A_P-}$ de G. Les sommets n'appartenant pas à l'occurrence de motif sont annotés par leur type, et les sommets de l'occurrence de motif sont annotés par leur numéro dans le graphe (1,2,3,4,5,6).

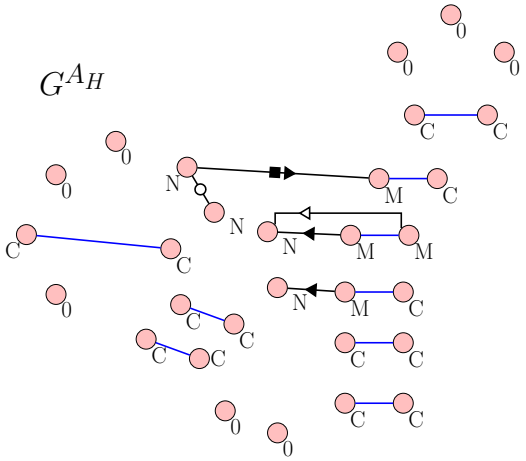


Figure 2.3 – Graphe d'interactions G^{A_H} du graphe d'ARN G de la Figure 2.1. Les sommets sont annotés par leur type ρ.

Définition 2.2.6 *Etant donnés une occurrence de motif O, un graphe G d'ARN, un sous-ensemble S de ses sommets et un entier k, la* **k*-extension d'une occurrence de motif* O *dans le graphe* G *selon* S** *est le sous-graphe $G_O = (V_O, A_{P,O}, A_{H,O})$ de G induit par les trois ensembles de sommets suivants :*

- *l'ensemble $V^{\mathscr{O}}$ des sommets de l'occurrence O (voir Définition 2.2.3)*
- *l'ensemble $V_k^{\mathscr{O}}$ des sommets se trouvant à une distance strictement inférieure à k de l'un des sommets de S, dans le graphe de séquence $G^{\mathscr{O},A_P-}$ (voir Définition 2.2.4)*
- *l'ensemble $V_k^{\mathscr{O}+}$ des sommets successeurs ou prédécesseurs d'un des sommets de $V_k^{\mathscr{O}}$, dans le graphe d'interactions G^{A_H} (voir Définition 2.2.5), sauf ceux appartenant à $V^{\mathscr{O}}$.*

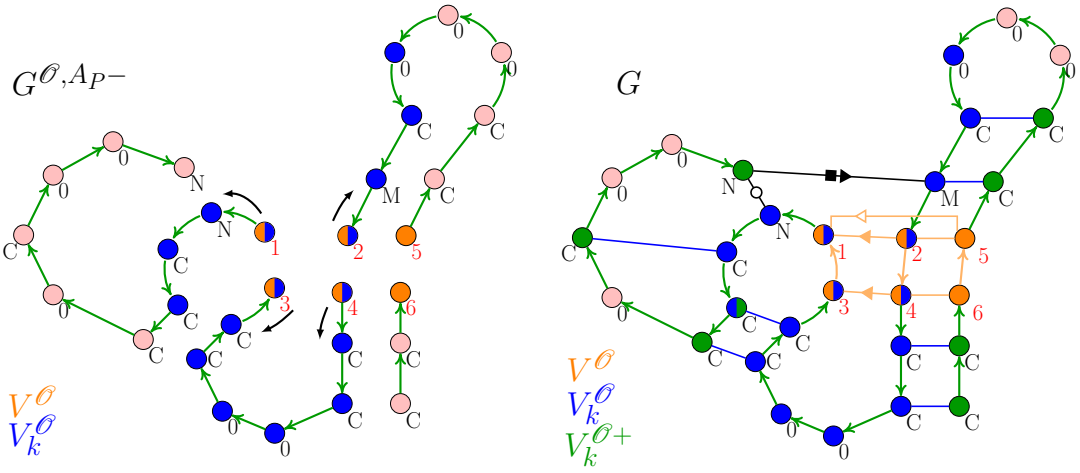
En Figure 2.4, sont présentés les trois sous-ensembles de sommets sur l'exemple du graphe d'ARN G et de l'occurrence de motif A-minor de la Figure 2.2. En Figure 2.5, est présentée la k-extension G_O obtenue, avec $k = 4$ et $S = \{1, 2, 3, 4\}$.

L'ensemble S contient les sommets de l'occurrence de motif O à partir desquels on souhaite étendre le motif (voir exemple dans la Figure 2.4). Par exemple, pour le motif A-minor, l'extension ne se fait qu'à partir des 4 premiers sommets, c'est-à-dire l'ensemble $\{1, 2, 3, 4\}$ (voir les flèches noires dans la Figure 2.4). Ce choix sera expliqué dans le chapitre 3 (section 3.2.2).

Notons également que les sous-ensembles de sommets $V^{\mathscr{O}}$ et $V_k^{\mathscr{O}}$ d'une part, et $V_k^{\mathscr{O}}$ et $V_k^{\mathscr{O}+}$ d'autre part peuvent ne pas être deux à deux disjoints (voir Figure 2.4).

Les sommets de V_O dans la k-extension peuvent être couverts par plusieurs sous-ensembles, que nous appellerons **branches** et définirons de la façon suivante (Figure 2.5). D'après la définition 2.2.4, le graphe de séquence $G_O^{\mathscr{O},A_P-}$ est le sous-graphe couvrant de la k-extension G_O ne contenant aucun arc de $A_{H,O}$ et uniquement les arcs de $A_{P,O}$ n'appartenant pas à l'occurrence de motif O. Ce sous-graphe de G_O consiste en un ensemble de chemins, dont au moins une des extrémités est un sommet de l'occurrence de motif. Nous considérons les chemins dont l'une des extrémités est un sommet de l'ensemble S, notons l'ensemble de ces chemins $C = \{C_1, C_2, ..., C_n\}$ avec $n \leq |S|$, et définissons les sous-ensembles de sommets à partir de ces chemins. Les sommets de $V_k^{\mathscr{O}}$ appartenant au même chemin C_i ($i \in \{1, 2, ..., n\}$) constituent les sommets de la branche i. Les sommets de $V_k^{\mathscr{O}+}$ appartiennent à la même branche de sommets que leur(s) successeur(s) et prédécesseur(s) dans $V_k^{\mathscr{O}}$. Notons que nous excluons des branches ainsi définies les sommets de l'occurrence de motif n'appartenant pas à l'ensemble S (par exemple, les sommets notés 5 et 6 dans l'exemple de la Figure 2.4). Nous considérerons que chacun de ces sommets constitue une branche à part entière.

Dans la k-extension d'une occurrence de motif A-minor, il y a ainsi en général 6 branches (4 branches constituées de plusieurs sommets, et 2



(a) graphe de séquence $G^{\mathscr{O},A_P-}$ avec les sommets de $V^{\mathscr{O}}$ en orange et les sommets de $V_k^{\mathscr{O}}$ en bleu.

(b) graphe d'ARN G avec les sommets de $V^{\mathscr{O}}$ en orange, les sommets de $V_k^{\mathscr{O}}$ en bleu et les sommets de $V_k^{\mathscr{O}+}$ en vert.

Figure 2.4 – Sous-ensembles de sommets permettant de définir une 4-extension ($k = 4$) à partir d'une occurrence de motif O selon le sous-ensemble $S = \{1, 2, 3, 4\}$.

branches constituées d'un seul sommet), comme illustré dans la Figure 2.5.

Dans une k-extension G_O, nous définissons le second type $\boldsymbol{\rho_O}$ de sommet, tel que pour tout sommet $u \in V_O$:

- $\rho_O(u)$ est égal à u si u appartient à $V^{\mathscr{O}}$, c'est-à-dire à l'occurrence de motif
- $\rho_O(u)$ est égal à $\rho(u)$ si u appartient à $V_k^{\mathscr{O}}$ (et pas à $V_{\mathscr{O}}$),
- $\rho_O(u)$ est égal à $\perp$ si u appartient à $V_k^{\mathscr{O}+}$ (et pas à $V_k^{\mathscr{O}}$) pour différencier ces sommets des sommets de $V_k^{\mathscr{O}}$,

Dans la suite de ce chapitre, lorsque nous parlerons de type de sommet dans les k-extensions, nous ferons référence à ρ_O et non à ρ (sauf mention contraire).

2.2.3 Définition d'une k-extension contractée

Les structures d'ARN sont sujettes à des variations apparaissant au cours de l'évolution. De légères modifications, comme un seul nucléotide différent dans une boucle ou une hélice de structure secondaire, peuvent ne pas changer drastiquement la structure tridimensionnelle de la molécule considérée. C'est la raison pour laquelle nous présentons dans cette partie une représentation contractée du contexte structural, permettant de représenter de manière similaire des contextes proches mais légèrement différents.

Pour cela, nous définissons d'abord la notion de *chemin contractable* qui va déterminer les sommets et les arcs qui seront contractés.

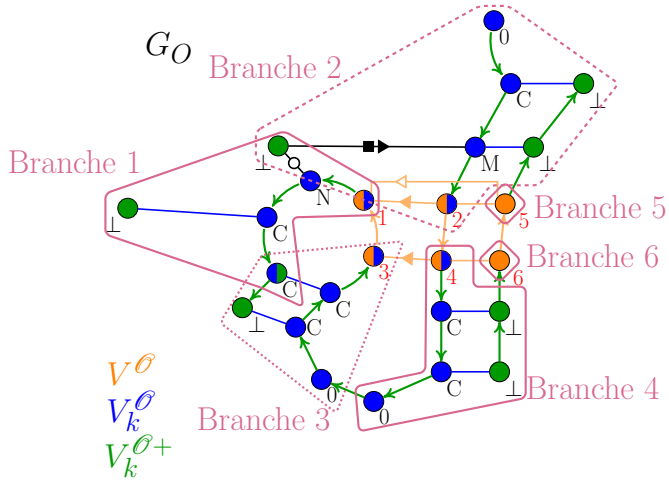


Figure 2.5 – 4-extension G_O de l'occurrence O de motif A-minor, selon l'ensemble de sommets $S = \{1, 2, 3, 4\}$. Les sommets sont annotés selon leur type ρ_O. Les 6 branches, définies selon leur position par rapport aux sommets de l'occurrence de motif, sont représentées.

Définition 2.2.7 ***Chemin contractable***

Soient une k-extension G_O, les sous-ensembles de sommets $V_k^{\mathscr{O}}$ et $V_k^{\mathscr{O}+}$ dans G_O, et le sous-graphe couvrant $G_O^{\mathscr{O},A_P-}$ de G_O ne contenant que les arcs de $A_{P,O}$ sauf ceux de l'occurrence de motif, et aucun arc de $A_{H,O}$.

*Un **chemin contractable** $\mathscr{C}$ est un chemin maximal dans le graphe $G_O^{\mathscr{O},A_P-}$, tel que :*

- *les sommets de $\mathscr{C}$ sont tous de type $\rho_O = C$ ou tous de type $\rho_O = 0$ et appartiennent tous à l'ensemble de sommets $V_k^{\mathscr{O}}$ ou tous à l'ensemble de sommets $V_k^{\mathscr{O}+}$,*
- *et si les sommets sont tous de type $\rho_O = C$, alors les voisins canoniques de ces sommets (voir définition 2.2.1) induisent aussi un chemin contractable dans $G_O^{\mathscr{O},A_P-}$.*

Les chemins contractables contiennent des sommets qui ne sont l'extrémité d'aucun arc de $A_{P,O}$ (type 0) ou bien des sommets qui possèdent des arcs entrants et/ou sortants de type covalent et canonique uniquement (type C). Ces ensembles de sommets représentent les boucles et les hélices de structure secondaire. Des exemples de chemins contractables sont présentés en Figure 2.6.

Définition 2.2.8 ***k-extension contractée d'une occurrence de motif*** O

Soient une k-extension G_O et le sous-graphe couvrant $G_O^{\mathscr{O},A_P-}$ de G_O.

*La **k-extension contractée**, notée $\tilde{\boldsymbol{G}}_{\boldsymbol{O}}$, est le graphe dérivé de la k-extension G_O, dans lequel les sommets de chaque chemin contractable de $G_O^{\mathscr{O},A_P-}$ sont contractés en un seul sommet. Si ces sommets sont de type C, leurs voisins canoniques induisent aussi un chemin contractable dans $G_O^{\mathscr{O},A_P-}$ (selon la définition 2.2.7) et seront donc*

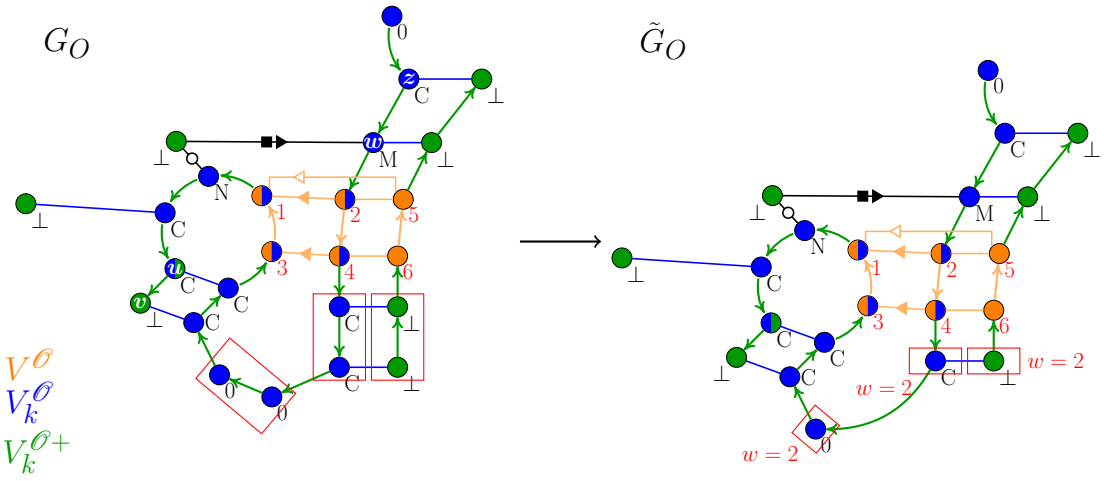


Figure 2.6 – Construction de la 4-extension contractée $\tilde{G}_O$ (à droite) à partir de contractions de sommets de la 4-extension G_O (à gauche). Les chemins contractables sont entourés en rouge dans G_O. Le chemin du sommet u au sommet v n'est pas un chemin contractable car seul le sommet u appartient à $V_k^{\mathscr{O}}$. De la même façon, les sommets w et z n'induisent pas un chemin contractable car ils ne sont pas de même type (C pour z et M pour w). Leurs voisins canoniques n'induisent donc pas non plus de chemin contractable. Le poids w des sommets issus de contractions est indiqué dans $\tilde{G}_O$. Le type ρ_O des sommets est indiqué.

contractés également. Dans ce cas, les deux sommets contractés sont reliés par un arc de type canonique.

Chaque sommet contracté dans $\tilde{G}_O$ possède le même type ρ_O que les sommets dont il est issu dans G_O. Le nombre de sommets de G_O regroupés dans $\tilde{G}_O$ en un seul sommet $x \in \tilde{V}_O$ est appelé **poids de x** et noté $w(x)$. Un exemple de graphe $\tilde{G}_O$ est présenté en Figure 2.6.

Comme précisé dans la définition 2.2.7, les chemins contractables sont maximaux au sens de l'inclusion, ce qui signifie qu'un ensemble de sommets appartenant à un chemin contractable ne peut pas être inclus dans un ensemble plus grand de sommets appartenant également à un chemin contractable. Ainsi, le graphe $\tilde{G}_O$ que l'on obtient est unique.

Enfin, remarquons que les sommets de type $\rho_O = 0$, qui ne sont l'extrémité d'aucun arc de $A_{P,O}$, représentent des nucléotides non appariés, formant des boucles dans la structure secondaire. L'absence d'arcs de type canonique ou non canonique ici peut être une information structurale en tant que telle. Pour pouvoir prendre en compte cette information, nous ajoutons aux k-extensions, pour chaque sommet u de type $\rho_O = 0$, un sommet fictif u' lié à u par deux arcs fictifs (u, u') et (u', u) appartenant à $A_{H,O}$, et tels que $t((u, u')) = 0$ et $t((u', u)) = 0$ (ces arcs ne sont pas représentés dans la Figure 2.6).

Tous les graphes définis dans la partie 2.2 sont récapitulés dans la Table 2.1.

Nom du graphe	Description
$G = (V, A_P, A_H)$	graphe d'ARN
$O = (V^{\mathscr{O}}, A_P^{\mathscr{O}}, A_H^{\mathscr{O}})$	occurrence d'un motif F dans G (sous-graphe de G, isomorphe à un motif F)
$G^{\mathscr{O},A_P-} = (V, A_P \backslash A_P^{\mathscr{O}})$	graphe de séquence de G (sous-graphe couvrant de G ne contenant aucun arc de A_H dans G et contenant tous les arcs de A_P sauf ceux de l'occurrence de motif)
$G^{A_H} = (V, A_H)$	graphe d'interactions de G (sous-graphe couvrant de G ne contenant aucun arc de A_P dans G et contenant tous les arcs de A_H dans G)
$G_O = (V_O, A_{P,O}, A_{H,O})$	k-extension d'une occurrence de motif O dans G (sous-graphe de G)
$G_O^{\mathscr{O},A_P-} = (V_O, A_{P,O} \backslash A_P^{\mathscr{O}})$	graphe de séquence de G_O (sous-graphe couvrant de G_O ne contenant aucun arc de $A_{H,O}$ dans G_O et contenant tous les arcs de $A_{P,O}$ dans G_O sauf ceux de l'occurrence de motif)
$\tilde{G}_O = (\tilde{V}_O, \tilde{A}_{P,O}, \tilde{A}_{H,O})$	k-extension contractée d'une occurrence de motif O dans G (obtenue à partir de la contraction de certains sommets et arcs de G_O)

Table 2.1 – Récapitulatif des différents graphes définis dans la partie 2.2

2.3 Définition et calcul de similarité entre k-extensions

Après avoir modélisé la notion de contexte structural topologique par des graphes, nous allons utiliser ce modèle pour déterminer si deux contextes structuraux topologiques sont similaires. Nous nous intéressons en particulier aux similarités en termes d'interactions entre nucléotides, donc aux arcs de $A_{H,O}$ des k-extensions, car ce sont elles qui contribuent le plus à la formation de la structure tridimensionnelle.

Dans cette partie, nous allons ainsi définir le sous-graphe commun à deux k-extensions, qui maximise le nombre d'arcs de $A_{H,O}$. Ensuite, nous verrons que rechercher un tel sous-graphe commun revient en réalité à résoudre le problème de recherche d'un sous-graphe commun, à deux graphes non orientés issus de sous-graphes particuliers des k-extensions, contenant un nombre maximum d'arêtes (en anglais, *Maximum Common Edge Subgraph*, soit MCES)[38] . Après avoir décrit le problème MCES et les difficultés en termes de complexité qu'il soulève, nous détaillerons un algorithme permettant de le résoudre dans le cas général, ainsi que les contraintes particulières que nous avons dû ajouter pour l'adapter à notre recherche de sous-graphe commun. Cet algorithme procède d'une méthode de résolution exacte.

Notons que les définitions et algorithmes présentés dans cette partie s'appliquent aussi bien aux k-extensions contractées que non contractées.

Nous parlerons ainsi de *k-extensions* pour faire référence aux deux types de sous-graphes d'ARN. Des précisions seront apportées lorsque des différences de traitement sont de mise entre les deux.

2.3.1 Définition du sous-graphe commun maximum à deux k-extensions et lien avec le problème MCES

Nous cherchons ainsi à déterminer le sous-graphe commun à deux k-extensions contractées ou non contractées, possédant le plus grand nombre d'arcs de type canonique ou non canoniques.

Pour cela, nous allons d'abord définir des correspondances entre les sommets et entre les arcs des k-extensions à l'aide de fonctions sur les ensembles de sommets et les ensembles d'arcs. Puis, nous pourrons définir un isomorphisme de graphes en utilisant ces fonctions. Ensuite, nous définirons le sous-graphe commun à deux k-extensions, respectant cet isomorphisme, et maximisant le nombre d'arcs de type canonique ou non canoniques. Nous finirons par introduire une métrique de similarité entre deux k-extensions, qui est maximum lorsque le nombre d'arcs de type canonique ou non canoniques du sous-graphe commun à ces deux k-extensions est maximum.

Soient $O_1 = (V^{\mathscr{O}_1}, A_P^{\mathscr{O}_1}, A_H^{\mathscr{O}_1})$ et $O_2 = (V^{\mathscr{O}_2}, A_P^{\mathscr{O}_2}, A_H^{\mathscr{O}_2})$ deux occurrences distinctes de motif (voir définition 2.2.3). Les deux graphes O_1 et O_2 peuvent être des sous-graphes du même graphe d'ARN ou de deux graphes d'ARN différents.

Etant donné un entier k défini au préalable, on appelle $G_{O_1} = (V_{O_1}, A_{P,O_1}, A_{H,O_1})$ et $G_{O_2} = (V_{O_2}, A_{P,O_2}, A_{H,O_2})$ les deux k-extensions contractées ou non contractées obtenues à partir de O_1 et de O_2 (voir sections 2.2.2 et 2.2.3).

Définition d'un (π_s, $\pi_{æ}$)-isomorphisme de graphes

Pour définir un isomorphisme de graphes, on considère d'abord deux fonctions définies comme suit

- La fonction π_s associe à chaque sommet $u \in V_{O_1}$ de G_{O_1} l'ensemble des sommets $\pi_s(u) \subset V_{O_2}$ de G_{O_2} tel que, pour tout $v \in \pi_s(u)$:
  - u et v sont de même type ($\rho_O(u) = \rho_O(v)$) (voir sous-section 2.2.2)
  - u et v appartiennent à la même branche de sommets i ($i \in \{1, 2, ..., n\}$) (voir sous-section 2.2.2)

- La fonction $\pi_{æ}$ associe :
  - à chaque arc $(u_1, v_1) \in A_{H,O_1}$ de G_{O_1} l'ensemble des arcs $\pi_{æ}((u_1, v_1)) \subset A_{H,O_2}$ de G_{O_2} tel que pour tout arc $(u_2, v_2) \in \pi_{æ}((u_1, v_1))$, les arcs (u_1, v_1) et (u_2, v_2) aient le même type t,
  - et à chaque arc de A_{P,O_1} l'ensemble des arcs de A_{P,O_2}

Définition 2.3.1 ***($\pi_s, \pi_{œ}$)-isomorphisme***

Une bijection i entre V_{O_1} et V_{O_2} est un $(\pi_s, \pi_{œ})$-isomorphisme entre G_{O_1} et G_{O_2} si et seulement si :

- $(u, v) \in A_{H,O_1} \iff (i(u), i(v)) \in A_{H,O_2}$
- $(u, v) \in A_{P,O_1} \iff (i(u), i(v)) \in A_{P,O_2}$
- $\forall\ u \in V_{O_1}$, $i(u) \in \pi_s(u)$
- $\forall\ (u, v) \in A_{H,O_1}$, $(i(u), i(v)) \in \pi_{œ}((u, v))$
- $\forall\ (u, v) \in A_{P,O_1}$, $(i(u), i(v)) \in \pi_{œ}((u, v))$

Définition du sous-graphe commun maximum à deux k-extensions

Tout d'abord, on définit :

$G'_{O_1} = (V'_{O_1}, A'_{P,O_1}, A'_{H,O_1})$ un sous-graphe de G_{O_1} tel que G'_{O_1} contient les sommets de l'occurrence de motif O_1,

$G'_{O_2} = (V'_{O_2}, A'_{P,O_2}, A'_{H,O_2})$ un sous-graphe de G_{O_2} tel que G'_{O_2} contient les sommets de l'occurrence de motif O_2.

Rappelons que le graphe $G_{O_1}^{\mathscr{O}_1, A_P-}$ (resp. $G_{O_2}^{\mathscr{O}_2, A_P-}$), appelé graphe de séquence, est le sous-graphe couvrant de G_{O_1} (resp. G_{O_2}) ne contenant que les arcs de A_{P,O_1} (resp. A_{P,O_2}) sauf ceux de l'occurrence de motif O_1 (resp. O_2) (voir définition 2.2.4).

On définit alors une notion de compatibilité entre couples de sommets de G'_{O_1} et de G'_{O_2}, en utilisant ces sous-graphes :

Définition 2.3.2 *Soient $< u_1, v_1 >$ un couple de sommets de G'_{O_1} et $< u_2, v_2 >$ un couple de sommets de G'_{O_2}*

*Les couples de sommets $< u_1, v_1 >$ et $< u_2, v_2 >$ seront dits **compatibles** si et seulement si :*

- *$u_1 \in \pi_s(u_2)$ et $v_1 \in \pi_s(v_2)$ et,*
- *s'il existe un chemin dans $G_{O_1}^{\mathscr{O}_1, A_P-}$ allant de u_1 à v_1, alors il existe un chemin dans $G_{O_2}^{\mathscr{O}_2, A_P-}$ allant de u_2 à v_2, et*
- *s'il n'existe pas un chemin dans $G_{O_1}^{\mathscr{O}_1, A_P-}$ allant de u_1 à v_1, alors il n'existe pas de chemin dans $G_{O_2}^{\mathscr{O}_2, A_P-}$ allant de u_2 à v_2.*

Le graphe G'_{O_1} doit être $(\pi_s, \pi_æ)$-isomorphe à G'_{O_2}. Le graphe G'_{O_1} est donc un sous-graphe commun à G_{O_1} et G_{O_2}.

Le sous-graphe commun à G_{O_1} et G_{O_2} que l'on recherche est ainsi un sous-graphe G'_{O_1} de G_{O_1} tel que chaque couple de sommets soit compatible avec le couple de sommets équivalent au sens de l'isomorphisme de G'_{O_2}, et tel que le nombre d'arcs de A'_{H,O_1} soit maximum. La notion de compatibilité entre couples de sommets est nécessaire pour prendre en compte les chemins définis par les arcs de A'_{P,O_1}.

Un exemple de sous-graphe commun est présenté en Figure 2.7.

Définition de la métrique de similarité

Nous définissons également une métrique de similarité particulière pour caractériser un sous-graphe commun maximum à deux k-extensions, que nous appelons *similarité contextuelle*.

La similarité contextuelle, associée au sous-graphe commun G'_{O_1} à G_{O_1} et G_{O_2}, est définie comme suit :

$$sim(G'_{O_1}, G_{O_1}, G_{O_2}) = \frac{\sum_{(u,v) \in A'_{O_1,H} \setminus A_H^{\mathscr{O}_1}} min(w(u), w(\pi_s^{-1}(u)))}{max(\sum_{(u,v) \in A_{O_1,H} \setminus A_H^{\mathscr{O}_1}} w(u), \sum_{(u,v) \in A_{O_2,H} \setminus A_H^{\mathscr{O}_2}} w(u))} \quad (2.1)$$

La valeur $w(u)$ est égale au poids du sommet u dans les k-extensions contractées et à 1 dans les k-extensions non contractées.

Cette valeur de similarité contextuelle compte la proportion d'arcs de $A'_{O_1,H}$ se trouvant dans G'_{O_1} par rapport au nombre d'arcs maximum entre $A_{O_1,H}$ dans le graphe G_{O_1} et $A_{O_2,H}$ dans le graphe G_{O_2}. On ne prend pas en compte les arcs de $A_H^{\mathscr{O}_1}$ c'est-à-dire les arcs de l'occurrence O_1 du motif (ou O_2 comme les occurrences sont isomorphes). En effet, par définition, toute k-extension d'un graphe d'ARN contient ces arcs. Dans le cas des k-extensions contractées, chaque arc est pondéré par le minimum des poids de ses extrémités dans les deux graphes G'_{O_1} et G'_{O_2}, noté w. Cela correspond au nombre de nucléotides que représente le sommet dans les k-extensions contractées. Pour les k-extensions non contractées, ce poids sera égal à 1 pour tous les sommets. Un exemple de calcul de valeur de similarité contextuelle est présenté en Figure 2.7.

Nous pouvons noter que la valeur maximum de la métrique de similarité pour deux k-extensions peut être obtenue à l'aide de plusieurs sous-graphes communs différents.

Cette métrique de similarité permet donc de prendre en compte uniquement les arcs de type canonique ou non canoniques en commun, en les pondérant par le poids des sommets dans le cas des k-extensions contractées. Pour les k-extensions contractées, cette métrique peut être paramétrée pour prendre ou non en compte la différence qui peut exister entre les poids de sommets équivalents au sens de l'isomorphisme. En effet, l'isomorphisme de

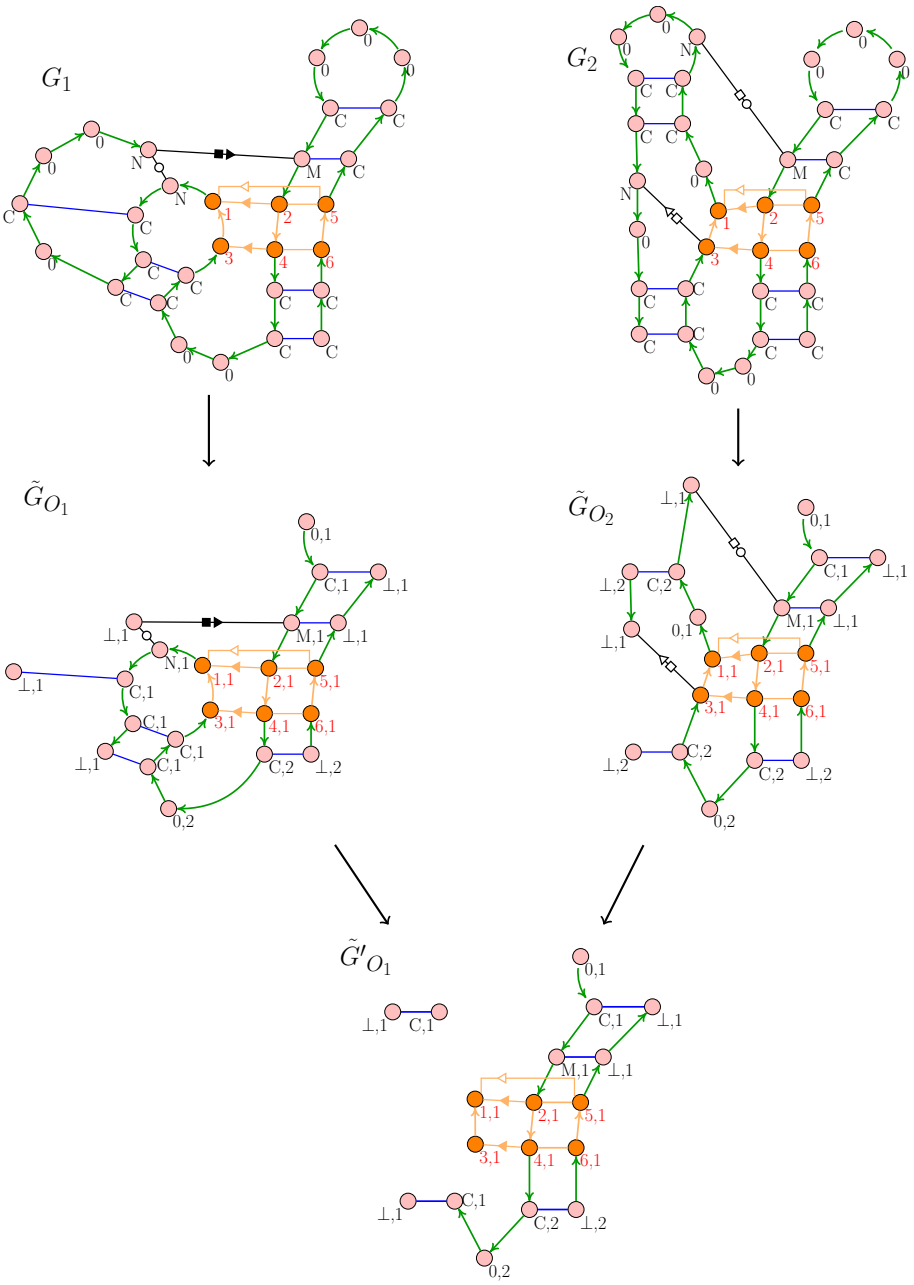


Figure 2.7 – Exemple d'un sous-graphe commun maximum à deux 4-extensions contractées. En haut sont présentés deux graphes d'ARN G_1 et G_2 possédant chacun une occurrence de motif A-minor. Chaque sommet y est annoté par son type ρ_O. Au milieu sont présentés les deux 4-extensions contractées $\tilde{G}_{O_1}$ et $\tilde{G}_{O_2}$ des occurrences de motif A-minor présentes respectivement dans G_1 et G_2. Chaque sommet y est annoté par son type ρ_O et son poids w. En bas est présenté le sous-graphe commun maximum à $\tilde{G}_{O_1}$ et $\tilde{G}_{O_2}$. La valeur de similarité contextuelle est indiquée.

graphes que l'on considère ne tient pas compte des poids des sommets. Notons que pour les k-extensions que nous étudierons dans les chapitres suivants, cette différence de poids sera petite (au maximum k), étant donné que les valeurs de k considérées seront petites également.

Ainsi, le sous-graphe commun G'_{O_1} aux deux k-extensions G_{O_1} et G_{O_2} maximisant le nombre d'arcs de A'_{H,O_1} que l'on a défini précédemment, permet également de maximiser la métrique de similarité contextuelle. Par conséquent, nous pouvons rechercher un tel graphe pour obtenir un sous-graphe commun maximisant cette métrique.

De plus, rechercher un tel sous-graphe commun à deux k-extensions revient en réalité à rechercher un sous-graphe commun maximisant le nombre d'arêtes (MCES) à deux graphes non orientés issus de nos k-extensions. C'est ce que nous allons voir dans la sous-section suivante.

Définition du sous-graphe commun maximum à des graphes non orientés issus de sous-graphes de k-extensions

Comme nous recherchons un sous-graphe commun à deux k-extensions maximisant le nombre d'arcs de type canonique et non canoniques et non le nombre d'arcs de la séquence primaire, nous considérons pour chaque k-extension $G_O = (V_O, A_{P,O}, A_{H,O})$, le sous-graphe d'interactions $G_O^{A_H} = (V_O, A_{H,O})$ (voir définition 2.2.5) ne contenant que les arcs de $A_{H,O}$ et aucun arc de $A_{P,O}$. Par définition, ce sous-graphe est orienté symétrique, c'est-à-dire que pour chaque arc $(x, y) \in A_{H,O}$, il existe un arc $(y, x) \in A_{H,O}$. Il est ainsi possible de considérer le graphe non orienté possédant les mêmes sommets et une arête entre deux sommets s'il y a un arc entre les deux sommets dans le graphe $G_O^{A_H}$, et ce, sans perdre d'information. Notons cependant que les types t de deux arcs symétriques dans $G_O^{A_H}$ peuvent être différents (voir définition 2.2.1). Il nous faut donc définir une règle pour les types d'arêtes du graphe non orienté issu de $G_O^{A_H}$. Nous définissons ce graphe non orienté de la manière suivante :

Définition 2.3.3 *Soit $G_O^{A_H} = (V_O, A_{H,O})$ le sous-graphe d'interactions d'une k-extension G_O.*

Le graphe $\boldsymbol{G_{O,E} = (V, E)}$ est le graphe non orienté sous-jacent à $G_O^{A_H}$ tel que :

- *L'ensemble des sommets V de $G_{O,E}$ est égal à l'ensemble des sommets de $G_O^{A_H}$, et chaque sommet de V possède les mêmes étiquettes que le sommet correspondant dans $G_O^{A_H}$ (type et poids, le cas échéant)*
- $E = \{[x, y] \ \forall (x, y) \in A_{H,O}\}$
  Chaque arête $[x, y] \in E$ possède un type $t([x, y])$ tel que :

- $t([x,y]) = t((x,y))$ *avec* $t((x,y))$ *le type de l'arc* (x,y) *dans* $G_O^{A_H}$ *si :*

  $\rho_O(x) \in \{C,N,M,0\}$ *et* $\rho_O(y) = \perp$ *ou si*

  $\rho_O(x) \in \{C,N,M,0\}$ *et* $\rho_O(y) \in \{C,N,M,0\}$*, et* x *appartient à une branche de plus petit numéro que* y*, ou si*

  $\rho_O(x) \in \{C,N,M,0\}$ *et* $\rho_O(y) \in \{C,N,M,0\}$*, et si* x *et* y *appartiennent à une même branche,* x *est à une distance plus petite que* y *d'un des sommets de l'occurrence de motif dans* G_O

- $t([x,y]) = t((y,x))$ *sinon*

Le sous-graphe commun à deux graphes non orientés quelconques G_1 et G_2 maximisant le nombre d'arêtes (MCES) [90], est un graphe isomorphe à un sous-graphe de G_1 et à un sous-graphe de G_2 ayant le plus grand nombre d'arêtes.

Si on considère les deux graphes non orientés $G_{O_1,E}$ et $G_{O_2,E}$ issus respectivement de deux k-extensions G_{O_1} et G_{O_2} comme définies précédemment, trouver le MCES à $G_{O_1,E}$ et $G_{O_2,E}$ revient à trouver le sous-graphe commun maximum aux deux k-extensions G_{O_1} et G_{O_2} maximisant le nombre d'arcs de A_{H,O_1} comme défini précédemment. On peut ainsi résoudre le problème de recherche d'un MCES sur ces graphes non orientés, pour trouver un sous-graphe commun maximum à deux k-extensions tel que nous le souhaitons.

2.3.2 Résolution exacte du problème MCES

Dans cette partie, nous allons brièvement présenter les difficultés en termes de complexité que soulève la recherche d'un MCES dans des graphes quelconques, puis nous détaillerons la méthode exacte de la littérature que nous avons utilisée pour résoudre le problème de recherche d'un MCES dans notre cas.

Algorithme général permettant de résoudre le problème MCES, utilisé en particulier sur des graphes représentant des molécules

Le problème MCES a été introduit en 1981 dans [10], pour modéliser le problème d'assignation de tâches, dans un contexte de programmation distribuée. Il a ensuite été démontré que le problème de décision associé à la recherche d'un MCES entre deux graphes quelconques était NP-difficile [38]. Par réduction au problème de clique maximum, Kann a montré que ce problème est APX-difficile [53], c'est-à-dire qu'il existe une constante $c > 1$ en-dessous de laquelle on ne peut pas l'approximer.

De nombreuses méthodes de résolution ont été étudiées. Nous nous sommes particulièrement intéressés aux méthodes utilisées dans la recherche de similarité entre molécules. Dans ce contexte, les graphes étudiés sont des graphes non orientés simples où les sommets représentent les atomes des molécules, et les arêtes représentent les interactions entre atomes. Parmi les méthodes exactes de la littérature, nous pouvons citer des algorithmes basés

sur la détection de clique maximum [90], des algorithmes de backtracking énumérant de manière exhaustive tous les sous-graphes possibles [58], ou encore des algorithmes se ramenant à un problème de couverture par sommets [1], ou bien résolvant le problème pour certaines classes de graphes possédant des propriétés particulières en termes de largeur arborescente et de degré [2, 122]. Nous avons choisi de nous baser sur la première de ces méthodes, car elle est la plus directement implémentable, et a montré son efficacité en pratique, bien que sa complexité asymptotique ne soit pas meilleure que celle d'un algorithme de type «force brute». De plus, elle permet de rechercher un MCES qui peut ne pas être connexe.

Cette méthode consiste en plusieurs étapes pour trouver le MCES entre deux graphes G_1 et G_2 :

- Le calcul des graphes adjoints de G_1 et de G_2, que nous appellerons *linegraphes*
- Le calcul du graphe produit de ces deux linegraphes
- La recherche d'une clique maximum dans ce graphe produit
- L'extraction d'un sous-graphe commun maximum à G_1 et G_2 à partir de cette clique maximum

Dans la partie qui suit, nous allons détailler chacune de ces étapes dans un contexte de graphe moléculaire, avant d'indiquer les adaptations que nous avons dû effectuer pour correspondre à notre modèle. En effet, un certain nombre de contraintes doivent être ajoutées.

Les définitions suivantes seront donc données sur des graphes non orientés où les sommets correspondent à des atomes, et les arêtes aux interactions entre atomes. Chaque sommet est étiqueté par un type d'atome, et chaque arête par un type de liaisons.

Définition 2.3.4 *Soit $G = (V, E)$ un graphe non orienté, avec un étiquetage sur les sommets et sur les arêtes. Le **linegraphe** du graphe G, noté $L(G) = (V_L, E_L)$, est défini par :*

- *L'ensemble V_L est égal à E : les arêtes de G sont les sommets de $L(G)$. Chaque sommet de $L(G)$ est étiqueté par la même étiquette que l'arête correspondante de E dans G.*
- *Soient deux sommets $e, f \in V_L$. L'arête $[e, f] \in E_L$ si et seulement si les arêtes de G associées aux sommets e et f sont incidentes au même sommet. On étiquette alors $[e, f]$ avec la même étiquette que ce sommet.*

Un exemple de linegraphe d'un graphe moléculaire est présenté en Figure 2.8.

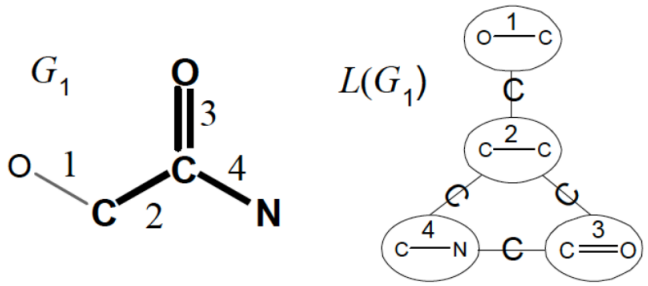


Figure 2.8 – Un graphe moléculaire G_1 et son linegraphe $L(G_1)$ (exemples issus de [90])

Définition 2.3.5 *Le graphe produit $GP = (V(GP), E(GP))$, noté $L(G_1) \diamond L(G_2)$, des graphes non orientés G_1 et G_2 est tel que :*

- $V(GP) = V(L(G_1)) \times V(L(G_2))$
- *Soient deux sommets $l_i = (e_i, f_i)$ et $l_j = (e_j, f_j)$ de GP. L'arête $[l_i, l_j] \in E(GP)$ si et seulement si les sommets e_i et f_i (resp. e_j et f_j) ont la même étiquette et si l'une des conditions suivantes est vérifiée :*
  - $[e_i, e_j] \in E(L(G_1))$ *et* $[f_i, f_j] \in E(L(G_2))$ *et les étiquettes associées à ces arêtes sont identiques*
  - $[e_i, e_j] \notin E(L(G_1))$ *et* $[f_i, f_j] \notin E(L(G_2))$

Il faut ensuite rechercher une clique maximum dans ce graphe produit. Dans le contexte de comparaison de molécules, des méthodes de résolution exacte (Branch and Bound [90] ou programmation linéaire) sont utilisées.

Finalement, le nombre de sommets de cette clique maximum est égal au nombre d'arêtes que comporte un MCES. Il suffit ensuite de considérer les arêtes, correspondant aux sommets de cette clique, dans l'un des graphes de départ pour obtenir le sous-graphe commun ayant le plus grand nombre d'arêtes, noté $G_{12} = (V_{12}, E_{12})$. Une valeur de similarité entre les deux graphes de départ G_1 et G_2 est ensuite calculée.

Dans le cas de recherche d'un MCES noté $G_{12} = (V_{12}, E_{12})$ entre deux graphes moléculaires G_1 et G_2, la métrique de similarité associée est la suivante [90] :

$$sim(G_{12}, G_1, G_2) = \frac{(|V_{12}| + |E_{12}|)^2}{(|V_1| + |E_1|) \times (|V_2| + |E_2|)}$$

Cependant, cette métrique prend en compte les sommets du sous-graphe commun de la même façon que les arêtes, et n'est pas assez discriminante pour notre modèle.

C'est la raison pour laquelle nous avons choisi une métrique calculant uniquement la proportion d'arêtes en commun par rapport au nombre d'arêtes maximum entre G_{O_1} et G_{O_2} (voir équation 2.1).

Adaptation de cet algorithme au modèle de k-extensions

Pour rechercher un sous-graphe commun maximum à deux k-extensions G_{O_1} et G_{O_2}, nous appliquons la même méthode que celle développée dans la sous-section précédente, sur les deux graphes non orientés $G_{O_1,E}$ et $G_{O_2,E}$ issus respectivement de G_{O_1} et G_{O_2} (voir sous-section 2.3.1).

Nous construisons alors les linegraphes de $G_{O_1,E}$ et de $G_{O_2,E}$, en prenant en compte les types de sommets et d'arêtes et les branches de sommets, comme indiqué par le ($\pi_s, \pi_{æ}$)-isomorphisme défini en 2.3.1. Cela signifie que chaque sommet des linegraphes $L(G_{O_1,E})$ et $L(G_{O_2,E})$ est étiqueté par le type de l'arête correspondante dans $G_{O_1,E}$ ou $G_{O_2,E}$, c'est-à-dire le type d'interactions canonique ou non canoniques. De plus, chaque arête des linegraphes $L(G_{O_1,E})$ et $L(G_{O_2,E})$ est étiquetée avec le type du sommet correspondant dans $G_{O_1,E}$ ou $G_{O_2,E}$, ainsi qu'avec les numéros de branches auquel le sommet appartient (voir sous-section 2.2.2). Un exemple de construction de deux linegraphes est présenté en Figure 2.9.

Le graphe produit GP de $L(G_{O_1,E})$ et $L(G_{O_2,E})$ est ensuite calculé avec une contrainte supplémentaire sur les arêtes pour prendre en compte le sens des arcs de la séquence (A_P), absents de $G_{O_1,E}$ et $G_{O_2,E}$. Etant donnés e_i et e_j deux sommets de $L(G_{O_1,E})$ et f_i et f_j deux sommets de $L(G_{O_2,E})$, pour qu'une arête $[(e_i, f_i), (e_j, f_j)]$ appartienne à $E(GP)$, alors que les sommets e_i et f_i (resp. e_j et f_j) ont la même étiquette et que les arêtes $[e_i, e_j] \notin E(L(G_{O_1,E}))$ et $[f_i, f_j] \notin E(L(G_{O_2,E}))$ (voir définition 2.3.5), il faut en plus que les deux sommets incidents à e_i (resp. e_j) dans la k-extension contractée G_{O_1} et les deux sommets incidents à f_i (resp. f_j) dans la k-extension contractée G_{O_2} soient compatibles deux par deux (voir définition 2.3.2).

La recherche d'une clique maximum dans ce graphe produit permet de trouver un MCES aux k-extensions contractées G_{O_1} et G_{O_2}. La valeur de similarité contextuelle (voir équation 2.1) entre les k-extensions G_{O_1} et G_{O_2} peut ensuite être calculée.

Complexité et motivations pour développer une méthode heuristique

Ce type de méthode de recherche d'un MCES, recherchant une clique maximum dans le graphe produit des linegraphes des graphes de départ, a une complexité exponentielle dans le pire des cas. Comme nous le verrons dans les chapitres suivants, les k-extensions contractées que nous utilisons dans les exemples étudiés sont de petite taille (environ 20 sommets). Il est ainsi possible d'utiliser un tel algorithme, pour comparer deux k-extensions d'une telle taille. Cependant, il nous faut pouvoir comparer rapidement un grand nombre de k-extensions entre elles, issues des bases de données. Comme évoqué au début de cette partie, il existe des méthodes de calcul d'un MCES de plus faible complexité, s'appuyant sur les propriétés des graphes étudiés. Par exemple, des méthodes permettent de trouver un MCES en temps polynomial pour des graphes possédant une largeur arborescente inférieure à une constante k et possédant un degré maximum borné [122]. Cependant, ces algorithmes ne sont pas directement utilisables avec notre modèle, car il existe

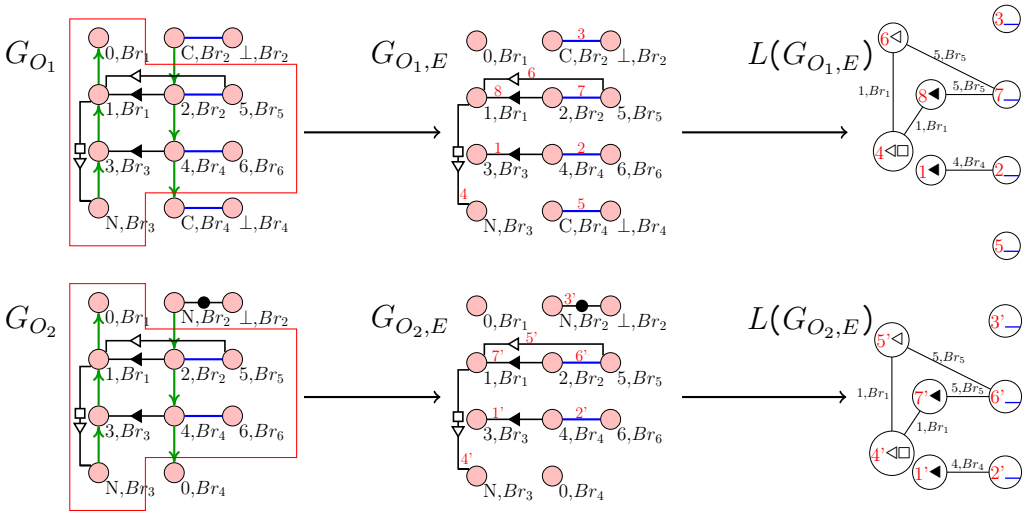


Figure 2.9 – Exemples de deux 2-extensions contractées (à gauche), et des linegraphes correspondants (à droite). Au milieu sont présentés les graphes d'interactions non orientés issus des 2-extensions contractées, c'est-à-dire les graphes privés des arcs de la séquence (A_H), à partir desquels sont construits les linegraphes. Le sous-graphe commun maximum est encadré en rouge dans les 2-extensions contractées G_{O_1} et G_{O_2}. Chaque sommet des k-extensions est annoté par son type ρ_O et la branche à laquelle il appartient (voir section 2.2.2).

de nombreux cas particuliers dans nos k-extensions. Il n'est ainsi pas facile de caractériser la famille de graphes la plus petite possible contenant toutes les k-extensions. C'est pourquoi nous avons développé une heuristique, spécifique aux k-extensions, permettant de diminuer le temps de calcul global, et garantissant une bonne qualité des résultats, comme nous le verrons par la suite. La partie qui suit décrit cette heuristique, ainsi que sa complexité. Nous y présentons également la comparaison de résultats et de temps de calcul pour les deux algorithmes.

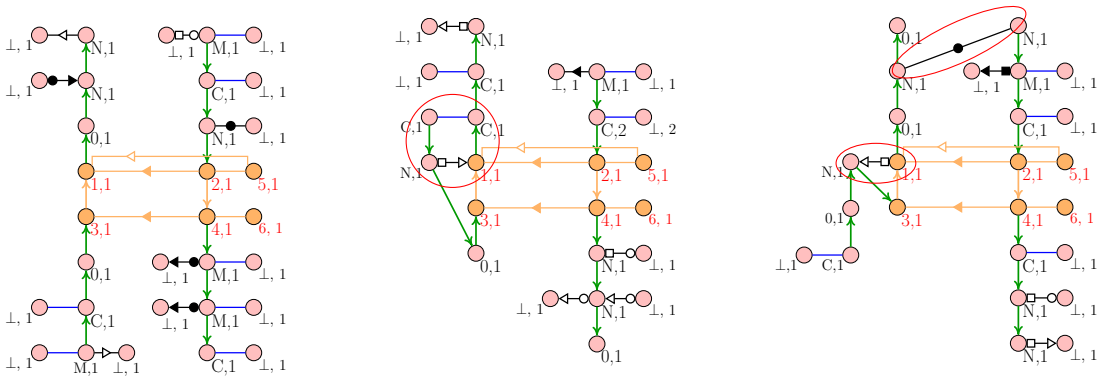
2.4 Heuristique de recherche de similarité entre k-extensions

2.4.1 Concept de l'heuristique

Dans les paragraphes précédents, nous avons présenté une méthode exacte permettant d'obtenir le sous-graphe commun entre deux k-extensions, respectant des conditions particulières sur les sommets et les arcs, et maximisant la similarité contextuelle.

La complexité de cette précédente méthode étant élevée, en particulier quand il s'agit de comparer les k-extensions issues d'occurrences de motif stockées dans les bases de données, nous avons développé une heuristique.

Pour cela, nous nous sommes appuyés sur une caractéristique de



(a) 4-extension avec aucun arc entre les branches

(b) 4-extension avec des arcs entre les branches 1 et 3 (entourés en rouge)

(c) 4-extension avec des arcs entre les branches 1 et 3 et les branches 1 et 2 (entourés en rouge)

Figure 2.10 – Exemples de 4-extensions contractées possédant des liens entre 0 ou plusieurs sous-ensembles de sommets.

construction des k-extensions, i.e. le fait que les sommets d'une k-extension peuvent être couverts par plusieurs sous-ensembles de sommets, appelés branches, en fonction de leur position par rapport aux sommets de l'occurrence du motif (voir section 2.2.2). En pratique, ces sous-ensembles ont souvent peu, voire pas, de sommets en commun (voir Figure 2.10a), et les sommets de différents sous-ensembles ne sont parfois reliés que par l'occurrence du motif. Lorsque des liens existent entre les sommets de différents sous-ensembles qui n'appartiennent pas à l'occurrence du motif, il est rare que cela concerne plus de deux sous-ensembles entre eux (Figure 2.10b et 2.10c). De plus, l'une des contraintes sur les sommets pour construire le sous-graphe commun maximum est justement l'appartenance à une même branche. En effet, cela n'aurait pas de sens biologique de faire correspondre des interactions se trouvant dans des branches différentes.

Comme l'occurrence du motif est présente dans toutes les k-extensions, on peut la supprimer et comparer séparément les composantes connexes obtenues, sans perdre d'information. On ne comparera que les paires de composantes connexes, issues des deux k-extensions distinctes, possédant des sommets qui appartiennent à une même branche (voir Figure 2.11).

Nous pourrions alors utiliser notre méthode exacte de recherche du sous-graphe commun maximum sur ces composantes connexes, qui sont donc des graphes de plus petite taille que les k-extensions complètes. Cependant, pour diminuer davantage la complexité, nous avons développé un algorithme permettant de rechercher un sous-graphe commun de proche en proche, en commençant par les sommets ayant le plus grand nombre d'arcs entrants et sortants. Notons que, contrairement à la méthode exacte décrite dans la section 2.3.2, nous utilisons ici les k-extensions en entrée, qui sont donc des graphes orientés. Nous décrivons l'heuristique dans la partie qui suit.

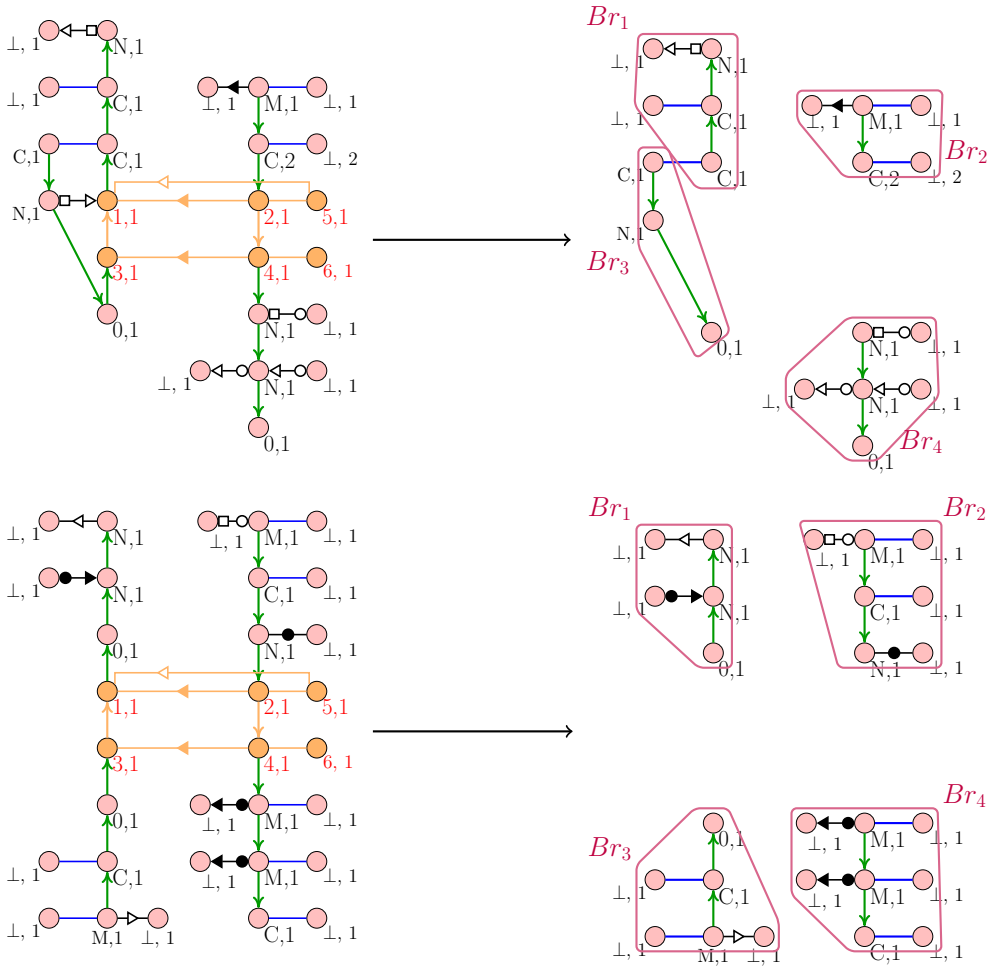


Figure 2.11 – Deux exemples de 4-extensions contractées (à gauche) et les composantes connexes obtenues en supprimant les sommets de l'occurrence de motif (à droite). Les branches de sommets sont indiquées pour les graphes de droite. La composante connexe formée des branches Br_1 et Br_3 pour la 4-extension du haut sera comparée avec la composante connexe formée de la branche Br_1 de la k-extension du bas d'une part, et avec la composante connexe formée de la branche Br_3 d'autre part. La composante connexe formée de la branche Br_3 (resp. Br_4) de la 4-extension du haut sera comparée avec la composante connexe formée de la branche Br_3 (resp. Br_4) de la k-extension du bas.

2.4.2 Description de l'heuristique de recherche d'un sous-graphe commun maximum entre k-extensions

Tout d'abord, nous allons définir une notion de compatibilité entre composantes connexes, à l'aide des fonctions d'isomorphisme définies dans la partie précédente.

Soient $G_1 = (V_1, A_{P,1}, A_{H,1})$ et $G_2 = (V_2, A_{P,2}, A_{H,2})$ deux k-extensions contractées ou non contractées distinctes.
Nous noterons $G_{1\bar{M}}$ (resp. $G_{2\bar{M}}$) le sous-graphe de G_1 (resp. G_2) induit par les sommets de G_1 (resp. G_2) n'appartenant pas à l'occurrence de motif (voir exemples en Figure 2.11).

Définition 2.4.1 *Deux composantes connexes $C_1 = (V_1^c, A_{P,1}^c, A_{H,1}^c)$ de G_1 et $C_2 = (V_2^c, A_{P,2}^c, A_{H,2}^c)$ de G_2 sont compatibles si et seulement s'il existe au moins un couple de sommets $u_1 \in V_1^c$ et $u_2 \in V_2^c$ tel que $u_1 \in \pi_s(u_2)$.*

L'heuristique de recherche d'un sous-graphe commun maximum entre deux k-extensions est présentée en pseudo-code ci-après (algorithme 1). Nous appellerons ici voisins, tous les successeurs et prédécesseurs d'un sommet.

Pour un couple donné C_1 et C_2 de composantes connexes compatibles entre $G_{1\bar{M}}$ et $G_{2\bar{M}}$, nous recherchons, parmi les sommets de C_1 qui ne sont pas de type $\perp$, le sommet u_{choisi} ayant le plus grand nombre de voisins non covalents, c'est-à-dire reliés au sommet u_{choisi} par un arc de $A_{H,1}$. Nous pouvons noter que ce nombre est pondéré par le poids du sommet : un sommet de poids 2 comptera pour 2 dans le nombre de voisins non covalents (voir les valeurs indiquées dans les sommets de la Figure 2.12). Si plusieurs sommets ont le même nombre de voisins non covalents, nous prenons celui qui se trouve à la plus petite distance d'un des sommets de l'occurrence du motif.

Nous recherchons ensuite tous les sommets v de C_2 tel que le couple (u_{choisi}, v) soit compatible avec tous les couples de sommets appartenant au sous-graphe commun déjà construit (voir définition 2.3.2). A partir de chaque couple $< u_{choisi}, v >$ tel que $v \in C_2$ soit compatible avec u_{choisi}, on réalise un parcours (en largeur) de C_1 et C_2 simultanément en recherchant des couples de sommets compatibles, et tel que les arcs des deux parcours soient de même type deux par deux. Lorsque plusieurs choix sont possibles, c'est-à-dire lorsque deux couples de sommets compatibles sont reliés deux à deux dans C_1 et C_2 par des arcs de même type, nous testons chaque possibilité séparément et considérons le sous-graphe commun ayant le plus grand nombre d'arcs de $A_{H,1}$ pondérés par le poids des sommets. Notons cependant que ce cas arrive rarement, car il est rare qu'un sommet soit lié à deux autres sommets par des arcs de même type. Un exemple de comparaison entre deux composantes connexes compatibles est présenté en Figure 2.12.

Chacun de ces parcours en largeur induit un sous-graphe commun à C_1 et à C_2 (c'est-à-dire un sous-graphe qui est $(\pi_s, \pi_{æ})$-isomorphe à un sous-graphe de C_1 et à un sous-graphe de C_2). Nous considérons celui qui possède le plus grand nombre d'arêtes non covalentes, pondérées par le poids des sommets (et donc associé à une similarité contextuelle maximale). Les sommets et les

Algorithme 1 : Heuristique de recherche du sous-graphe commun maximum à deux k-extensions G_1 et G_2

Entrées : deux k-extensions $G_{1\bar{M}}$ et $G_{2\bar{M}}$ privées des sommets de leur occurrence de motif

Sorties : Sous-graphe commun à $G_{1\bar{M}}$ et $G_{2\bar{M}}$ ayant le plus grand nombre d'arcs de $A_{H,1}$ pondérés par le poids des sommets, selon l'heuristique

G'_1 := graphe vide (sous-graphe de G_1 de sortie)
G'_2 := graphe vide (sous-graphe de G_2 de sortie)
$\mathscr{C}_1$:= Composantes connexes de $G_{1\bar{M}}$
$\mathscr{C}_2$:= Composantes connexes de $G_{2\bar{M}}$
Pour *chaque composante connexe C_1 de $\mathscr{C}_1$* **faire**
  G_{C_1} := graphe vide (sous-graphe de C_1 commun à G_{C_2})
  G_{C_2} := graphe vide (sous-graphe de $G_{2\bar{M}}$ commun à G_{C_1})
  $u_{choisi} \in V_1^c$ = un sommet de C_1 (de type différent de $\perp$) ayant le plus grand nombre de voisins non covalents, pondéré par le poids du sommet
  Pour *chaque composante connexe C_2 de $\mathscr{C}_2$* **faire**
    G_{u_{choisi},C_1} := graphe vide (sous-graphe de C_1, contenant u_{choisi} et commun à G_{u_{choisi},C_2})
    G_{u_{choisi},C_2} := graphe vide (sous-graphe de C_2 commun à G_{u_{choisi},C_1})
    si *C_1 et C_2 sont compatibles (voir définition 2.4.1)* **alors**
      $V_{comp} \subset V_2^c$ = l'ensemble des sommets de C_2 compatibles avec u_{choisi} pour être ajouté au graphe G_{C_2} ainsi qu'au graphe G'_2
      Pour *tout $v \in V_{comp}$* **faire**
        - Parcourir (en largeur) simultanément C_1 à partir de u_{choisi} et C_2 à partir de v, en recherchant des couples de sommets compatibles (pour être ajouté à G_{C_1} ou G_{C_2} et G'_1 ou G'_2) et de façon à ce que les deux arcs à chaque étape du parcours soient de même type.
        (Si plusieurs couples de sommets sont possibles à une étape du parcours, considérer les différents sous-graphes communs possibles et choisir le sous-graphe correspondant à la similarité contextuelle maximale)
        → $G_{u_{choisi},v}$ est le sous-graphe de C_1 résultant de ce parcours
        → $G_{v,u_{choisi}}$ est le sous-graphe de C_2 résultant de ce parcours
        si $sim(G_{u_{choisi},v}, C_1, C_2) > sim(G_{u_{choisi},C_1}, C_1, C_2)$ **alors**
          $G_{u_{choisi},C_1} := G_{u_{choisi},v}$
          $G_{u_{choisi},C_2} := G_{v,u_{choisi}}$
    $G_{C_1} := G_{C_1} \cup G_{u_{choisi},C_1}$
    $G_{C_2} := G_{C_2} \cup G_{u_{choisi},C_2}$
    $C_2 := C_2$ privé des sommets de G_{u_{choisi},C_2}
  si G_{C_1} *est vide* **alors**
    $C_1 := C_1 / u_{choisi}$
  sinon
    $C_1 := C_1$ privé des sommets de G_{C_1}
  $G'_1 = G'_1 \cup G_{C_1}$
  $G'_2 = G'_2 \cup G_{C_2}$
renvoyer G'_1, G'_2

arcs de ce sous-graphe commun sont ajoutés au sous-graphe commun en construction. Les sommets et les arcs de C_2 appartenant au sous-graphe auquel le sous-graphe commun est $(\pi_s, \pi_{æ})$-isomorphe sont supprimés de C_2.

Nous effectuons ainsi cette recherche entre C_1 et toutes les composantes connexes dans $G_{2\bar{M}}$ qui sont compatibles avec C_1, puis supprimons de C_1 tous les sommets et les arcs appartenant au sous-graphe de C_1 auquel le sous-graphe commun construit est $(\pi_s, \pi_{æ})$-isomorphe.

A la fin de cet algorithme, pour obtenir le sous-graphe commun à G_1 et G_2 recherché, il suffit d'ajouter au sous-graphe commun de sortie les sommets et les arcs de l'occurrence de motif (communs à toutes les k-extensions) et ajouter également les arcs dont l'une des extrémités est l'un des sommets du motif si ces arcs existent et sont de même type dans G_1 et dans G_2.

On peut également noter que comparer le graphe G_1 au graphe G_2 peut ne pas être équivalent à comparer le graphe G_2 au graphe G_1, car les deux graphes ne sont pas traités de la même manière dans l'algorithme. Ainsi, pour s'assurer de trouver le sous-graphe commun maximisant autant que possible la similarité contextuelle, il vaut mieux appliquer l'algorithme sur les deux couples de graphes et conserver le meilleur résultat.

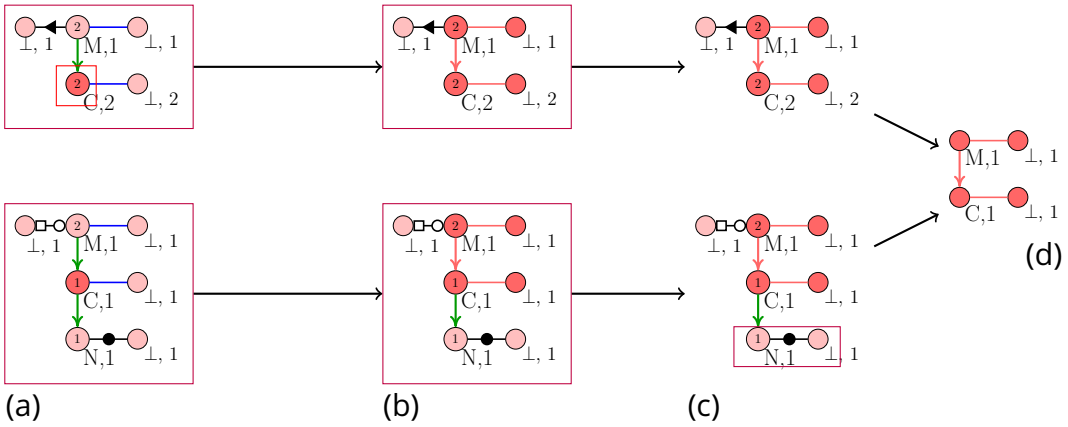


Figure 2.12 – Recherche d'un sous-graphe commun de proche en proche entre les composantes connexes de la branche Br_2 des 4-extensions de la Figure 2.11. Le nombre de voisins covalents de chaque sommet qui n'est pas de type $\perp$ est indiqué. Les parties encadrées dans chaque graphe correspondent aux composantes connexes qu'on considère à chaque étape. Le premier u_{choisi} est le sommet encadré en rouge dans la composante connexe en haut en (a). Il est compatible avec le sommet rouge du deuxième graphe (en-dessous). A partir de ces deux sommets, on peut construire le sous-graphe commun représenté en rouge dans les graphes en (b). Puis, si on supprime les sommets, et les arcs de ce sous-graphe commun, il ne reste plus de sommet de type différent de $\perp$ dans le graphe du dessus (en (c), pas de sommet encadré), et seulement un seul sommet dans le graphe du dessous (en (c), sommets encadrés). La recherche de sous-graphe commun entre ces deux composantes connexes s'arrête donc à cette étape, et le sous-graphe commun aux deux k-extensions est présenté en (d).

2.4.3 Etude de la qualité et de la complexité de la méthode heuristique par rapport à la méthode exacte

Nous avons comparé la méthode exacte et la méthode heuristique sur un jeu de données de 389 k-extensions contractées d'occurrences de motif A-minor de type I/II (voir chapitre 1, page 32), issues de la base de données de motifs CaRNAval [92]. C'est un motif à longue distance qui reste très difficile à prédire et qui est très fréquemment retrouvé dans les structures d'ARN (voir chapitre 1, section 1.7.3). Nous avons choisi une valeur de $k = 4$ et un ensemble de sommets du motif $S = \{1, 2, 3, 4\}$ (voir section 2.2.2). Ce choix sera expliqué dans le chapitre 3 (section 3.6).

Nous étudierons plus avant ce motif dans les chapitres suivants, ainsi que deux autres motifs de la base de données CaRNAval. Parmi ces trois motifs que nous étudierons par la suite, le motif A-minor comporte le nombre le plus élevé d'occurrences.

Dans cette section, nous allons comparer les résultats obtenus avec les deux méthodes, ainsi que la complexité et le temps d'exécution de ces deux méthodes.

Comparaison des résultats

Pour le motif A-minor, sur les 75466 ($(389 \times 388)/2$) paires de k-extensions, seules 75 paires ont une similarité contextuelle différente entre les deux méthodes. Ainsi, 99,9% des paires possèdent la même similarité contextuelle avec les deux méthodes. Les valeurs de similarité contextuelle étant comprises entre 0 et 1, la différence moyenne sur ces 75 paires est de 0,05, et la différence maximale de 0,095. Dans les sous-graphes, ces différences ne correspondent ainsi qu'à une ou deux interactions en moins.

Ainsi, dans la très grande majorité des cas, l'heuristique permet d'obtenir la similarité contextuelle associée à un sous-graphe commun maximum entre les k-extensions.

Comparaison des complexités

Dans le cas général, dans le pire des cas, la complexité de notre méthode heuristique est exponentielle en fonction du nombre de sommets des k-extensions. En effet, considérons une paire de k-extensions $G_1 = (V_1, A_{p,1}, A_{H,1})$ et $G_2 = (V_2, A_{P,2}, A_{H,2})$ dans lesquelles tous les sommets sont de même type (et appartiennent à la même branche) et tous les arcs non covalents sont également de même type sans contrainte sur les degrés des sommets. Dans de tels graphes, à partir d'un couple de sommets <u_1,u_2> avec u_1 appartenant à G_1 et u_2 appartenant à G_2, le nombre d'arbres que l'on peut obtenir par un parcours en largeur tel qu'effectué dans l'algorithme est exponentiel, dans $O(Max(\Delta^+_{G_1}, \Delta^+_{G_2})^{|V_1|})$ avec $\Delta^+_{G_1}$ (resp. $\Delta^+_{G_2}$) le degré sortant maximum du graphe G_1 (resp. G_2). En effet, chaque successeur de u_1 dans G_1 est compatible avec chaque successeur de u_2 dans G_2. Dans un parcours

	Temps d'exécution total (en min) pour les k-extensions non contractées (75466 paires)	Temps d'exécution total (en min) pour les k-extensions contractées (75466 paires)
Méthode exacte	960	240
Heuristique	17,4	13,5

Table 2.2 – Temps d'exécution de la recherche de sous-graphe commun maximum sur le jeu de données de 389 motifs A-minor, sur une machine Intel Core i5-7440HQ 4x2.80GHzCPU. Les algorithmes sont implémentés en Python3.

donné, on choisit un couple de successeurs donné, puis on recherche parmi les successeurs de ces successeurs non encore vus un autre couple de successeurs compatible.

Ces parcours sont effectués à partir de chaque couple de sommets compatibles dans les deux k-extensions de départ, ce qui donne une complexité dans $O(|V_1| \times |V_2| \times Max(\Delta^+_{G_1}, \Delta^+_{G_2})^{|V_1|})$ dans le pire des cas.

Cependant, des contraintes inhérentes aux k-extensions existent, empêchant ce genre de cas d'arriver. Le degré sortant d'un sommet est en effet inférieur à 5 comme un nucléotide ne peut interagir avec plus de 5 autres nucléotides. Dans les k-extensions de nos jeux de données, le degré entrant ou sortant d'un sommet est rarement supérieur à 3. De plus, il n'y a en général que $k-1$ sommets par branche qui possèdent un type différent de $\perp$ (et qui peuvent donc être choisis). Enfin, dans la plupart des cas, il y a un seul parcours possible à partir d'un couple de sommets compatibles donné, car il arrive très rarement qu'un sommet soit connecté à deux autres sommets par des arcs de même type.

Du point de vue pratique, sur l'exemple du motif A-minor, le temps d'exécution de la méthode heuristique est réduit de 50 fois pour les k-extensions non contractées et de 17 fois pour les k-extensions contractées, par rapport à la méthode exacte (voir table 2.2).

2.5 Définition d'un représentant à un sous-ensemble de k-extensions

Dans les sections précédentes, nous avons décrit le sous-graphe commun maximum à deux k-extensions contractées ou non contractées. Nous pouvons étendre cette définition à un sous-ensemble de taille n de k-extensions, pour obtenir un sous-graphe commun à ces n k-extensions. Cette partie définit ce sous-graphe commun, que l'on appellera *représentant* d'un sous-ensemble de k-extensions, ainsi que la manière de l'obtenir.

Dans cette partie, nous appellerons $\mathscr{E}_i$ un sous-ensemble de k-extensions contractées de taille n_i. Nous le notons $\mathscr{E}_i = \{G_{i,1}, G_{i,2}, ..., G_{i,n_i}\}$ avec $G_{i,j} = (V_{i,j}, A_{P,i,j}, A_{H,i,j})$ pour tout $j \in \{1, 2, ..., n_i\}$

Définition du représentant

Définition 2.5.1 *On appelle représentant de $\mathscr{E}_i$, le sous-graphe commun maximum à chaque paire de k-extensions contractées de $\mathscr{E}_i$.*

Ce sous-graphe commun maximum est défini de la même manière qu'un sous-graphe commun maximum entre deux k-extensions contractées (voir section 2.3), mais pour n_i k-extensions contractées.

Construction du représentant

Pour déterminer le sous-graphe commun maximum à un sous-ensemble $\mathscr{E}_i$, on construit tout d'abord un graphe non orienté $G_C = (V_C, E_C)$ tel que :

- l'ensemble des sommets $V_C = \bigcup_{j=1}^{n} V_{i,j}$ et
- pour toute paire de sommets $\{u, v\} \in V_C^2$ avec $u \in V_{i,l}$ et $v \in V_{i,m}$ ($l \neq m$, $l \in \{1, 2, .., n_i\}$, $m \in \{1, 2, ..., n_i\}$), $[u, v] \in E_C$:

  si u appartient au sous-graphe de $G_{i,l}$ isomorphe au sous-graphe commun maximum de $G_{i,l}$ et de $G_{i,m}$ (déterminé par l'algorithme de recherche d'un MCES) et,

  si v appartient au sous-graphe de $G_{i,m}$ isomorphe au sous-graphe commun maximum de $G_{i,l}$ et de $G_{i,m}$ également,

Dans ce graphe G_C, on recherche toutes les cliques de taille n_i. Une telle clique contient nécessairement un et un seul sommet de chaque k-extension contractée du sous-ensemble $\mathscr{E}_i$. Chacun des sommets d'une clique de taille n_i appartient au sous-graphe que l'on recherche, qui est commun aux n_i k-extensions contractées du sous-ensemble $\mathscr{E}_i$. Pour l'obtenir, il faut considérer le sous-graphe d'une k-extension contractée $G_{i,j}$ induit par l'ensemble des sommets de ces cliques appartenant à $G_{i,j}$.

Pour énumérer toutes les cliques de taille n_i du graphe, nous avons utilisé l'algorithme de Bron-Kerbosch [125]. Cet algorithme a une complexité exponentielle dans le cas général, mais, comme nous le verrons dans les chapitres suivants, la taille n_i des sous-ensembles de k-extensions contractées que l'on considère est petite (en général inférieure à 10), ce qui rend la recherche de cliques de taille n_i possible en temps polynômial.

Notons également qu'étant donné que la recherche de MCES donne un seul sous-graphe commun maximum à deux k-extensions, et pas tous les sous-graphes communs maximum à ces deux k-extensions, il n'y a pas de garantie que le sous-graphe commun à un sous-ensemble de k-extensions que nous définissons ici soit maximum. Pour le garantir, il faudrait rechercher l'intersection maximum de tous les sous-graphes communs à toutes les paires de k-extensions du sous-ensemble. Ce problème est alors bien plus complexe à résoudre. Nous n'avons cependant pas adressé ce problème dans cette thèse, comme la méthode décrite ici permet d'obtenir de bons résultats pour les sous-ensembles de k-extensions que nous étudions.

2.6 Conclusion

Dans ce chapitre, nous avons défini un modèle de graphes permettant de représenter le contexte structural d'un motif d'ARN, constitué des interactions canoniques et non canoniques. Deux types de graphes ont ainsi été définis : l'un représentant le contexte structural topologique comme un graphe, appelé k-extension, avec chaque sommet correspondant à un nucléotide et chaque arc à une interaction covalente, canonique ou non canonique, et l'autre, appelé k-extension contractée, regroupant certaines parties du contexte structural topologique en un seul sommet. Notons que la taille du contexte structural considérée (valeur k) peut être paramétrée.

Pour déterminer si deux contextes topologiques sont similaires ou non, nous avons utilisé la notion de MCES entre deux graphes, ce qui nous permet d'obtenir un sous-graphe commun maximisant le nombre d'arcs non covalents, c'est-à-dire le nombre d'interactions canoniques et non canoniques communes. Pour quantifier cette similarité, nous avons défini une métrique, appelée similarité contextuelle, calculant la proportion d'arcs non covalents en commun par rapport aux nombres d'arcs non covalents dans les deux graphes de contexte structural de départ. Nous avons présenté deux méthodes permettant d'obtenir un sous-graphe commun maximum, une méthode exacte et une méthode heuristique, et montré, sur l'exemple du motif A-minor, que l'heuristique permet d'obtenir dans la majorité des cas un sous-graphe commun maximum et est significativement plus rapide que la méthode exacte en termes de temps d'exécution.

Nous avons également défini un sous-graphe commun à un sous-ensemble de k-extensions, que l'on pourra utiliser dans les chapitres suivants (chapitre 5).

Dans les chapitres suivants, nous allons appliquer cette approche à des jeux de données de motifs à longue distance d'ARN, comme le motif A-minor par exemple. Nous pourrons ainsi faire évoluer les différents paramètres associés à notre approche (taille de contexte, contraction, seuil de similarité, etc.), et classifier les contextes selon notre métrique, dans le but de déterminer dans quelle mesure les similarités de topologies du contexte d'un motif d'ARN peuvent expliquer les similarités apparaissant entre les structures 3D associées.

Chapitre 3

Cohérence entre la similarité contextuelle et la similarité 3D sur un jeu de données de trois motifs complexes

3.1 Introduction

Dans ce chapitre, nous allons présenter une étude portant sur trois motifs à longue distance d'ARN : le motif A-minor, bien connu, très répandu et réputé difficile à prédire par les méthodes algorithmiques actuelles, et deux autres motifs, mis en évidence dans [92] et présents dans la base de données CaRNAval. Pour ces trois motifs, nous allons comparer les topologies de contexte, comme définies dans le chapitre 2, avec une notion de contextes 3D que nous définirons précisément dans ce chapitre. La structure 3D locale d'un motif d'ARN et de son contexte (contexte 3D) contient les positions relatives dans l'espace des atomes les uns par rapport aux autres. L'information contenue dans cette structure 3D contient donc, notamment, la topologie du contexte, c'est-à-dire les interactions canoniques et non canoniques du contexte, représentées par nos k-extensions. Nous souhaitons comprendre le lien entre la topologie de contexte et la formation du motif, en comparant la topologie de contexte au contexte 3D. Pour cela, nous allons étudier la corrélation entre la similarité contextuelle, décrivant les similarités de topologies de contexte, et une métrique décrivant les similarités de contextes 3D dans un alignement de structures optimal.

Nous allons donc d'abord présenter, dans la section 3.2, les données que nous avons utilisées, et le traitement préalable que nous avons effectué sur ces données. Pour cela, nous présenterons les trois motifs, en indiquant de quelles bases de données ils proviennent, quelles sont leurs caractéristiques et pourquoi nous les avons choisis. Puis, nous expliquerons comment nous avons obtenu toutes les occurrences de ces motifs stockées dans la PDB.

Nous définirons ensuite, dans la section 3.3, une manière de comparer

entre elles les structures 3D associées aux contextes d'occurrences de motif, et de quantifier leurs similarités, dans le but de comparer cette métrique à celle utilisée pour comparer les contextes topologiques.

Dans la section 3.4, nous nous intéresserons à une première manière de regrouper les occurrences d'un motif selon leur homologie.

Nous détaillerons ensuite, dans la section 3.5, une méthode permettant de classifier les occurrences de motifs selon la similarité contextuelle des k-extensions, et selon la métrique permettant de comparer les structures 3D locales.

Après cette présentation des outils utilisés, dans la section 3.6, nous étudierons l'effet de la variation de certains paramètres des k-extensions (contraction, taille d'extension k, seuil de similarité), sur la corrélation entre les deux métriques de similarité.

Et enfin, dans la section 3.7, après avoir choisi les valeurs de ces paramètres permettant d'obtenir la meilleure corrélation entre les deux métriques, nous comparerons les deux métriques entre elles, et les deux métriques avec les groupes d'homologie que nous aurons définies au préalable.

Les modèles et algorithmes du chapitre précédent, ainsi que les expérimentations présentées dans ce chapitre et les suivants ont été implémentés en Python3, notamment à l'aide du package networkx pour la représentation des graphes.

3.2 Présentation des données utilisées

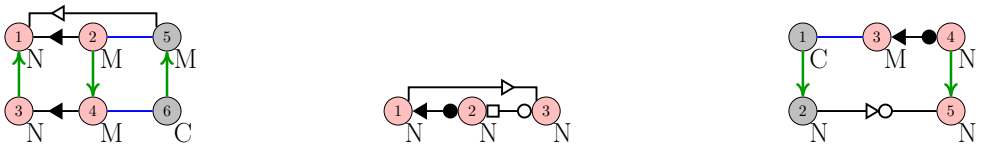
3.2.1 Description des trois motifs structuraux étudiés

Nous avons étudié plusieurs motifs apparaissant dans les structures 3D d'ARN (Figure 3.1) :

- Le *motif A-minor de type I/II* [63] (Figure 3.1a). Comme évoqué dans le chapitre 1 (section 1.5.3), ce motif apparaît entre une boucle de structure secondaire et une hélice, et relie ces deux éléments par des interactions non canoniques (cis Sugar-Sugar et trans Sugar-Sugar). Ces interactions relient souvent des régions de la molécule qui sont éloignées sur la séquence.

  Le motif A-minor est présent dans de nombreuses familles d'ARN non codants (ARN ribosomiques, ARN de transfert, riboswitches, ribozymes, etc.), et représente plus de 80% des interactions non canoniques à longue distance apparaissant dans les structures d'ARN, d'après la base de données de motifs CaRNAval [92]. Il a été démontré que ce motif était important dans le repliement dans l'espace des molécules d'ARN, ainsi que dans des mécanismes cellulaires comme la reconnaissance codon-anticodon au moment de la traduction [66].

- Le motif *trans Watson-Crick/Hoogsteen* (Figure 3.1b), que nous noterons trans WC/H, provenant de la base de données de motifs à longue



(a) Motif A-minor (b) Motif trans WC/Hoogsteen (c) Motif G

Figure 3.1 – Les trois motifs que l'on étudie. Ils sont représentés de la même façon que les graphes d'ARN du chapitre 2. Les sommets roses sont les sommets à partir desquels les k-extensions seront construites (sous-ensemble S dans la définition 2.2.6)

distance CaRNAval [92]. Ce motif est composé de trois nucléotides reliés entre eux par des interactions non canoniques (en particulier, une interaction de type trans Watson-Crick/Hoogsteen qui lui a donné son nom). On retrouve ce motif le plus souvent dans les ARN de transfert, seuls ou en complexes avec des protéines. On le trouve également dans des ARN ribosomiques et des introns.

- un autre motif de la base de données CaRNAval que nous avons appelé *motif G* (Figure 3.1c). Il est également présent dans plusieurs familles d'ARN, comme les ARN ribosomiques, les riboswitchs et les ribozymes.

Le motif A-minor et son implication dans des mécanismes biologiques ont fait l'objet de nombreuses études [79, 66, 40] mais ce motif reste pour le moment difficile à prédire. Les deux autres motifs ont été caractérisés pour la première fois par la méthode de détection de motifs récurrents , utilisée pour construire la base de données CaRNAval [92] (voir chapitre 1, section 1.5.3). Ces deux motifs ont été choisis dans cette étude, car ce sont les plus petits motifs de la base CaRNAval impliquant des interactions non canoniques, et ils sont observés dans des molécules d'ARN de différentes familles. Ces deux motifs n'ont jamais été étudiés de manière exhaustive à ma connaissance.

3.2.2 Obtention et filtrage des données de la PDB

Nous avons récupéré les occurrences de ces 3 motifs, stockées dans la base de données CaRNAval. Ces occurrences correspondent à toutes les occurrences de ces motifs se trouvant dans les structures de la PDB (Protein Data Bank) (récupérées en août 2019).

En amont de cette étude, les interactions canoniques et non canoniques ont été annotées dans les structures comportant ces occurrences, à l'aide du programme FR3D [98]. Rappelons que cette méthode utilise les caractéristiques géométriques des nucléotides dans l'espace pour déterminer la présence de ces interactions. A partir de ces informations, nous représentons les contextes structuraux de ces occurrences à l'aide de graphes, par la méthode présentée dans le chapitre 2. Pour cela, il nous faut déterminer à partir de quels sommets du motif nous voulons construire les k-extensions (sous-ensemble S dans la définition 2.2.6). Nous avons décidé, lorsqu'une

Intramoléculaire									
Famille d'ARN / Organisme	ARNr 23S	ARNr 25S	ARNr 28S	ARNr 16S	ARNr 18S	Riboswitch	Intron	Ribozyme	**Total**
D. radiodurans (B)	18								**18**
E. coli (B)	47			21					**68**
H. marismortui (A)	79								**79**
M. jannaschii (A)	1								**1**
S. oleracea (E)	17			9					**26**
S. aureus (B)	19			5					**24**
T. thermophilus (B)	46			33					**79**
S. cerevisiae (E)		19			9				**28**
H. sapiens (E)			8		4				**12**
L. donovani (E)					7				**7**
T. petrophila (E)						1			**1**
T. tengcongensis (B)						10			**10**
T. thermophila (E)							1		**1**
O. iheyensis (B)							6		**6**
D. iridis (E)								1	**1**
Hepatitis delta virus								1	**1**
Non spécifié	2					8		5	**15**
Total	**229**	**19**	**8**	**68**	**20**	**19**	**7**	**7**	**377**

Intermoléculaire			
Famille d'ARN / Organisme	ARNr 16S – ARNt	16S – ARNt – ARNm	**Total**
T. thermophilus/E. coli/E. coli (B)	3		**3**
E. coli/ E.coli/E.coli (B)	4		**4**
T. thermophilus/T. thermophilus/T. thermophilus (B)	1		**1**
T. thermophilus/synthetic/E.coli (B)		1	**1**
T. thermophilus/?/? (B)		1	**1**
T. thermophilus/Enterobacteria phage T4/E.coli (B)		1	**1**
T. thermophilus/E.coli/E.coli (B)		1	**1**
Total	**8**	**4**	**12**

Table 3.1 – Nombre d'occurrences non redondantes intramoléculaires et intermoléculaires de motif A-minor, par organisme et famille d'ARN. Pour chaque organisme, il est indiqué entre parenthèses à quel règne du Vivant il appartient (Bactéries (B), Archées (A) ou Eucaryotes (E)).

Intramoléculaire

Famille d'ARN / Organisme	ARNr 23S	ARNr 5S	ARNr 16S	ARNr alpha	ARNr 25S	ARNr 28S	ARNr 18S	Ribo-switch	Ribo-zyme	Intron	ARN viral pseudonoeud (28 nts)	ARNsg	ARNm	**Total**
T.thermophilus (B)	11	4	3					2	1					**21**
E.coli (B)	9	4	4					1						**18**
H.marismortui (A)	16													**16**
T.tengcongensis (B)								15						**15**
L.donovani (E)				5			1							**9**
S.cerevisiae (E)		1			4		1	2						**8**
S.aureus (B)	5		2											**7**
H.sapiens (E)						4	2							**6**
D.radiodurans (B)	4	1												**5**
S.oleracea (E)	3		1											**4**
T.cruzi (E)		1		2										**4**
B.subtilis (B)								3						**3**
C.subterraneus (E)								3						**3**
V.vulnificus (B)								3						**3**
O.iheyensis (B)										3				**3**
R.solanacearum (B)								2						**2**
D.iridis (E)									1					**1**
C.jejuni (B)												1		**1**
Non spécifié					1			22	3		3		1	**30**
Total	**48**	**11**	**10**	**7**	**5**	**4**	**4**	**53**	**5**	**3**	**3**	**1**	**1**	**159**

Intermoléculaire

Famille d'ARN / Organisme	ARNr alpha - ARNr zeta	ARNr alpha - ARN sr3	**Total**
L. donovani (E)	2	1	**3**
T. cruzi (E)		1	**1**
Total	**2**	**2**	**4**

Table 3.2 – Nombre d'occurrences non redondantes intramoléculaires et intermoléculaires du motif G, par organisme et familles d'ARN. Pour chaque organisme, il est indiqué entre parenthèses à quel règne du Vivant il appartient (Bactéries (B), Archées (A) ou Eucaryotes (E)).

hélice de structure secondaire est impliquée dans le motif, c'est-à-dire lorsqu'un (ou plusieurs) arc(s) canonique(s) est (sont) présent(s), de choisir les sommets impliqués dans l'un des deux brins de l'hélice seulement. Nous choisissons en priorité le brin contenant les sommets impliqués dans un nombre maximum d'arcs non canoniques du motif. Cela nous permet de ne pas considérer les informations de l'hélice de manière redondante sur les deux brins. De plus, comme la structure 3D des hélices est aisément prédictible, nous souhaitons nous intéresser davantage aux interactions non canoniques. Ainsi, les sommets non impliqués dans une hélice et impliqués dans des arcs non canoniques sont tous choisis également. Pour chaque motif, les sommets choisis sont indiqués en rose dans la Figure 3.1.

La PDB comporte de nombreuses structures redondantes, c'est-à-dire des structures qui correspondent à la même molécule, mais obtenues dans des conditions expérimentales différentes. Ainsi, les occurrences de motif récupérées comportent elles aussi une grande redondance.

Nous avons alors filtré les données pour ne conserver que des occurrences non redondantes. Pour cela, nous avons créé des classes d'occurrences identiques, en considérant deux occurrences de motif comme identiques si les

Famille d'ARN / Organisme	ARNt	ARNr 18S	ARNr 23S	ARNr 25S	Intron	ARNr 28S	ARNr alpha	Autres	**Total**
E.coli (B)	25	0	1	0	0	0	0	0	**26**
S.cerevisiae (E)	8	1	0	1	0	0	0	0	**10**
H.sapiens (E)	5	0	0	0	0	1	0	0	**6**
T.thermophilus (B)	5	0	1	0	0	0	0	0	**6**
T.maritima (B)	4	0	0	0	0	0	0	0	**4**
L.donovani (E)	0	1	0	0	0	0	2	0	**3**
O.cuniculus (E)	1	1	0	0	0	1	0	0	**3**
S.aureus (B)	0	0	2	0	0	0	0	0	**2**
T.kodakarensis (A)	0	0	2	0	0	0	0	0	**2**
A.baumannii (B)	0	0	1	0	0	0	0	0	**1**
A.fulgidus (A)	1	0	0	0	0	0	0	0	**1**
B.subtilis (B)	1	0	0	0	0	0	0	0	**1**
G.kaustophilus (B)	1	0	0	0	0	0	0	0	**1**
L.bacterium (B)	0	0	0	0	0	0	0	1	**1**
M.tuberculosis (B)	1	0	0	0	0	0	0	0	**1**
O.iheyensis (B)	0	0	0	0	1	0	0	0	**1**
P.abyssi (A)	1	0	0	0	0	0	0	0	**1**
S.enterica (B)	0	0	0	0	0	0	0	1	**1**
T.cruzi (E)	0	0	0	0	0	0	1	0	**1**
?	15	0	0	0	0	0	0	0	**15**
Total	**68**	**3**	**7**	**1**	**1**	**2**	**3**	**2**	**87**

Table 3.3 – Tableau du nombre d'occurrences non redondantes de motif trans WC/H, par organisme et famille d'ARN. Toutes les occurrences sont intramoléculaires. Pour chaque organisme, il est indiqué entre parenthèses à quel règne du Vivant il appartient (Bactéries (B), Archées (A) ou Eucaryotes (E)).

séquences de 30 nucléotides de part et d'autre de chaque nucléotide de l'occurrence de motif étaient identiques. Nous avons ensuite choisi le représentant d'une classe selon deux critères : la meilleure résolution de la structure 3D globale d'abord, puis le nombre maximum d'arcs non covalents dans la topologie du contexte structural de même taille k.

Dans les tables 3.1, 3.2 et 3.3, sont présentées les occurrences des différents motifs après filtrage, triées par organisme et famille d'ARN.

3.3 Recherche de similarité 3D entre contextes structuraux

Dans le chapitre précédent, nous avons défini une manière de comparer les contextes topologiques entre eux (chapitre 2, section 2.3). Dans ce chapitre, nous cherchons en particulier à déterminer si des contextes topologiques similaires correspondent à des structures 3D similaires ou non, et inversement. En effet, la structure tridimensionnelle d'une molécule d'ARN est l'ensemble des coordonnées dans l'espace des atomes composant cette molécule. L'information de topologie est donc incluse dans la structure 3D. Nous cherchons ainsi à savoir si l'information apportée par la topologie de contexte uniquement (sans toute l'information fournie par la structure 3D) est suffisante pour décrire les contextes structuraux d'un motif d'ARN, et en

expliquer les similarités.

Pour cela, nous devons définir une manière de comparer les structures 3D des contextes d'occurrences de motif. Nous définissons la sous-structure 3D correspondant au contexte d'une occurrence de motif, à partir d'une k-extension. Puis, nous utilisons une mesure de similarité, appelée RMSD, pour comparer ces sous-structures 3D.

3.3.1 Représentation des contextes 3D de motifs d'ARN

Soit G_O une k-extension non contractée, S le sous-ensemble de sommets de l'occurrence de motif à partir duquel est définie la k-extension (voir chapitre 2, section 2.2.2), et $G_O^{\mathscr{O},A_P-}$ le sous-graphe de séquence de G_O (voir chapitre 2, définition 2.2.4).

Nous considérons la *sous-structure 3D locale* associée à cette k-extension, contenant tous les nucléotides correspondant aux sommets qui se trouvent, dans $G_O^{\mathscr{O},A_P-}$, à une distance strictement inférieure à k d'un sommet de S. Rappelons-le, ces sommets induisent des chemins formés uniquement d'arcs covalents contenant au moins un sommet de l'occurrence de motif (voir sommets en vert dans la Figure 3.2). Cela signifie que les nucléotides correspondants sont reliés entre eux par des interactions covalentes.

Nous pouvons noter que la plupart des sous-structures 3D locales, pour un motif donné et pour une valeur de k donnée, auront le même nombre de nucléotides.

Pour comparer deux sous-structures 3D locales, définies à partir de deux sous-graphes de séquence de k-extensions non contractées $G_{O_1}^{\mathscr{O}_1,A_P-}$ et $G_{O_2}^{\mathscr{O}_2,A_P-}$, nous définissons une *correspondance* entre les nucléotides des deux sous-structures :

- les nucléotides correspondant aux sommets de même type ρ_O de l'occurrence de motif dans $G_{O_1}^{\mathscr{O}_1,A_P-}$ et $G_{O_2}^{\mathscr{O}_2,A_P-}$ sont associés deux à deux. Par exemple, pour le motif A-minor, le nucléotide associé au sommet de type $\rho_O = 1$ (resp. $\rho_O = 2,3,4,5,6$) dans $G_{O_1}^{\mathscr{O}_1,A_P-}$ est mis en correspondance avec le nucléotide associé au sommet de type $\rho_O = 1$ (resp. $\rho_O = 2,3,4,5,6$) dans $G_{O_2}^{\mathscr{O}_2,A_P-}$.

- un nucléotide correspondant à un sommet dans $G_{O_1}^{\mathscr{O}_1,A_P-}$, se trouvant à une distance i ($i \in \{1,2,...,k-1\}$) du sommet de l'occurrence de motif d'une branche j, est associé au nucléotide correspondant au sommet dans $G_{O_2}^{\mathscr{O}_2,A_P-}$, se trouvant à la même distance i du sommet de l'occurrence de motif de la branche j (voir exemple des sommets x et y dans la Figure 3.2)

A partir de cette correspondance, nous pouvons *aligner* les deux sous-structures locales, c'est-à-dire les superposer de telle sorte que chaque nucléotide de l'une des structures soit superposé à son correspondant dans l'autre structure. Nous allons à présent voir comment quantifier la similarité entre ces deux sous-structures 3D locales.

3.3.2 La RMSD comme mesure de similarité entre contextes 3D

Calcul de la RMSD dans le cas général

La mesure que nous utilisons pour comparer deux sous-structures 3D locales est la RMSD (Root Mean Square Deviation) [16]. Cette mesure est la plus utilisée pour comparer des structures 3D en biologie.

La RMSD mesure la qualité d'un alignement entre deux structures 3D de molécules, en calculant les distances en Angstrom (Å) entre les atomes de ces deux structures, pris deux à deux. Plus la RMSD est faible, meilleur est l'alignement et plus les deux structures sont similaires.

Plus formellement, la RMSD est calculée selon la formule suivante :
Soit a_1 l'ensemble des n atomes de la première structure.
Soit a_2 l'ensemble des n atomes de la deuxième structure.
Chaque atome numéro i dans l'ensemble d'atomes a_1 correspond à l'atome numéro i dans l'ensemble d'atomes a_2.
Chaque $a_{1i} = (a_{1ix}, a_{1iy}, a_{1iz})$ ($i \in \{1, 2, ..., n\}$) est un triplet comportant les coordonnées de l'atome numéro i dans la première structure.
Il en va de même pour chaque a_{2i}.

$$RMSD(a_1, a_2) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}((a_{1ix} - a_{2ix})^2 + (a_{1iy} - a_{2iy})^2 + (a_{1iz} - a_{2iz})^2)}$$

Le calcul de la RMSD nécessite que les deux structures possèdent le même nombre d'atomes. Pour comparer des structures 3D d'ARN dont les séquences primaires sont différentes, on représente souvent chaque nucléotide par un seul atome du squelette ribose-phosphate. Dans notre étude, nous représenterons chaque nucléotide par l'atome C3' du ribose, comme cela est souvent le cas dans la littérature [20, 71].

Calcul de la RMSD pour les sous-structures 3D locales

Pour comparer deux sous-structures 3D locales, nous alignons d'abord uniquement les nucléotides de l'occurrence de motif deux à deux, de manière à minimiser la RMSD. Puis, nous conservons cet alignement, c'est-à-dire la position de chaque atome dans l'espace pour les deux structures, et calculons la RMSD associée aux sous-structures locales complètes, c'est-à-dire avec les nucléotides qui n'appartiennent pas à l'occurrence de motif.

Comme dit précédemment, pour une valeur de k donnée, ces sous-structures locales auront, en général, toutes le même nombre de nucléotides ($|S| \times k + c$, avec S le sous-ensemble de sommets de l'occurrence du motif considérés pour étendre le motif, et c le nombre de sommets de l'occurrence du motif non considérés pour étendre le motif), et pourront donc être comparées, chaque nucléotide étant représenté par un unique atome dans l'espace.

L'une des limitations de l'utilisation de la RMSD est cependant que la RMSD augmente rapidement lorsque la taille des structures 3D considérées augmente. Nous verrons par la suite qu'il faut alors prendre cela en considération si on souhaite comparer les valeurs de RMSD obtenues pour des valeurs de k différentes.

3.4 Détection des homologues

Pour caractériser les occurrences de motif, nous avons besoin de déterminer des relations d'évolution entre occurrences de motifs. Des molécules dérivant d'un même ancêtre commun, que l'on appelle alors homologues, possèdent des structures 3D souvent très conservées, en particulier pour les sous-structures qui ont un rôle fondamental dans la fonction de la molécule. Ces molécules sont apparentées d'un point de vue structural et fonctionnel. Ainsi, dans le cas d'occurrences de motif se trouvant dans des molécules homologues et s'alignant dans les structures 3D, les similarités de contexte structural s'expliquent par cette parenté. En revanche, retrouver des contextes structuraux similaires parmi des occurrences qui n'ont pas cette propriété pourrait indiquer que les molécules ont convergé au cours de l'évolution pour aboutir à ce même contexte.

C'est ainsi pour faire la différence entre ces deux cas que nous avons besoin de déterminer quelles occurrences sont homologues. Des séquences de molécules biologiques sont souvent considérées comme homologues lorsqu'elles présentent plus de similarité que ce qui serait attendu par hasard [86]. Cependant, nous avons trop peu de données pour pouvoir effectuer une étude statistique de ce type. Nous allons alors utiliser d'autres méthodes.

Nous allons répartir les occurrences de motifs en *groupes d'homologie*, notion que nous définissons de la manière suivante :

Définition 3.4.1 ***Groupe d'homologie***
Deux occurrences de motif sont dans le même groupe d'homologie si et seulement si :

- *ces deux occurrences appartiennent à des molécules homologues et,*
- *ces deux occurrences sont alignées dans les structures 3D.*

Nous utilisons deux méthodes pour détecter des occurrences homologues, que nous expliquons dans les deux paragraphes suivants.

Alignements séquence-structure de Gutell

Pour détecter les occurrences homologues de nos jeux de données, nous avons utilisé tout d'abord les alignements séquence-structure de Gutell [18] (récupérés en avril 2021). Ce sont des alignements de séquence de molécules homologues dans lesquels des informations de structure ont été utilisées pour

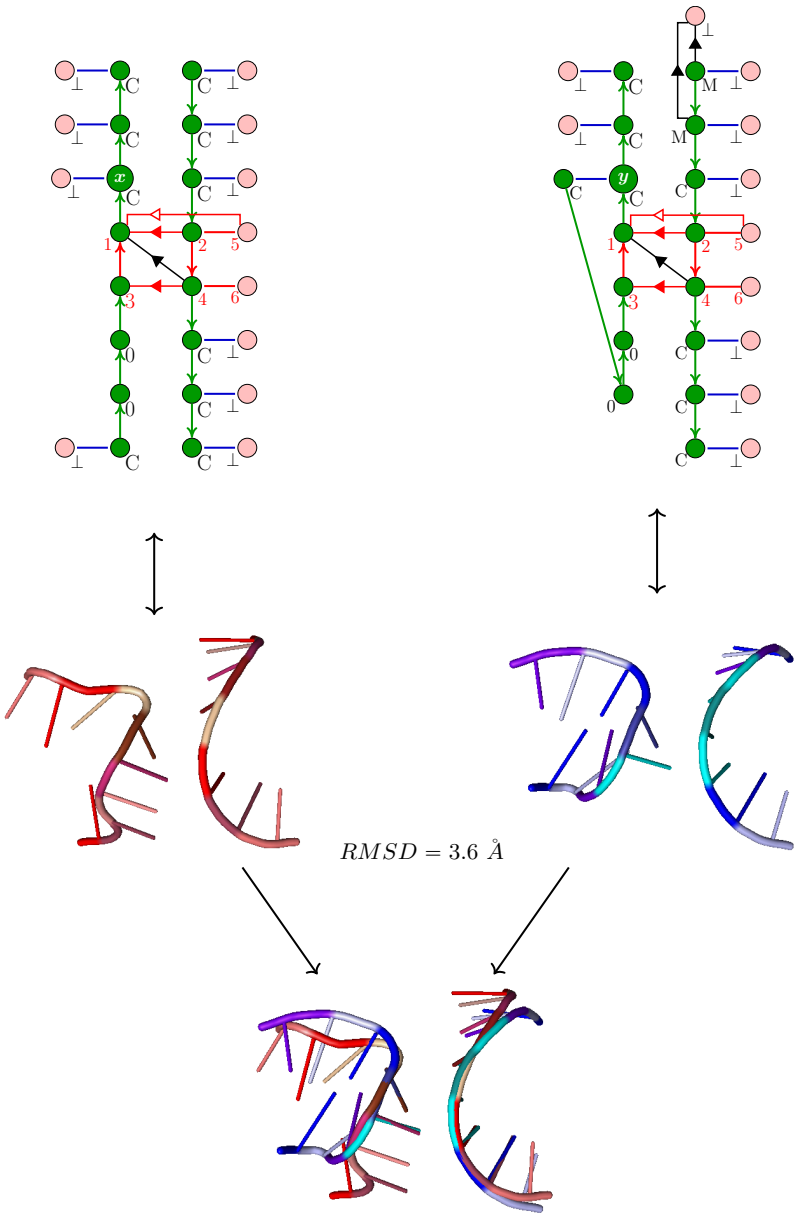


Figure 3.2 – Exemple de comparaison 3D entre deux sous-structures 3D locales associées à deux 4-extensions. En haut, sont présentés les deux 4-extensions non contractées. Les sommets sont annotés par leur type ρ_O. Les sommets en vert dans les 4-extensions correspondent aux nucléotides des sous-structures 3D locales. Par exemple, les deux sommets des occurrences de motif de type 1 (resp. 2,3,4,5,6) en rouge sont mis en correspondance, de même que le sommet x de la 4-extension de gauche et le sommet y de la 4-extension de droite, étant tous deux à distance 1 du sommet 1 de l'occurrence de motif. Au milieu, les sous-structures 3D locales sont représentées avec une couleur différente pour chaque nucléotide (seuls le squelette ribose-phosphate et l'atome de la base azotée relié au ribose sont représentés). En bas, est présenté l'alignement 3D des deux structures, avec la valeur de RMSD correspondante. Cette valeur de RMSD est assez élevée pour refléter les différences observées surtout dans la partie gauche des deux sous-structures 3D.

en améliorer la qualité. Ces alignements ont été réalisés par l'expertise humaine à partir des connaissances biologiques d'un certain nombre de familles d'ARN (ARN ribosomiques et ARN de transfert surtout).

Des occurrences de motif alignées dans ces alignements seront considérées comme homologues dans notre étude. Cependant, toutes les structures PDB de nos jeux de données ne se trouvent pas dans ces alignements. En particulier, certaines familles d'ARN, comme les ribozymes ou les riboswitches, n'y sont pas du tout représentées. Pour le motif A-minor, 75% des occurrences de motif se trouvent dans ces alignements, tandis que 40% (resp. 11%) seulement des occurrences de motif G (resp. du motif trans WC/H) s'y trouvent. Nous avons alors développé une autre méthode de détection d'occurrences homologues, basée également sur la recherche de similarité de séquence et de structure.

Utilisation d'autres alignements de séquence et de structure

Dans le but de détecter les occurrences homologues non présentes dans les alignements de Gutell, nous avons donc recherché des similarités de structure et de séquence entre les contextes des occurrences de motif issues de molécules homologues (i.e. de la même famille d'ARN).

Pour les similarités de structure, nous avons utilisé la RMSD entre sous-structures 3D locales définies dans la section 3.3. Nous considérons toutes les k-extensions avec $k = 4$, et calculons la RMSD pour toutes les paires de sous-structures 3D locales associées à ces k-extensions. Nous justifierons la valeur de k choisie par la suite (section 3.6).

Alignements de séquence. Pour effectuer des alignements de séquence, nous allons définir quelles séquences nous utilisons. Pour cela, nous définissons tout d'abord un sous-graphe d'ARN particulier.

Définition 3.4.2 ***Graphe de sous-séquence primaire*** $\mathbf{G_p}$

Soit un graphe $G = (V, A_P, A_H)$ *d'ARN et une occurrence* O *de motif dans* G*.*

Le ***graphe*** $\mathbf{G_p} = (\mathbf{V_P}, \mathbf{A_{P,30}})$ *est un sous-graphe partiel de* G *tel que* $V_P \subseteq V$ *contient tous les sommets appartenant à un chemin formé uniquement d'arcs covalents de* A_P *contenant 30 sommets, et ayant comme extrémité initiale ou finale un sommet de* O*.*

Ce sous-graphe G_p consiste donc en un ensemble de chemins correspondant à une partie de la séquence primaire de la molécule représentée par le graphe G d'ARN. À chaque sommet peut être associé le type de nucléotide (A,C,G,U) correspondant. Nous considérons alors les séquences en nucléotides induites par ces chemins. Notons que nous considérons ici l'ensemble des sommets d'une occurrence de motif et non pas uniquement les sommets appartenant au sous-ensemble S permettant de construire la k-extension (voir chapitre 2, section 2.2.2).

Pour chaque occurrence de motif A-minor, nous avons donc en général 3 séquences de taille 60 : l'une contenant les nucléotides correspondant aux

Paramètre	Valeur
Matrice de substitution	EDNAFULL
Pénalité d'ouverture d'un gap	10
Pénalité d'extension d'un gap	0.5

Table 3.4 – Paramètres utilisés pour l'alignement global de séquence avec l'algorithme de Needleman-Wunsch

sommets 1 et 3 de la Figure 3.1a, la deuxième contenant les nucléotides correspondant aux sommets 2 et 4 de la Figure 3.1a, et la troisième contenant les nucléotides correspondant aux sommets 5 et 6 de la Figure 3.1a. De la même façon, pour chaque occurrence de motif G, nous avons en général 3 séquences de tailles respectives 60, 60 et 59, et pour chaque occurrence de motif trans WC/H, nous avons en général 3 séquences de taille 59. Des séquences plus courtes peuvent exister si l'occurrence de motif se trouve près d'une extrémité de la séquence primaire de la molécule, ou bien si l'occurrence se trouve dans une petite molécule d'ARN.

Nous effectuons alors un alignement global des paires de séquences pour chaque paire d'occurrences de motif, à l'aide de l'algorithme de Needleman-Wunsch [78]. Dans notre cas, il y a donc 3 alignements effectués pour chaque paire d'occurrences. Nous obtenons ainsi un score d'alignement de séquence pour chaque paire testée, dépendant de paramètres standard utilisés par l'algorithme de Needleman-Wunsch, qui sont décrits dans la Table 3.4. Pour la suite, nous considérerons le score minimum entre les 3 alignements effectuées pour une paire d'occurrences de motif, c'est-à-dire la paire de séquences la moins similaire.

Définitions de seuils sur ces mesures. Pour déterminer les paires d'occurrences de motifs homologues, nous définissons deux seuils : un seuil sur la RMSD et un seuil sur le score d'alignement de séquence. Les paires d'occurrences ayant une RMSD inférieure au seuil et un score d'alignement de séquence supérieur au seuil seront considérées comme homologues. Pour définir ces seuils, nous observons la répartition des valeurs de RMSD et de score d'alignement. L'exemple des occurrences de motif A-minor dans les grandes sous-unités du ribosome est présenté en Figure 3.3. Nous remarquons dans cette figure deux principaux ensembles de points. Les paires d'occurrences associées à une RMSD élevée (environ supérieure à 2Å) et un score d'alignement faible (environ inférieur à 60) forment le premier ensemble de points. Le deuxième ensemble de points est constitué de paires d'occurrences ayant une RMSD environ inférieure à 2.5Å et un score d'alignement environ supérieur à 40. Ces deux ensembles de points correspondent respectivement aux occurrences non homologues et homologues. Nous avons alors choisi des seuils qui permettent de séparer ces deux ensembles. Pour choisir les seuils le plus justement possible, nous avons observé à l'oeil l'alignement 3D de structures contenant certaines paires d'occurrences associées à des valeurs de RMSD et de score proches des seuils

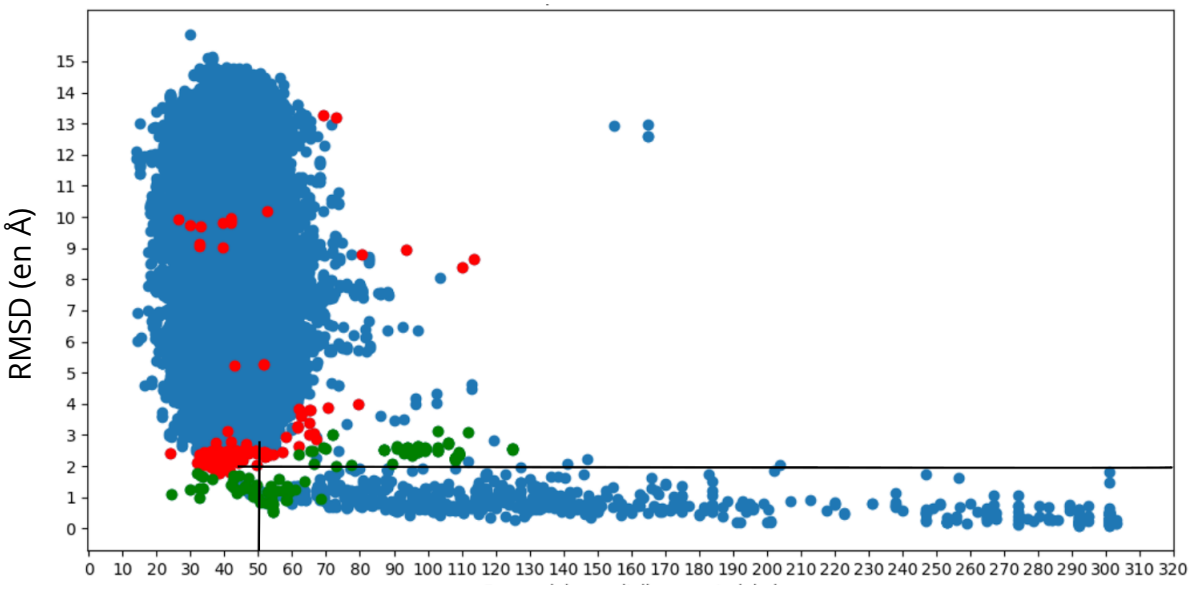
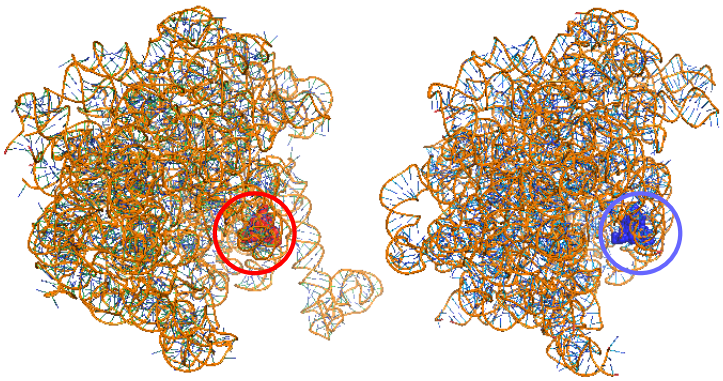


Figure 3.3 – Distribution des valeurs de RMSD et de score d'alignement de séquence pour les paires d'occurrences de motifs A-minor, se trouvant dans la grande sous-unité du ribosome (ARNr 23S, ARNR 25S, ARNr 28S). Les points verts et les points rouges correspondent à des paires d'occurrences testées manuellement. Les points verts sont des paires homologues et les points rouges des paires non homologues.



ARNr 23S de T. thermophilus (PDB : 4Y4O)

ARNr 23S de H. marismortui (PDB : 2QEX)

Figure 3.4 – Représentation de deux structures 3D de molécules contenant chacune un motif A-minor (entouré en rouge à gauche et en bleu à droite). Ces occurrences de motif sont homologues car elles sont alignées dans les structures 3D.

potentiels. Si les occurrences de motifs sont alignées dans ces alignements de structure, nous considérerons les occurrences comme homologues (voir un exemple dans la Figure 3.4). Ces paires testées manuellement correspondent aux points verts et aux points rouges dans la Figure 3.3. Les points verts sont des paires homologues et les points rouges des paires non homologues.

Nous avons ainsi défini un seuil de RMSD α_{rmsd} de 2Å et un seuil de score d'alignement de séquence α_{score} de 50, choisis pour le motif A-minor et restant cohérents pour le motif G, pour toutes les familles d'ARN comportant ces deux motifs. Ces seuils sont choisis expérimentalement à partir des répartitions de valeurs de RMSD et de scores d'alignement de séquence. A partir de ces seuils, nous construisons un graphe non orienté, que nous appellerons *graphe d'homologie*, dans lequel un sommet correspond à une occurrence de motif et une arête est présente entre deux sommets si la valeur de RMSD correspondante est inférieure au seuil α_{rmsd} et la valeur de score minimum d'alignement est supérieure au seuil α_{score}. Un exemple d'un tel graphe est présenté en Figure 3.5, toujours pour les occurrences de motif A-minor trouvées dans une grande sous-unité du ribosome. Deux occurrences sont alors considérées comme homologues si les sommets correspondants appartiennent à la même composante connexe, dans le graphe d'homologie. Notons qu'avec les seuils ainsi définis, les composantes connexes obtenues sont denses, comme illustré dans la Figure 3.5.

Ce choix de seuils permet d'obtenir des résultats cohérents pour les différentes familles d'ARN du motif A-minor et du motif G. Pour en valider la cohérence biologique, nous avons comparé les groupes d'homologie ainsi obtenus, avec ceux obtenus par l'approche des alignements de Gutell, pour les familles d'ARN présentes dans les alignements de Gutell (ARN ribosomiques,

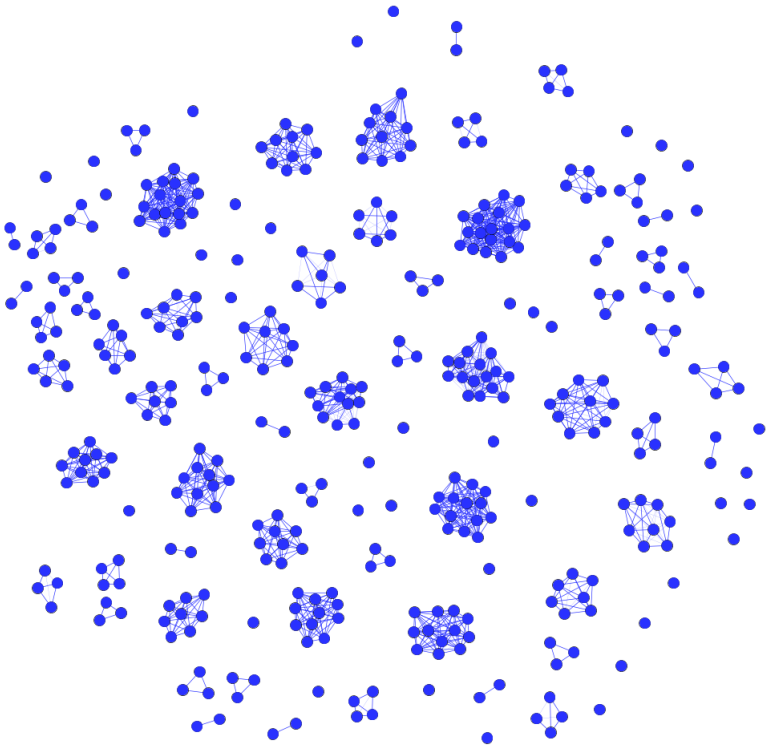


Figure 3.5 – Graphe d'homologie des occurrences de motif A-minor, trouvées dans une grande sous-unité de ribosomes (ARNr 23S, ARNr 25S, ARNr 28S). Chaque sommet correspond à une occurrence de motif, et il y a une arête entre deux sommets si la RMSD entre les sous-structures 3D des deux occurrences est inférieure au seuil α_{rmsd}, et si le score minimum d'alignement entre les séquences des deux occurrences est supérieur au seuil α_{score}. Chaque composante connexe représente un groupe d'occurrences homologues.

ARN de transfert et introns). Les deux classifications d'homologues sont cohérentes dans la plupart des cas, avec seulement 7 paires d'occurrences sur 35 000 non correctement classifiées pour le motif A-minor, et 16 paires d'occurrences sur 2000 non correctement classifiées pour le motif G.

Il est plus difficile de choisir des seuils pour le motif trans WC/H, surtout pour les occurrences se trouvant dans des ARN de transfert. Les occurrences du motif trans WC/H sont en effet en majorité présentes dans des ARN de transfert (voir Table 3.2). Les ARN de transfert sont de petites molécules (75 à 90 nucléotides), qui, bien qu'appartenant à la même famille d'ARN, possèdent de grandes diversités de séquence et de structure. C'est la raison pour laquelle notre approche de détection d'homologues donne de moins bons résultats pour cette famille d'ARN. Nous avons ainsi considéré comme molécules homologues, non pas tous les ARNt, mais seulement les ARNt fixant le même acide aminé, comme c'est le cas dans les alignements de Gutell. En appliquant la recherche de similarités de séquence et de structure sur les occurrences trouvées dans des ARNt homologues, nous obtenons des résultats cohérents

avec ceux obtenus avec les alignements de Gutell (toutes les paires d'occurrences présentes dans les alignements de Gutell sont correctement classifiées avec notre méthode).

3.5 Classification des contextes structuraux

Pour déterminer des caractéristiques structurales communes aux contextes structuraux des motifs que nous étudions, nous classifions les occurrences à partir des deux métriques que nous avons définies dans le chapitre 2 : la métrique de similarité contextuelle sur les k-extensions, et la RMSD sur les structures 3D locales.

Cette partie 3.3 présentera les graphes dans lesquels nous allons calculer de telles classifications. Puis elle décrira brièvement la méthode de clustering utilisée pour obtenir ces classifications, et les raisons du choix de cette méthode.

3.5.1 Définition de classifications dans deux graphes particuliers

Nous allons ici définir les graphes dans lesquels seront calculées les classifications.

Définition 3.5.1 ***Graphe de similarité de seuil*** s

Soit s *une valeur de seuil comprise entre 0 et 1.*

Le graphe de similarité $G_s = (V_s, E_s, \omega)$ *est un graphe non orienté dans lequel chaque sommet représente une occurrence de motif et son contexte structural topologique, et il y a une arête entre deux sommets si la valeur de la similarité contextuelle entre les deux k-extensions correspondantes est supérieure au seuil* s*. Chaque arête est pondérée par la valeur de similarité contextuelle. Cette pondération est dénotée par la fonction* $\omega : E_s \rightarrow [0, 1]$*.*

Définition 3.5.2 ***Graphe de RMSD de seuil*** r

Soit r *une valeur de seuil supérieure à 0.*

Le graphe de RMSD $G_r = (V_r, E_r, \omega)$ *est un graphe, dans lequel chaque sommet représente une occurrence de motif et son contexte structural, et il y a une arête entre deux sommets si la valeur de RMSD entre les deux sous-structures 3D locales correspondantes est inférieure au seuil* r*. Chaque arête est pondérée par la valeur de RMSD. Cette pondération est dénotée par la fonction* $\omega : E_r \rightarrow \mathbb{R}^+$*.*

Une classification dans un graphe de similarité G_s (resp. dans un graphe de RMSD G_r) est alors un ensemble de sous-ensembles de sommets $W = \{V_{s1}, V_{s2}, ..., V_{sn}\}$ (resp. $W = \{V_{r1}, V_{r2}, ..., V_{rn}\}$) tel que $V_s = \bigcup_{i=1}^{n} V_{si}$ (resp. $V_r = \bigcup_{i=1}^{n} V_{ri}$).

Au sein d'une classe, nous souhaitons que les occurrences de motif possèdent des contextes (topologiques ou 3D) similaires. Cela se traduit dans le graphe de similarité ou le graphe de RMSD par une forte densité du

sous-graphe correspondant à une classe, et par une similarité moyenne élevée (ou une RMSD moyenne faible) au sein de la classe. La méthode de clustering décrite ci-après permet de correspondre à ces critères.

Nous souhaitons également que la classification ainsi définie forme une couverture du graphe de départ et non une partition, ce qui signifie qu'un sommet peut appartenir à plusieurs classes distinctes. Cela nous permettra de mettre en lumière les cas où le contexte (topologique ou 3D) d'une occurrence de motif est similaire aux contextes de deux autres occurrences de motif, alors que ces deux autres occurrences possèdent, quant à elles, des contextes moins similaires.

3.5.2 Description de la méthode de clustering utilisée

Pour obtenir de telles classifications, nous avons donc besoin d'une méthode autorisant un sommet à appartenir à plusieurs classes, et même le favorisant.

Pour ce faire, nous avons choisi une méthode décrite dans [89], appelée OClust-R. Par rapport aux autres méthodes de clustering recouvrant, cette méthode a l'avantage d'avoir une faible complexité en temps ($O(n^2)$ pour n sommets), d'être facile à implémenter, et d'être composée d'une étape de post-traitement permettant de limiter le nombre de recouvrements, c'est-à-dire le nombre de classes auquel peut appartenir un même sommet. D'autres méthodes, comme [5] ou [108], ont tendance à générer de nombreux clusters et de nombreux recouvrements, et d'autres encore possèdent une complexité exponentielle [82]. Enfin, des méthodes comme la version recouvrante de k-moyennes [120] nécessite de choisir le nombre de clusters au préalable, ce qui n'est pas le cas dans OClust-R.

Cette méthode prend en entrée un graphe non orienté dont les arêtes sont pondérées, et un seuil sur ces valeurs. On peut donc choisir en entrée un graphe de similarité G_s avec le seuil s, ou un graphe de RMSD G_r avec le seuil r. Cette méthode nécessite donc de choisir un seuil de valeurs au préalable. Cependant, cette particularité nous a permis d'étudier plusieurs seuils possibles comme nous le verrons dans la partie 3.6.

Le principe de la méthode est le suivant. Dans un graphe de similarité G_s de seuil s (ou un graphe de RMSD G_r de seuil r), des sommets sont tout d'abord choisis pour être centres de classe. Un sommet est un bon candidat pour être un centre de classe s'il possède un fort degré et des valeurs de similarité contextuelle avec ses voisins plus élevées que ses voisins entre eux (ou des valeurs de RMSD plus faibles que ses voisins entre eux, si l'entrée est un graphe de RMSD). Une classe est ensuite définie par un sommet centre et tous ses voisins dans le graphe de départ, si le sommet centre n'appartient pas déjà à une autre classe et si la majorité des sommets voisins n'appartiennent pas également à une autre classe. Choisir un sommet centre selon son degré et la valeur de pondération moyenne des arêtes qui le relient à ses voisins, permet donc de couvrir un nombre maximal de sommets par classe, et permet d'assurer que la valeur moyenne des arêtes de la classe soit également maximisée. Une étape de post-traitement pour finir permet de supprimer les

classes redondantes pour minimiser les recouvrements et le nombre de classes.

Ainsi, cette méthode permet d'obtenir une couverture du graphe de départ dans laquelle les recouvrements sont limités. Elle ne garantit pas que les classes aient une densité maximale ou une similarité contextuelle moyenne maximale (ou RMSD moyenne minimale), mais choisit les centres de classes en fonction de ces caractéristiques. Comme nous le verrons dans la partie 3.6, l'utilisation de cette méthode permet d'obtenir des classes denses sur les exemples qui nous intéressent.

3.6 Choix de paramètres sur les k-extensions

Dans le but d'étudier la topologie des contextes structuraux à l'aide des k-extensions, et ainsi déterminer si cette topologie a une influence sur le contexte 3D des motifs, nous avons tout d'abord fait évoluer différents paramètres sur les k-extensions et choisi des valeurs selon la corrélation avec la similarité 3D et selon le gain en temps d'exécution. Comme précisé dans l'introduction de ce chapitre, les algorithmes ont été implémentés en Python3.

3.6.1 Contraction

Nous avons tout d'abord comparé les k-extensions contractées et non contractées, selon le temps de calcul nécessaire pour obtenir les sous-graphes communs maximum de toutes les paires de k-extensions, et selon la corrélation des valeurs de similarité contextuelle avec les valeurs de similarité 3D. Comme évoqué dans le chapitre 2 (section 2.2.3), contracter certains sommets et certains arcs des k-extensions a pour but de rendre le modèle plus flexible, en autorisant à considérer comme similaires certaines différences d'un ou deux nucléotides dans une boucle ou une hélice. En effet, la structure 3D n'est souvent que peu modifiée, voire pas modifiée du tout, par de légères différences comme celles-ci.

Dans cette partie, nous allons voir si la représentation contractée présente effectivement des avantages.

Nous présenterons ici les résultats obtenus avec une taille d'extension $k = 4$, car, comme nous le verrons dans la sous-section 3.6.2, cette taille donne les meilleurs résultats pour le motif A-minor, et des résultats très proches des meilleurs résultats pour les deux autres motifs. Cependant, nous pouvons noter que des conclusions similaires peuvent être tirées des autres tailles d'extensions que nous avons étudiées.

Comparaison des temps d'exécution entre les k-extensions contractées et non contractées

Nous comparons d'abord les temps d'exécution des deux algorithmes de recherche d'un MCES (méthode exacte et heuristique), pour les k-extensions

contractées et non contractées. En effet, la recherche d'un MCES est la partie possédant la plus grande complexité en temps.

Comme on peut le voir dans la Table 3.5, le gain de temps avec les extensions contractées est considérable en particulier si on utilise la méthode exacte. L'écart est évidemment moins important si on utilise la méthode heuristique. Cependant, si l'on souhaite effectuer cette recherche sur un plus grand nombre de données, comme sur tous les motifs structuraux d'une base de données par exemple, ce gain de temps peut être plus important.

Motif structural	Méthode exacte		Méthode heuristique	
	Temps d'exécution (en min) pour les 4-extensions contractées	Temps d'exécution (en min) pour les 4-extensions non contractées	Temps d'exécution (en min) pour les 4-extensions contractées	Temps d'exécution (en min) pour les 4-extensions non contractées
motif A-minor (391 occurrences)	240	960	13,5	17,5
motif G (159 occurrences)	48	132	1,8	2,8
motif trans WC/H (87 occurrences)	13	27	1,3	1,7

Table 3.5 – Temps d'exécution de l'algorithme de recherche d'un MCES pour toutes les paires de 4-extensions (contractées ou non contractées) selon les deux méthodes (exacte et heuristique) et pour les trois motifs étudiés, sur un PC Intel Core i5-7440HQ 4x2.80GHzCPU (implémentation en Python3).

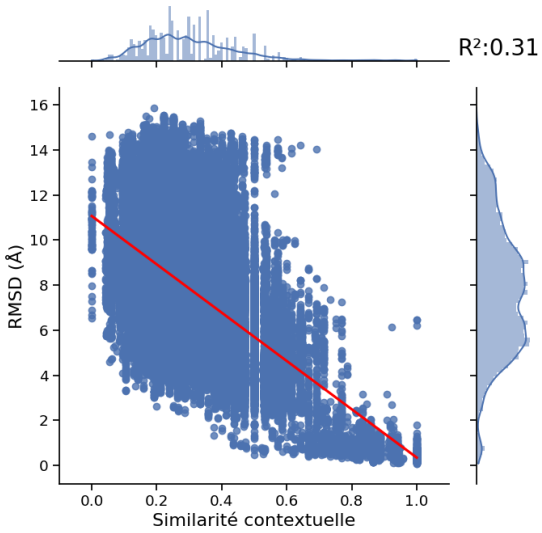
Comparaison des représentations contractées et non contractées selon leur corrélation avec la similarité 3D

Nous comparons ensuite les valeurs de similarité contextuelle des k-extensions contractées et non contractées, aux valeurs de similarité du contexte structural en 3D à l'aide de la RMSD.

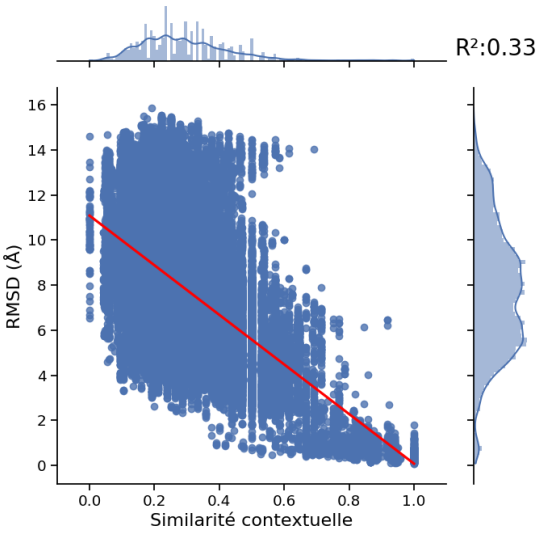
Pour les trois motifs que nous étudions, nous pouvons voir dans la Figure 3.6 que les valeurs de similarité contextuelle pour les deux représentations (contractées et non contractées) ont une distribution similaire, par rapport à la RMSD.

En observant plus précisément les différences entre les deux représentations, on peut remarquer que les valeurs de similarité contextuelle, bien que différentes pour 14,2% des paires de k-extensions pour le motif A-minor (resp. 27% pour le motif G, et 7% pour le motif trans WC/H), ne diffèrent en moyenne que de 0.06 (resp. 0.07 et 0.064), pour des valeurs comprises entre 0 et 1.

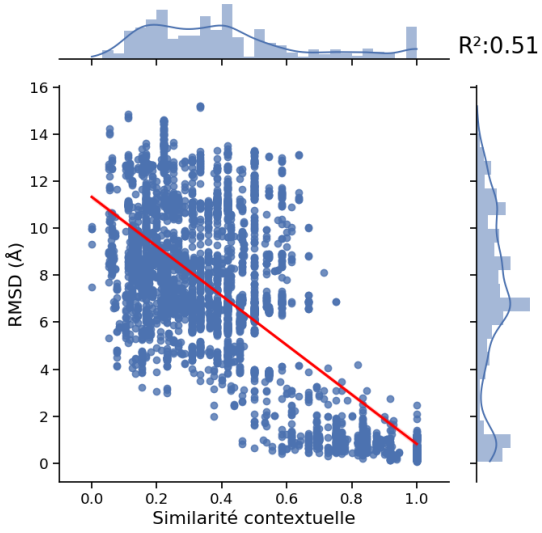
Ainsi, cette étude montre que les similarités de contexte 3D de ces motifs semblent pareillement capturées par la représentation contractée et par la représentation non contractée.



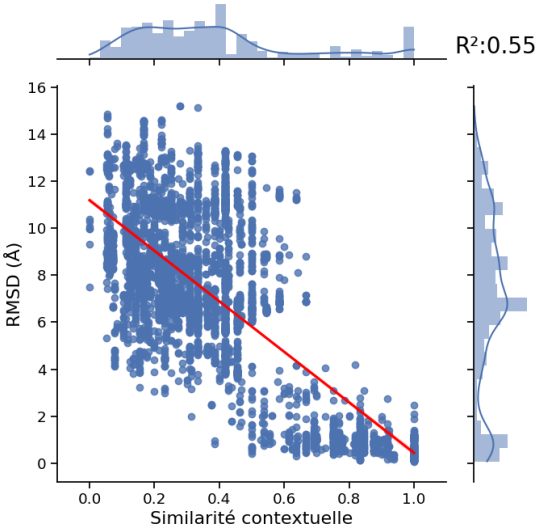
(a) Motif A-minor, 4-extensions non contractées



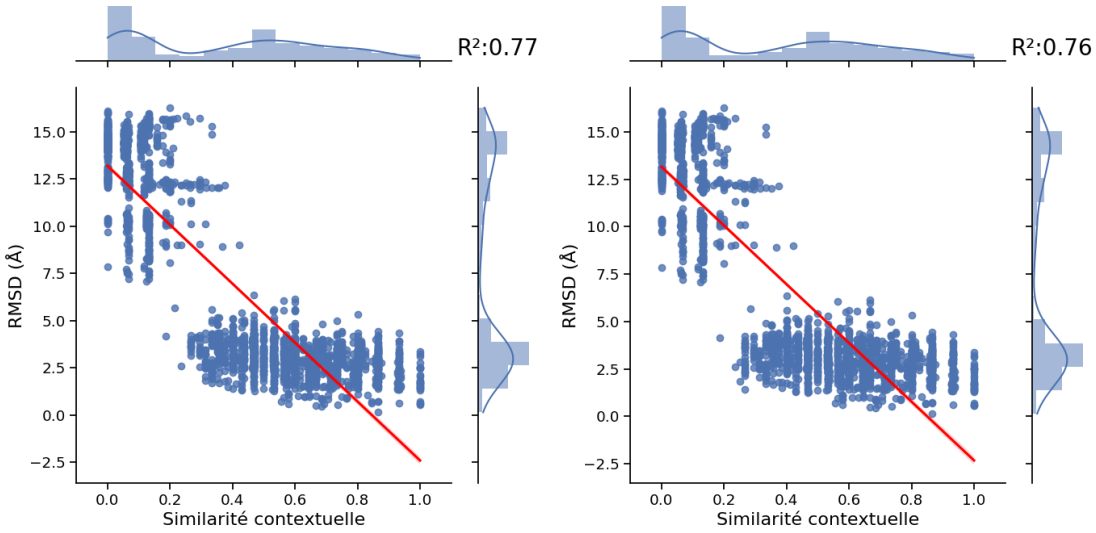
(b) Motif A-minor, 4-extensions contractées



(c) Motif G, 4-extensions non contractées



(d) Motif G, 4-extensions contractées



(e) Motif trans W/H, 4-extensions non contractées

(f) Motif trans W/H, 4-extensions contractées

Figure 3.6 – Distribution des valeurs de RMSD en fonction des valeurs de similarité contextuelle pour toutes les paires de 4-extensions contractées ou non contractées, dans chaque jeu de données de motif. Chaque point des différents graphiques correspond à une paire de 4-extensions. L'histogramme et la courbe au-dessus de chaque graphique indiquent la fonction de densité gaussienne des valeurs de similarité contextuelle, obtenue par la méthode d'estimation par noyau. L'histogramme et la courbe à droite de chaque graphique indiquent la fonction de densité des valeurs de RMSD, de la même façon. En rouge est indiquée la droite de corrélation sur chaque graphique, et le coefficient de corrélation R^2 associé est indiqué en haut à droite de chaque figure.

Nous pouvons ainsi conclure de cette première comparaison, que la représentation contractée présente un avantage du point de vue computationnel, et n'occasionne pas de perte d'information par rapport à la représentation non contractée.

3.6.2 Taille d'extension

Nous avons également étudié l'évolution des résultats de similarité contextuelle et de similarité 3D en fonction de la taille d'extension k.

Pour cela, de la même façon que pour comparer les k-extensions contractées et les k-extensions non contractées, nous avons étudié la corrélation entre les valeurs de similarité contextuelle des k-extensions contractées et les valeurs de RMSD, pour différentes valeurs de k, allant de 2 à 10 pour que les graphes étudiés aient une taille limitée.

Comme évoqué dans la section 3.3, la RMSD ne peut se calculer qu'entre des sous-structures 3D ayant le même nombre de nucléotides. Or, même pour une valeur de k donnée, les sous-structures 3D associées aux k-extensions n'ont pas toutes le même nombre de nucléotides. En effet, dans certains cas, les motifs étudiés relient des régions proches sur la séquence, et il y a parfois moins de $2k$ nucléotides consécutifs sur la séquence entre deux nucléotides du motif. De plus, le motif peut apparaître en début ou en fin de chaîne de molécule. Nous avons alors choisi d'exclure ces cas particuliers de cette étude sur les tailles d'extension. Notons cependant qu'ils ne sont pas exclus de l'étude globale de la corrélation entre la similarité contextuelle et la RMSD (voir section 3.7).

Le nombre de ces cas particuliers augmente lorsque la taille d'extension augmente. Ainsi, pour comparer les résultats des différentes tailles d'extension, nous avons considéré un sous-ensemble de paires de contextes 3D d'occurrences de motif, pour lequel la RMSD était calculable pour toute taille d'extension. Le nombre de paires considérées pour chaque motif est indiqué dans la Table 3.6.

De plus, la RMSD augmentant rapidement lorsque le nombre d'atomes augmente, nous avons normalisé les valeurs de RMSD en divisant par le nombre de nucléotides dans les sous-structures, pour que les valeurs de RMSD soient comparables d'une taille à l'autre.

Nous allons ainsi d'abord présenter les résultats obtenus sur les motifs G et trans WC/H, qui évoluent de façon similaire avec la taille d'extension. Puis, nous traiterons le cas du motif A-minor, dont les résultats diffèrent des deux autres motifs.

Cas du motif trans WC/H et du motif G

On peut remarquer dans la Figure 3.7, que le coefficient de corrélation augmente généralement avec la taille d'extension pour les deux motifs trans WC/H et G.

Motif structural	Nombre de paires d'occurrences de motif
Motif A-minor	30662/75466
Motif G	7690/12561
Motif trans WC/H	942/3741

Table 3.6 – Nombre de paires d'occurrences de motif considérées dans l'étude des tailles d'extension, sur le nombre total de paires, pour chacun des motifs étudiés.

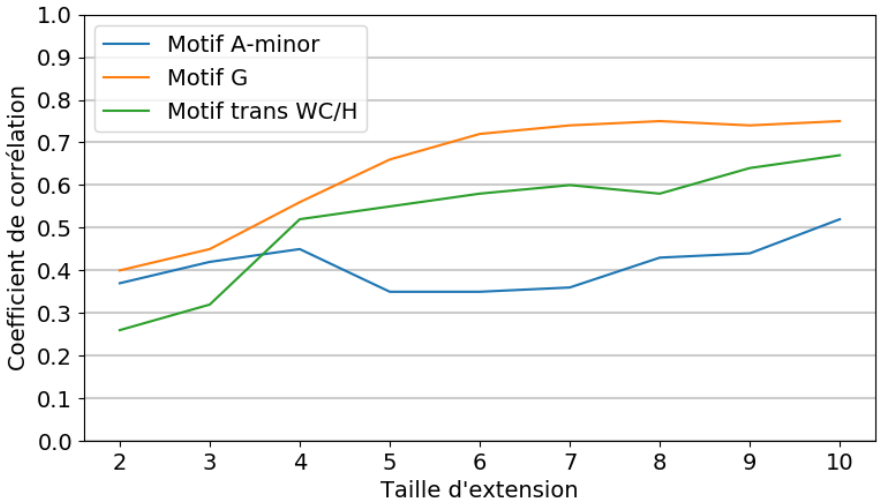


Figure 3.7 – Variations du coefficient de corrélation entre les valeurs de similarité contextuelle et de RMSD, pour des tailles d'extension k comprises entre 2 et 10, pour les trois motifs étudiés.

On peut noter également que l'évolution du coefficient de corrélation pour le motif G semble se stabiliser vers une taille d'extension de 6.

Pour ces deux motifs en particulier, il semble donc que plus la taille d'extension est élevée, plus la similarité contextuelle est cohérente avec la similarité 3D. Cependant, les classifications que l'on peut obtenir sur les différentes tailles d'extension à partir de 6 sont très proches, voire identiques pour le motif trans WC/H.

Ainsi, sachant que le but final est de classifier les occurrences de motif selon la similarité contextuelle, il n'est pas nécessaire de considérer une taille d'extension plus élevée que 6 pour ces deux motifs.

De plus, en observant les classifications obtenues, on remarque que plus la taille d'extension est élevée, plus les classes regroupent uniquement des occurrences homologues. Cette amélioration de la corrélation entre similarité contextuelle et RMSD avec la taille d'extension peut ainsi s'expliquer par ce fait. Pour une taille d'extension proche de 10 nucléotides, les contextes structuraux des occurrences homologues restent très similaires, tandis que les différences entre les contextes structuraux d'occurrences non homologues semblent s'accentuer. Ainsi, pour pouvoir observer des similarités qui seraient dues à un phénomène de convergence, une taille d'extension moins élevée est à privilégier, même si la corrélation avec la RMSD est légèrement moins bonne.

A partir de ces observations, nous avons choisi la taille d'extension 6 pour les deux motifs.

Cas du motif A-minor

Le motif A-minor semble se comporter différemment des deux autres. La corrélation entre les deux métriques est plus stable pour les différentes tailles d'extension, oscillant entre 0,35 et 0,5.

Nous pouvons noter une diminution de la corrélation pour les tailles 5 et 6, par rapport à la taille 4, puis une augmentation ensuite à partir de la taille 7. Nous n'avons pas d'explication pour ce comportement, si ce n'est l'hypothèse d'un biais statistique dû à la faible quantité de données.

Cependant, la même observation que pour les deux autres motifs peut être faite à propos des occurrences homologues : plus la taille d'extension est élevée, plus les classes formées regroupent uniquement des occurrences homologues. La taille d'extension possédant le meilleur compromis entre corrélation avec la RMSD et proportion d'occurrences non homologues est la taille $k = 4$, ce pourquoi nous l'avons choisie pour ce motif.

3.6.3 Seuil de valeur de similarité pour la classification

La méthode de classification que nous utilisons nécessite de choisir un seuil sur les valeurs de similarité, en dessous duquel les k-extensions sont considérées comme non similaires (voir section 3.5).

Pour choisir le seuil de similarité contextuelle qui est le plus cohérent avec la similarité 3D, nous avons comparé les classifications obtenues avec la similarité

contextuelle et avec la RMSD pour différents seuils.

Pour cela, nous avons utilisé l'indice de Jaccard [49, 110], qui permet de comparer la similarité et la diversité entre deux classifications d'un même ensemble d'éléments, en calculant la taille de l'intersection par rapport à la taille de l'union entre ces deux classifications.

Nous considérons la formule suivante pour le calculer :
Soient C et C' deux classifications du même ensemble E d'éléments.

n_{11} : le nombre de paires qui appartiennent à la même classe dans C et dans C'.
n_{10} : le nombre de paires qui appartiennent à la même classe dans C mais pas dans C'.
n_{01} : le nombre de paires qui appartiennent à la même classe dans C' mais pas dans C.

$$J(C, C') = \frac{n_{11}}{n_{11} + n_{10} + n_{01}}$$

L'indice de Jaccard est compris entre 0 et 1, et plus sa valeur est proche de 1 plus les classifications sont proches.

Pour chaque couple de seuils sur la RMSD et sur la similarité contextuelle, nous avons calculé l'indice de Jaccard entre les deux classifications. Les résultats obtenus sont présentés en Figure 3.9. Notons que les valeurs de RMSD ne sont ici pas normalisées, car la taille d'extension k est fixée (à 4 pour le motif A-minor et à 6 pour les deux autres), donc le nombre de nucléotides considérés dans les sous-structures 3D ne varie que peu.

Dans la Figure 3.9, nous pouvons remarquer que pour chaque motif, deux intervalles de seuils permettent d'obtenir des indices de Jaccard proches de 1. Pour des seuils faibles de similarité contextuelle et élevés de RMSD, on obtient des classifications proches, car les graphes de similarité et de RMSD considérés possèdent un grand nombre d'arêtes. Par conséquent, la classification obtenue contient peu de classes, et ces classes regroupent des contextes structuraux pouvant être très peu similaires. C'est un cas non pertinent dans notre étude, car nous cherchons à regrouper des contextes structuraux similaires, selon leur topologie ou selon la structure 3D locale. Le deuxième intervalle nous intéressera davantage : il correspond à des seuils élevés de similarité contextuelle et des seuils faibles de RMSD.

Pour le motif A-minor, le seuil de similarité contextuelle possédant la meilleure cohérence avec la RMSD est 0,75 (pour des seuils de RMSD autour de 2Å).

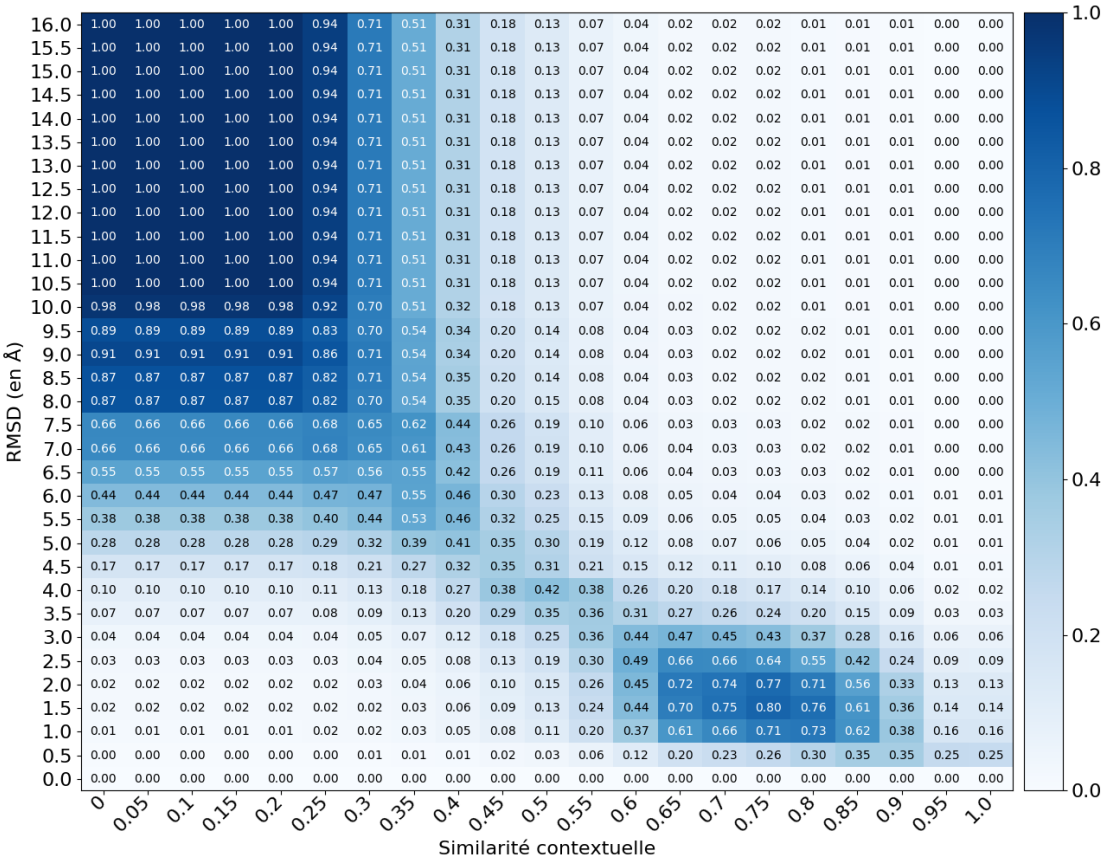
Pour le motif G (Figure 3.9b), les meilleurs indices de Jaccard sont obtenus pour des seuils de similarité contextuelle compris entre 0,55 et 0,65, avec des seuils de RMSD compris entre 3 et 5Å.

Pour le motif trans WC/H (Figure 3.9c), les meilleurs indices de Jaccard sont obtenus pour des seuils de similarité contextuelle entre 0,3 et 0,45 environ, avec des classifications identiques (indice de Jaccard égal à 1) avec ceux obtenus avec un seuil de RMSD entre 5 et 14Å.

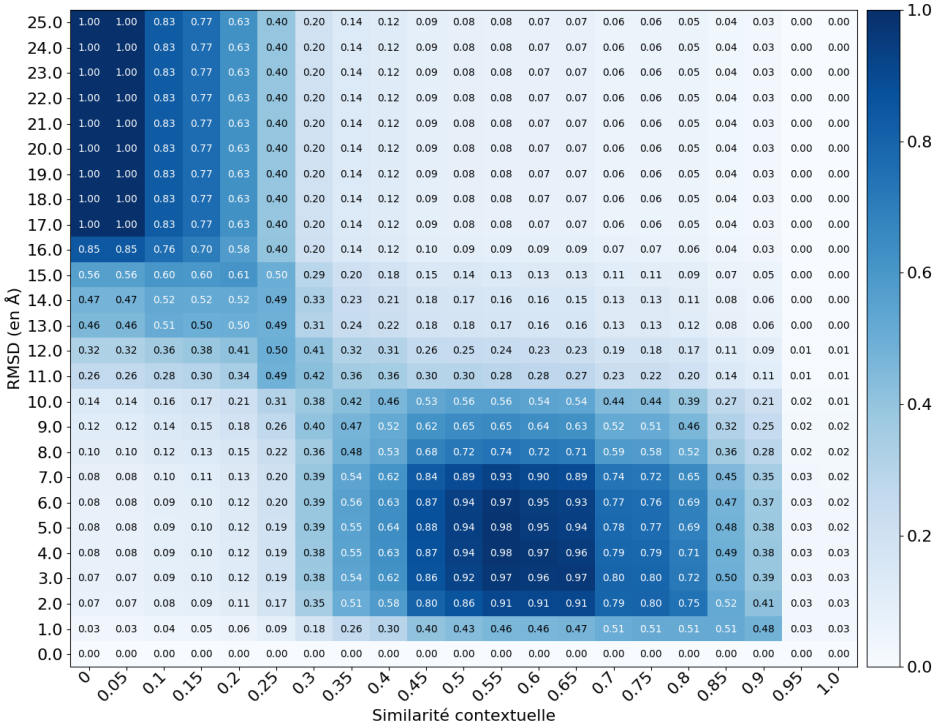
Pour nous assurer que ces seuils de similarité contextuelle induisent une classification qui soit vraiment cohérente avec la similarité 3D des contextes, nous présentons en Figure 3.10 des alignements 3D dont la valeur de RMSD est proche du seuil pour lequel la classification sur la RMSD est similaire à la classification sur la similarité contextuelle. Par exemple, pour le motif G, l'indice de Jaccard est de 0,96 entre la classification sur la similarité contextuelle avec un seuil de 0,6, et la classification sur la RMSD avec un seuil de 3Å. Si l'alignement 3D de deux sous-structures ayant une valeur de RMSD de 3Å indique que les sous-structures sont similaires, nous pourrons choisir le seuil de similarité contextuelle de 0,6.

Nous estimons que les alignements 3D que nous présentons en Figure 3.10 sont suffisamment bons pour considérer les sous-structures comme similaires.

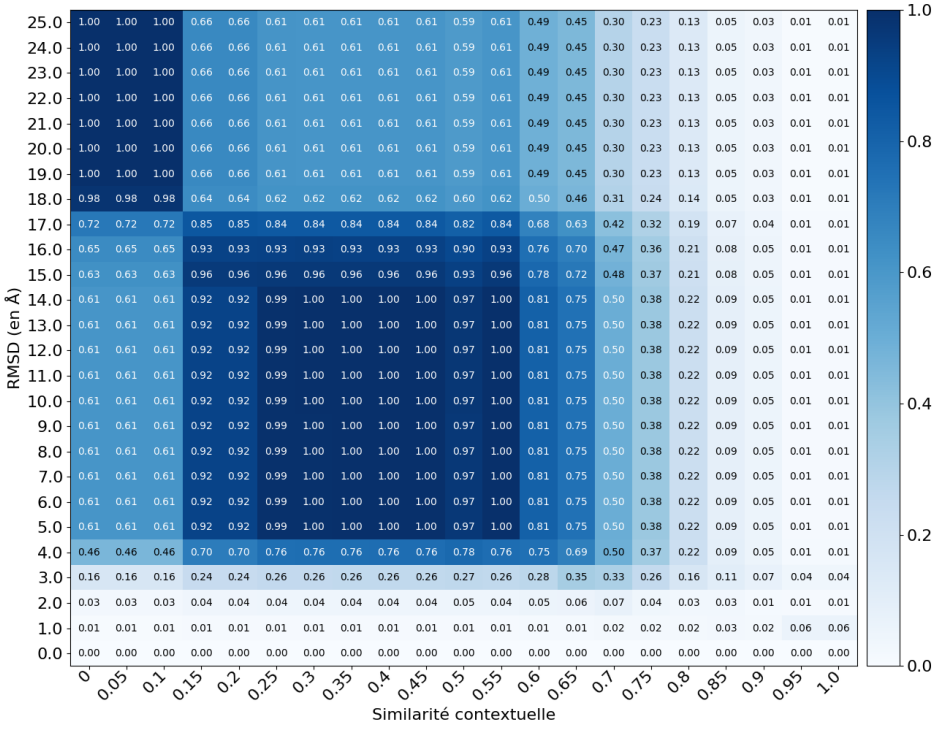
Ainsi, dans la suite de l'étude, nous utiliserons un seuil de similarité contextuelle de 0,75 pour le motif A-minor, de 0,6 pour le motif G et de 0,4 pour le motif trans WC/H.



(a) Motif A-minor (taille d'extension = 4)

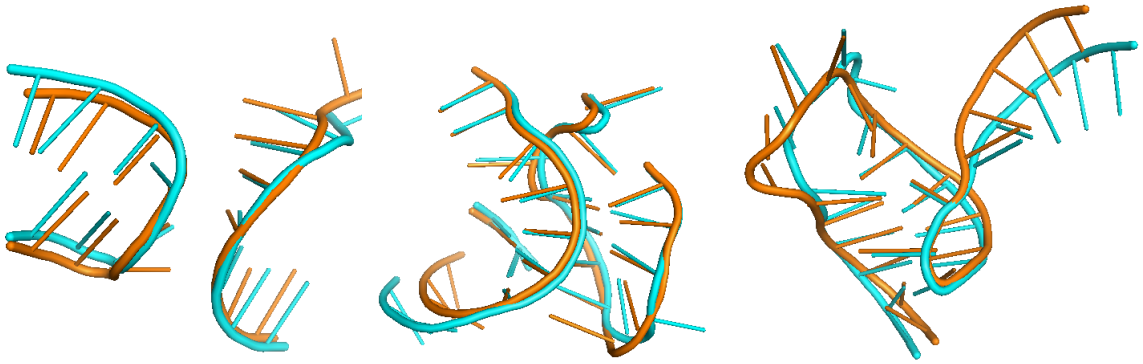


(a) Motif G (taille d'extension = 6)



(c) Motif trans WC/H (taille d'extension = 6)

Figure 3.9 – Distribution des indices de Jaccard entre des classifications obtenues avec différents seuils de similarité contextuelle et de RMSD, pour les différents motifs étudiés (en prenant la taille d'extension choisie dans la section 3.6.2).



(a) Deux sous-structures locales de motifs A-minor (taille 4), dont l'alignement a une RMSD de 2.4 Å

(b) Deux sous-structures locales de motifs G (taille 6), dont l'alignement a une RMSD de 3,05 Å

(c) Deux sous-structures locales de motifs Trans WC/H (taille 6), dont l'alignement a une RMSD de 4,99 Å

Figure 3.10 – Alignements 3D de paires de sous-structures locales des trois motifs, ayant une RMSD que l'on peut associer au seuil de similarité contextuelle choisi.

3.7 Cohérence entre similarité contextuelle et similarité 3D

A partir de ces paramètres ainsi définis sur les k-extensions, nous allons à présent comparer les valeurs de similarité contextuelle et de RMSD ainsi que les classifications avec ces deux métriques.

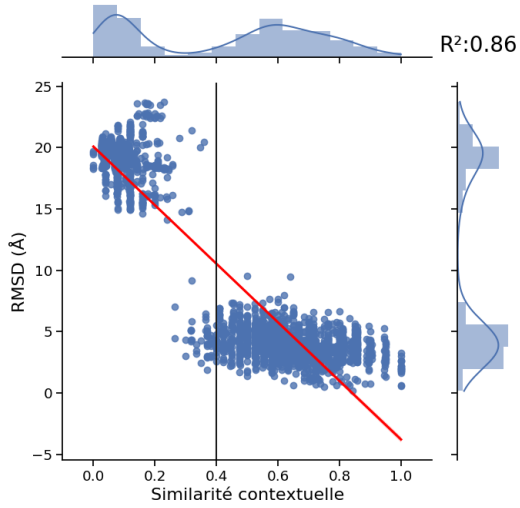
Cela nous permettra de déterminer si la topologie du contexte structural a une influence sur la structure tridimensionnelle locale autour de ces motifs.

3.7.1 Comparaison des deux métriques entre elles

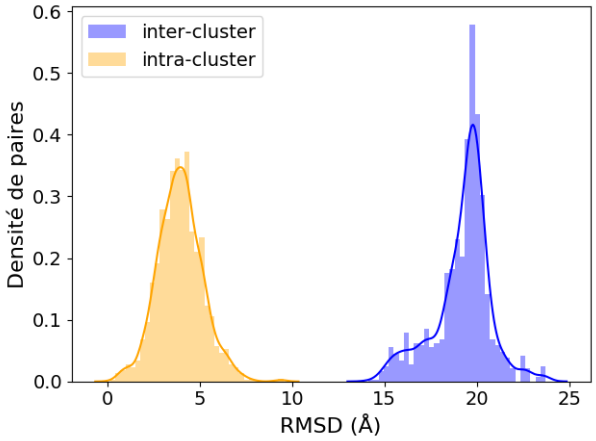
Nous avons étudié les distributions des deux métriques, avec les k-extensions contractées de taille $k = 4$ pour le motif A-minor et de taille $k = 6$ pour les deux autres motifs (voir Figure 3.11a,c,e).

Notons ici que, contrairement à l'étude réalisée pour choisir les tailles d'extension (section 3.6.2), les valeurs de RMSD ne sont pas normalisées et que sont considérées toutes les paires d'occurrences dont les sous-structures locales possèdent le même nombre de nucléotides à la taille d'extension choisie (le nombre de paires considérées pour chaque motif est indiqué dans la Figure 3.11 en légende des graphiques)

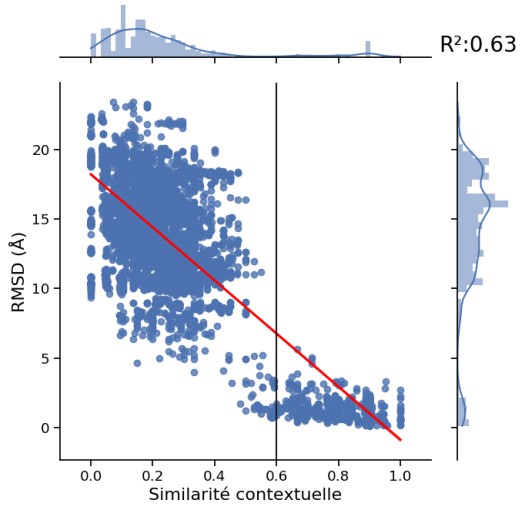
D'après la Figure 3.11, la corrélation entre les deux métriques est la plus élevée pour le motif trans WC/H (coefficient de corrélation égal à 0,86). On remarque en effet sur la Figure 3.11a deux principaux ensembles de points : l'un de ces ensembles correspond aux paires de k-extensions ayant une similarité contextuelle supérieure à 0,3 et une RMSD associée inférieure à 8Å, et l'autre ensemble correspond aux paires de k-extensions ayant une similarité



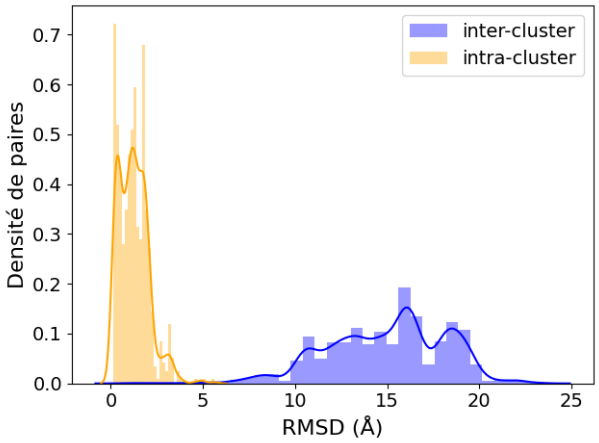
(a) Motif trans WC/H, taille d'extension $k = 6$, nombre de paires 2698



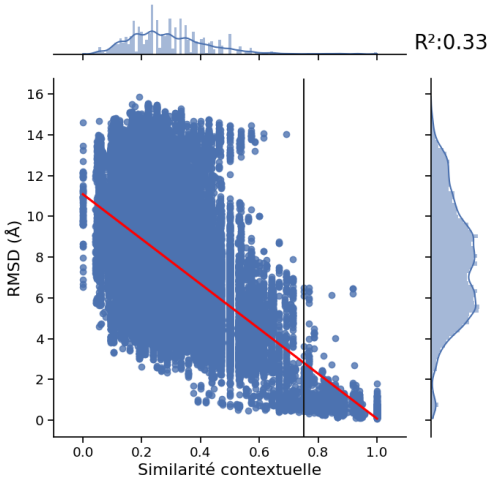
(b) Motif trans WC/H, seuil de similarité $s = 0,4$



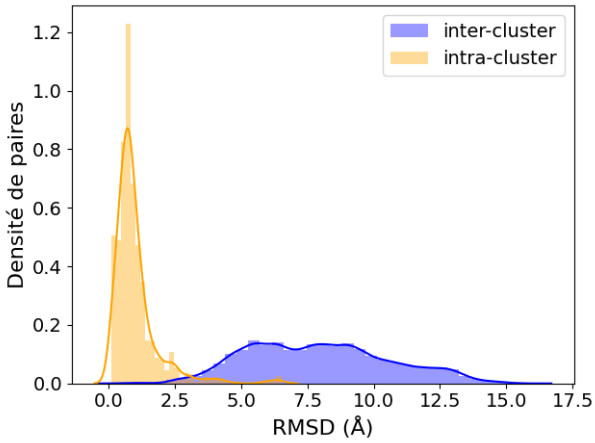
(c) Motif G, taille d'extension $k = 6$, nombre de paires 9686



(d) Motif G, seuil de similarité $s = 0,6$



(e) Motif A-minor, taille d'extension $k = 4$, nombre de paires 70898



(f) Motif A-minor, seuil de similarité $s = 0,75$

Figure 3.11 – Pour chaque motif, distribution des valeurs de RMSD en fonction des valeurs de similarité contextuelle (à gauche), et distribution des paires d'occurrences, en densité de paires (voir Figure 3.6), appartenant à une même classe (intra-cluster) ou à deux classes distinctes (inter-cluster), en fonction des valeurs de RMSD (à droite). La droite verticale en noir sur les graphiques de gauche correspond au seuil de similarité contextuelle choisi pour chaque motif.

contextuelle inférieure à 0,3 et une RMSD supérieure à 15Å. Le graphique en Figure 3.11b montre également que dans la classification sur la similarité contextuelle, avec un seuil de 0,4, les valeurs de RMSD sont réparties de manière bimodale, entre les paires de k-extensions intra-clusters et les paires de k-extensions inter-clusters.

Le motif G possède quant à lui un coefficient de corrélation de 0,63. Comme pour le motif trans WC/H, les paires d'occurrences sont séparées en deux groupes assez bien distincts (voir Figure 3.11c), avec un groupe comprenant des paires d'occurrences associées à une RMSD inférieure à 5Å et une similarité contextuelle supérieure à 0,5, et un autre groupe pour lequel la RMSD est supérieure à 5Å et la similarité contextuelle inférieure à 0,5. La distribution des valeurs de RMSD selon la classification sur la similarité contextuelle (voir Figure 3.11d) montre également une distribution bimodale entre les paires intra-clusters et les paires inter-clusters.

Le motif A-minor est celui pour lequel le coefficient de corrélation est le plus bas (0,33). Cependant, les deux métriques sont généralement corrélées, car pour la majorité des paires d'occurrences, plus la similarité contextuelle est élevée plus la RMSD est faible (voir Figure 3.11e). La distribution des valeurs de RMSD selon la classification sur la similarité contextuelle le montre également. La plupart des paires d'occurrences intra-cluster ont une RMSD inférieure à 2,5Å, tandis que la plupart des paires d'occurrences inter-clusters ont une RMSD supérieure à 3Å.

Cette première observation semble indiquer que le contexte 3D autour du

motif A-minor dépend moins de sa topologie que ce n'est le cas pour les deux autres motifs. On peut également noter que le motif trans WC/H, bien que possédant une bonne corrélation des deux métriques de similarité, nécessite de choisir un seuil de similarité contextuelle plus faible que les autres (0,4) pour obtenir la meilleure corrélation avec la similarité 3D. De plus, les valeurs de RMSD intra-clusters peuvent aller jusqu'à 7Å pour ce motif trans WC/H, alors que pour le motif G les valeurs de RMSD intra-clusters sont inférieures à 4Å (voir Figure 3.11b et Figure 3.11d).

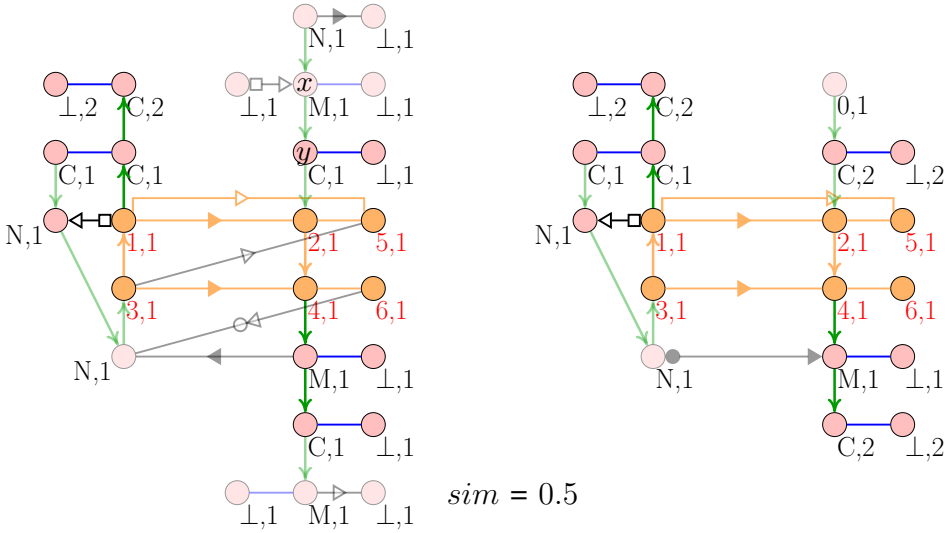
Dans la suite de cette étude, nous allons présenter et analyser quelques exemples de paires de k-extensions pour lesquels les valeurs de similarité contextuelle et de RMSD ne sont pas du tout corrélées.

Exemples de faux positifs et faux négatifs. Comme vu précédemment, certaines k-extensions possèdent des valeurs non corrélées de RMSD et de similarité contextuelle (voir Figure 3.11). C'est en particulier le cas pour le motif A-minor, dont nous allons donner quelques exemples.

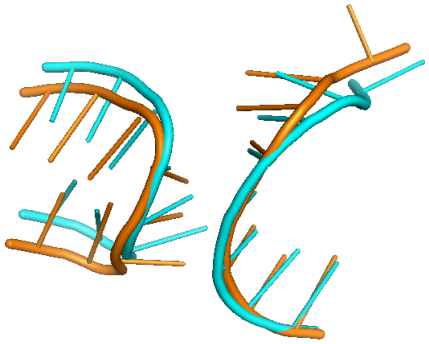
D'une part, certaines paires sont similaires selon le contexte structural en 3D et non similaires selon la topologie du contexte structural (valeur de similarité contextuelle faible et RMSD faible). Par exemple, sur 1547 paires de motifs A-minor possédant des contextes 3D similaires (les strutures 3D locales ont une RMSD inférieure à 2.5Å), 485 (soit 31%) possèdent des topologies de contexte non similaires (les k-extensions correspondantes possèdent une similarité contextuelle inférieure à 0.75). De la même façon, on trouve 82 paires de motifs G sur 797 (soit 10%) partageant des contextes 3D similaires et des topologies de contexte non similaires, et 96 paires de motifs trans WC/H sur 1419 (soit 6.8%) dans le même cas. Notons donc que cette proportion est non négligeable mais minoritaire parmi toutes les paires possédant une similarité de contexte 3D. Nous allons tenter d'expliquer pourquoi ces paires ne partagent pas une topologie de contexte similaire, en prenant quelques exemples d'occurrences de motif A-minor, pour lequel la proportion est la plus élevée.

Cette différence peut s'expliquer de diverses manières. Tout d'abord, il est possible que le modèle de graphes que nous utilisons ne parvienne pas à refléter certaines similarités de topologie de contexte. Par exemple, dans la k-extension de gauche de la Figure 3.12a, nous remarquons que les sommets x et y ne sont pas contractés, bien qu'étant tout deux une extrémité d'un arc canonique. Le sommet x est de type M, et possède donc un arc sortant (et un arc entrant) non canonique également, et l'autre sommet y est de type C et ne possède donc aucun arc sortant (ou entrant) non canonique. Nous avons choisi cette règle en considérant que ces deux sommets, bien qu'étant extrémité d'un arc de même type, ne possèdent pas le même contexte et ne peuvent donc pas être équivalents. Cependant, dans cet exemple comme dans d'autres, cette différence ne semble pas avoir d'incidence sur la position dans l'espace des nucléotides. Notons tout de même qu'en levant cette contrainte, la corrélation entre les deux métriques n'en devient pas meilleure.

Une deuxième explication pourrait être l'importance à accorder aux



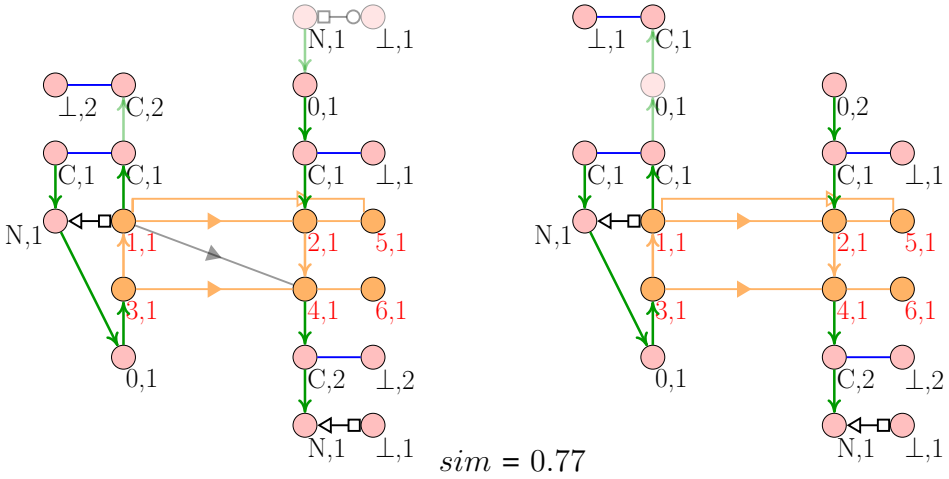
(a) Deux 4-extensions possédant une similarité contextuelle de 0,5. Les sommets et les arêtes en transparence sur les graphes n'appartiennent pas au sous-graphe commun maximum aux deux 4-extensions. Les sommets et les arcs de l'occurrence de motif sont en orange.



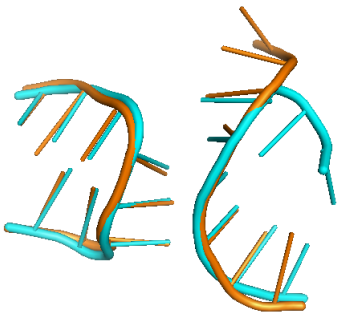
$RMSD$ = 2.41 Å

(b) Alignement 3D des sous-structures locales associées aux 4-extensions en (a).

Figure 3.12 – Exemple de deux occurrences de motif A-minor dont les contextes topologiques ne sont pas similaires (en (a)) mais qui partagent des sous-structures 3D locales similaires (en (b)). L'alignement 3D de ces sous-structures locales est présenté en (b).



(a) Deux 4-extensions possédant une similarité contextuelle de 0.77. Les sommets et les arêtes en transparence sur les graphes n'appartiennent pas au sous-graphe commun maximum aux deux 4-extensions. Les sommets et les arcs de l'occurrence de motif sont en orange.



(b) Alignement 3D des sous-structures locales associées aux 4-extensions en (a).

Figure 3.13 – Exemple de deux 4-extensions de motif A-minor (en (a)) qui sont similaires du point de vue topologique, mais possèdent des sous-structures 3D locales non similaires. L'alignement 3D de ces sous-structures locales est présenté en (b).

interactions non canoniques. Dans la Figure 3.12a, nous remarquons que la k-extension de gauche possède davantage d'arcs non canoniques que celle de droite et que la plupart d'entre eux ne semblent pas modifier outre mesure le contexte 3D (voir Figure 3.12b). Seuls ceux de la branche numéro 2 semblent avoir une légère incidence sur la structure tridimensionnelle. Ainsi, certaines interactions non canoniques semblent avoir plus d'importance que d'autres dans le contexte 3D.

Enfin, la dernière explication possible est que, dans certains cas au moins, la topologie du contexte structural ne détermine pas la structure tridimensionnelle locale. D'autres interactions que l'on ne considère pas dans la topologie (comme des interactions d'empilements qui apparaissent entre paires de nucléotides dans les hélices notamment), ou bien l'influence de la structure 3D à une échelle plus globale peuvent entrer en ligne de compte.

D'autre part, certaines paires de k-extensions sont similaires selon la topologie du contexte, mais non similaires selon le contexte en 3D. Ce cas arrive moins fréquemment. En effet, sur 1111 paires de motifs A-minor partageant des topologies de contexte similaires, 49 (soit 4.4%) ne possèdent pas de contextes 3D également similaires. De même, 25 paires de motifs G sur 737 (soit 3.3%) et 52 paires de motifs trans WC/H sur 1375 (soit 3.7%) partagent des topologies de contexte similaires et des contextes 3D non similaires. Les mêmes arguments que précédemment peuvent être avancés pour l'expliquer. Sur l'exemple présenté en Figure 3.13, les topologies sont très similaires, alors qu'une des branches des sous-structures locales est mal alignée (en haut à droite dans la Figure 3.13b). La topologie locale seule ne permet pas de capturer cette différence de contexte 3D.

3.7.2 Cohérence des deux métriques avec l'homologie

Dans cette sous-section, nous comparons les classifications obtenues sur les trois motifs avec les informations d'homologie entre occurrences. Notons que pour la classification selon la RMSD, nous choisissons un seuil sur les valeurs de RMSD pour chaque motif, proche de celui pris en exemple dans la Figure 3.10 (pour montrer des alignements 3D de structures que nous considérerons comme similaires).

Ici, nous dirons que deux topologies de contexte sont similaires si la valeur de similarité contextuelle entre les deux k-extensions correspondantes est supérieure au seuil que l'on a défini pour chaque motif. De même, nous dirons que deux contextes 3D sont similaires si la valeur de RMSD de l'alignement 3D des deux structures associées est inférieure au seuil défini pour chaque motif.

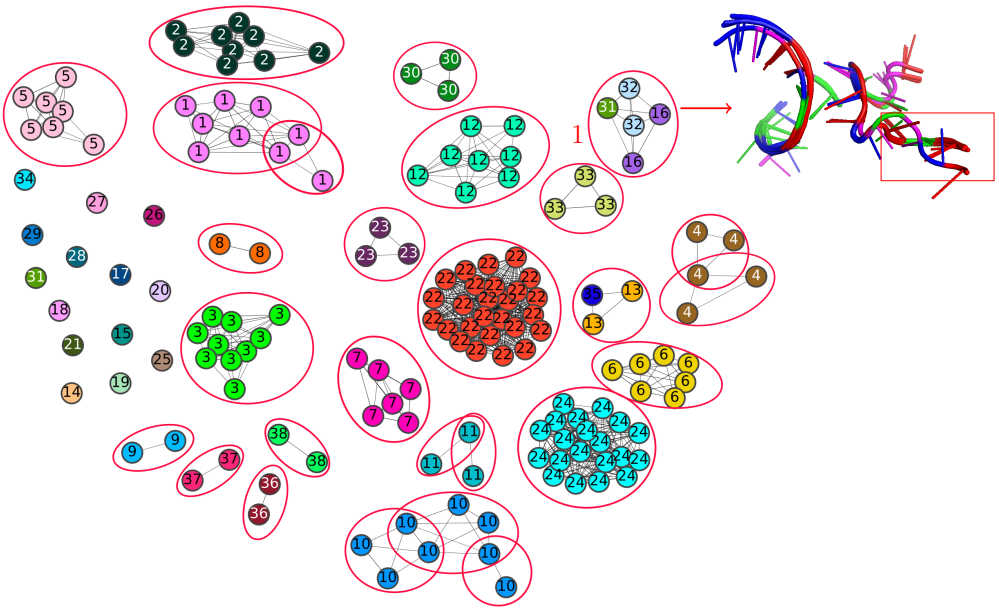
Pour le motif G tout d'abord, les graphes présentés en Figure 3.14 montrent que la majorité des classes regroupent uniquement des occurrences homologues, quelle que soit la métrique utilisée pour la classification (similarité contextuelle ou RMSD). Les deux classifications sont d'ailleurs très proches. De manière plus quantitative, les paires d'occurrences homologues

ou non homologues, ayant des topologies de contexte similaires ou non (resp. des contextes 3D similaires ou non), sont quantifiées dans la Table 3.7. On y remarque ainsi que peu de paires d'occurrences non homologues partagent des topologies de contexte similaires ou des contextes 3D similaires, et que peu d'occurrences homologues possèdent des topologies de contexte non similaires ou des contextes 3D non similaires. L'indice de Jaccard entre la classification selon la similarité contextuelle et les groupes d'homologie est d'ailleurs très élevé (0,96), comme l'indice de Jaccard entre la classification selon la RMSD et les groupes d'homologie (0,96 également).

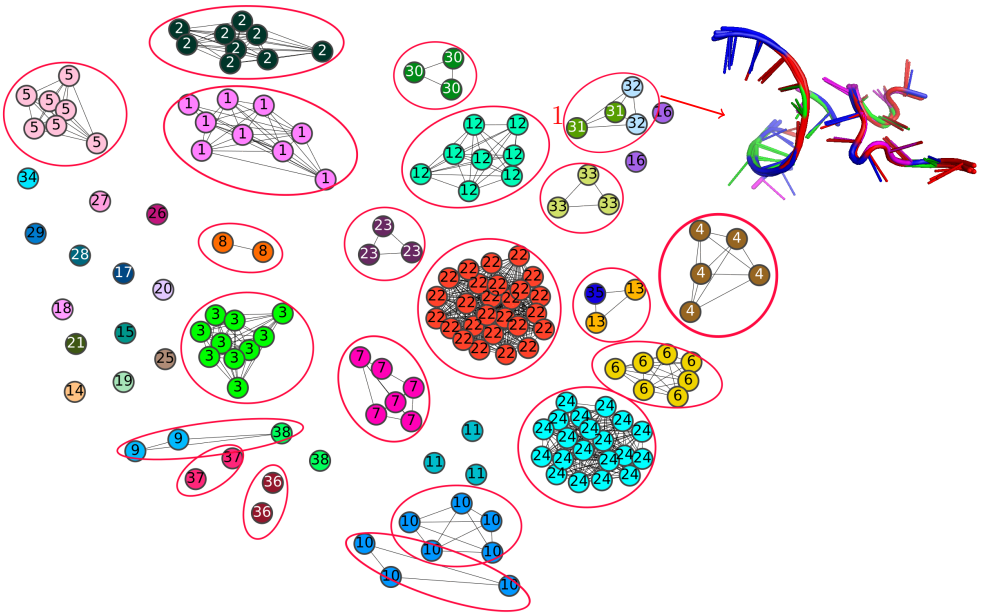
Les deux classes annotées 1 dans les graphes de la Figure 3.14 sont parmi les seules à regrouper des occurrences non homologues. La classe numéro 1 de la classification sur la similarité contextuelle (Figure 3.14a) regroupe des occurrences homologues issues de trois groupes d'homologie (numérotées 16, 31 et 32 dans la Figure), et la classe numéro 1 de la classification sur la RMSD (Figure 3.14b) regroupe uniquement les occurrences de deux classes d'homologie (31 et 32). Du point de vue biologique, les occurrences des groupes d'homologie 31 et 32 ont des particularités : ce sont les seules occurrences intermoléculaires du jeu de données, impliquant un ribosome et une autre molécule d'ARN pour les occurrences du groupe 32 et deux chaînes de ribosome pour les occurrences du groupe 31. De plus, les contextes 3D de ces occurrences sont très similaires, comme le montre l'alignement 3D présenté (Figure 3.14b). Les occurrences du groupe d'homologie numéro 16, quant à elles, sont intramoléculaires, et apparaissent dans des ribosomes également. On peut noter que les occurrences de ces trois groupes d'homologie, possédant des topologies de contexte donc similaires, possèdent également des contextes 3D relativement similaires, comme en témoigne l'alignement 3D qui diverge surtout pour la partie entourée en rouge (Figure 3.14a).

Pour ce motif G, certaines occurrences homologues n'ont ni une topologie de contexte structural très similaire, ni un contexte 3D très similaire. C'est le cas par exemple des groupes d'homologie 10 et 11 (en bas sur la Figure 3.14), qui contiennent des occurrences issues d'ARNr 5S.

Pour le motif trans WC/H (Figure 3.15), avec les deux métriques, 4 classes de plus de 2 occurrences émergent, dont une comprenant la majorité des occurrences. Cette classification correspond presque exactement aux familles d'ARN dans lesquelles sont trouvées les occurrences : la classe majoritaire est constituée des occurrences issues d'ARN de transfert, et les autres classes sont constituées d'occurrences issues d'ARN ribosomiques (petite ou grande sous-unité). Comme nous avons considéré comme homologues des ARN de transfert fixant le même acide aminé, la classe majoritaire contient de nombreuses occurrences non homologues dans les deux classifications (voir Table 3.7), puisque la plupart des occurrences de motif issues d'ARN de transfert sont dans une même classe dans les deux classifications. Nous pouvons également noter que certaines occurrences trouvées dans des ARNt fixant le même acide aminé, que l'on considère donc ici comme homologues, possèdent des différences de contexte 3D (91 paires dans la Table 3.7). Cependant, comme le montre l'alignement 3D présenté en Figure 3.15, toutes

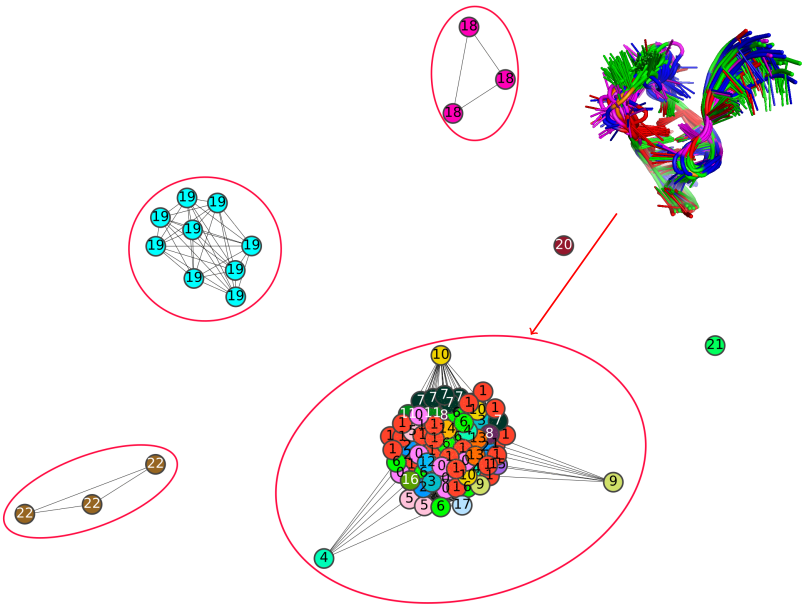


(a) seuil de similarité contextuelle = 0.6

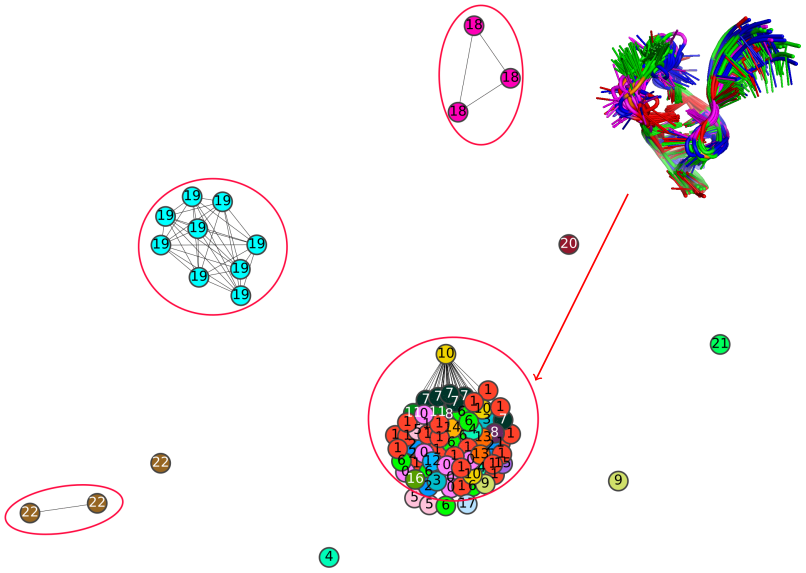


(b) seuil de RMSD = 3.5 Å

Figure 3.14 – Graphe de similarité de seuil $s = 0,6$ en (a) et graphe de RMSD de seuil $r = 3,5$Å en (b) du motif G. Dans ces deux graphes, chaque sommet correspond à une k-extension et il y a une arête entre deux sommets si la similarité contextuelle pour le graphe en (a) (resp. la RMSD pour le graphe en (b)) est supérieure au seuil (resp. inférieure). Les classes sont indiquées par des ellipses rouges. Les occurrences homologues de motif G sont de la même couleur, et annotées par le même numéro. L'alignement 3D d'une des deux classes regroupant des occurrences non homologues est présenté pour chaque classification, avec une couleur par type de nucléotides (A : rouge, C : bleu, G : vert, U : magenta). La partie encadrée en rouge dans l'alignement 3D en (a) est moins bien alignée entre les sous-structures locales que les autres parties.

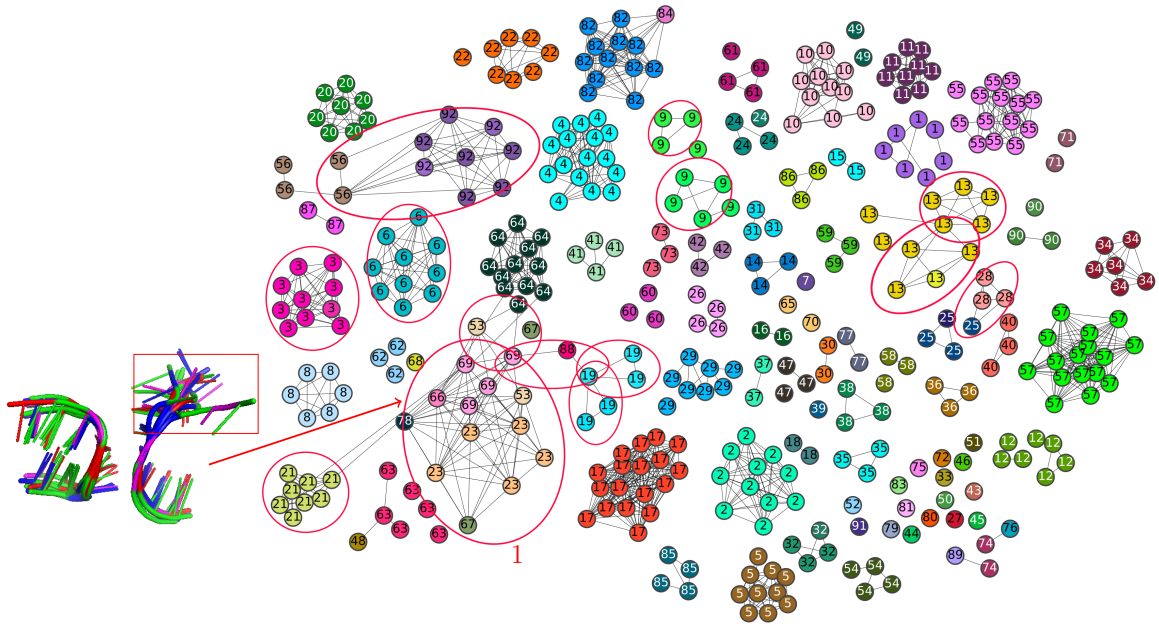


(a) seuil de similarité contextuelle = 0.4

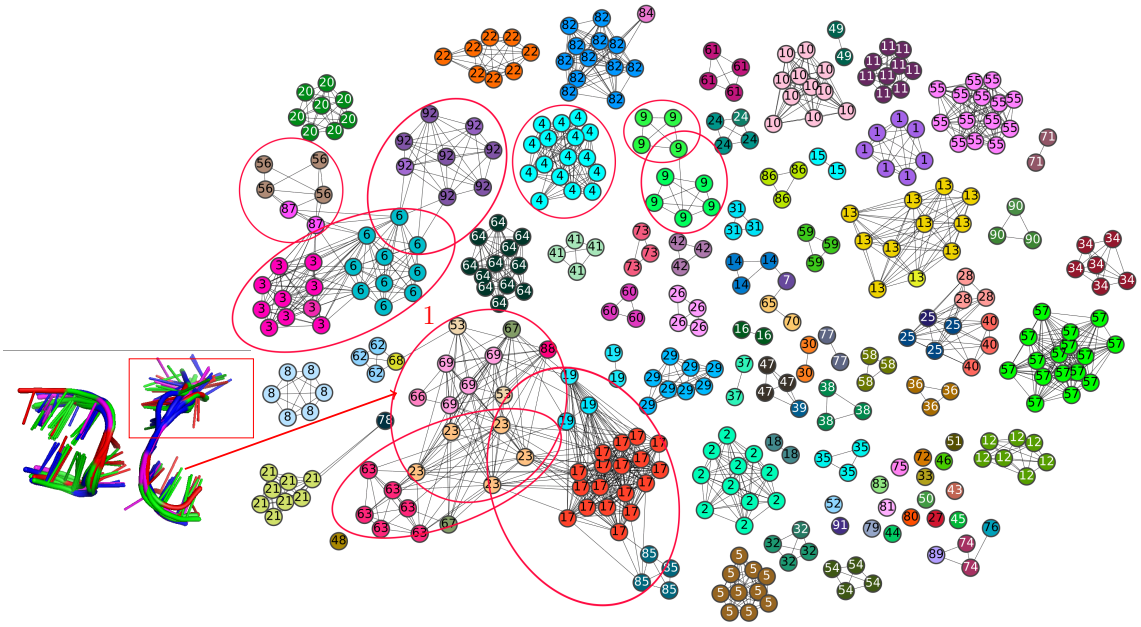


(b) seuil de RMSD = 5 Å

Figure 3.15 – Graphe de similarité de seuil $s = 0,4$ en (a) et graphe de RMSD de seuil $r = 5$Åen (b) du motif trans WC/H. Dans ces deux graphes, chaque sommet correspond à une k-extension et il y a une arête entre deux sommets si la similarité contextuelle pour le graphe en (a) (resp. la RMSD pour le graphe en (b)) est supérieure au seuil (resp. inférieure). Les classes sont indiquées par des ellipses rouges. Les occurrences homologues de motif trans WC/H sont de la même couleur, et annotées par le même numéro. L'alignement 3D de la classe composée d'occurrences non homologues est présenté pour chaque classification, avec une couleur par type de nucléotides (A : rouge, C : bleu, G : vert, U : magenta).



(a) seuil de similarité contextuelle = 0.75



(b) seuil de RMSD = 2.5 Å

Figure 3.16 – Graphe de similarité de seuil $s = 0,75$ en (a) et graphe de RMSD de seuil $r = 2.5$Å en (b) du motif A-minor. Dans ces deux graphes, chaque sommet correspond à une k-extension et il y a une arête entre deux sommets si la similarité contextuelle pour le graphe en (a) (resp. la RMSD pour le graphe en (b)) est supérieure au seuil (resp. inférieure). Les classes ne correspondant pas à une composante connexe du graphe sont indiquées par des ellipses rouges. Les occurrences homologues de motif A-minor sont de la même couleur, et annotées par le même numéro. L'alignement 3D d'une classe composée d'occurrences non homologues est présenté pour chaque classification, avec une couleur par type de nucléotides (A : rouge, C : bleu, G : vert, U : magenta). La partie encadrée en rouge dans les alignements 3D est moins bien alignée entre les sous-structures locales que les autres parties.

		Homologues	Non homologues
Motif G	Topologies similaires	752	10
	Topologies non similaires	24	
	Contextes 3D similaires	749	8
	Contextes 3D non similaires	27	
Motif trans WC/H	Topologies similaires	357	2170
	Topologies non similaires	0	
	Contextes 3D similaires	266	1370
	Contextes 3D non similaires	91	
Motif A-minor	Topologies similaires	1129	102
	Topologies non similaires	226	
	Contextes 3D similaires	1333	489
	Contextes 3D non similaires	22	

Table 3.7 – Nombre de paires d'occurrences homologues ou non homologues et partageant des contextes topologiques similaires ou non similaires ou des contextes 3D similaires ou non similaires, pour chacun des trois motifs. Le cas de paires d'occurrences non homologues, ne possédant pas de topologies de contexte similaires ou de contextes 3D similaires, n'est pas quantifié, mais correspond à toutes les autres paires des jeux de données.

les occurrences issues d'ARN de transfert ont des contextes 3D relativement similaires. Les autres classes ne regroupent que des occurrences homologues dans les deux classifications.

Pour le motif A-minor (Figure 3.16), le même constat que pour le motif G peut être fait : la majorité des classes regroupent uniquement des occurrences homologues dans les deux classifications. Cependant, davantage de différences existent entre les deux classifications que pour les deux motifs précédents. Dans la Table 3.7, nous remarquons qu'un nombre plus élevé de paires d'occurrences homologues possèdent des topologies de contexte non similaires (226), par rapport aux paires d'occurrences homologues possédant des contextes 3D non similaires (22). A l'inverse, le nombre de paires d'occurrences non homologues possédant des contextes 3D similaires (489) est plus élevé que le nombre de paires d'occurrences non homologues possédant des topologies de contextes similaires (102). Ainsi, la classification utilisant la RMSD semble mieux corrélée avec l'homologie que la classification utilisant la similarité contextuelle. En effet, trouver des contextes non similaires pour des occurrences homologues peut être le reflet des limites de notre modèle, comme évoqué dans la section 3.7.1, tandis que trouver des contextes similaires pour des occurrences non homologues nous apporte une nouvelle information.

Une classe de non homologues cependant est retrouvée partiellement dans les deux classifications (numéro 1 dans la Figure 3.16). Les occurrences qu'elle contient ont donc à la fois une topologie de contexte similaire et des

sous-structures 3D locales similaires. Cette classe regroupe des occurrences issues d'ARNr 16S et d'ARNr 23S. Leur topologie a comme particularité de contenir peu d'interactions non canoniques et l'alignement 3D diverge surtout pour une branche sur les 4 (voir Figure 3.16, partie entourée en rouge).

Comme attendu, les occurrences homologues possèdent d'une manière générale des topologies de contexte et des structures 3D locales similaires, puisque les structures de molécules homologues sont souvent conservées au cours de l'évolution. Cependant, nous trouvons quelques exemples d'occurrences non homologues qui possèdent des structures 3D similaires, ou bien des topologies de contexte similaires, ou bien les deux, plus rarement. Ces exemples pourraient indiquer que les structures des molécules qui contiennent ces occurrences ont convergé au cours de l'évolution pour devenir ainsi similaires. Ces cas seront davantage étudiés dans le chapitre 4.

Nous remarquons également dans cette étude que, pour les motifs G et trans WC/H, les deux classifications sont très proches de l'homologie (ou des familles d'ARN pour le motif trans WC/H), à quelques exceptions près. Pour le motif A-minor, davantage de classes formées d'occurrences non homologues émergent des deux classifications. De plus, nous pouvons conclure de cette étude que, pour les trois motifs étudiés, la classification selon la similarité contextuelle donne des résultats très similaires à ceux obtenus avec la classification selon la RMSD. Cela démontre que, dans la plupart des cas, la topologie du contexte structural d'un motif d'ARN est suffisante pour expliquer les similarités de contexte 3D et d'homologie.

3.8 Conclusion

Nous avons donc étudié les contextes structuraux de trois motifs à longue distance difficiles à prédire, selon leur topologie et selon leur forme 3D.

Nous avons ainsi observé que notre modèle de contraction de certains sommets et arcs des k-extensions permet avant tout de réduire la taille des graphes considérés et ainsi le temps d'exécution des algorithmes, tout en conservant la même cohérence avec le contexte 3D. Les tailles d'extension, quant à elles, semblent avoir une influence sur la distance entre occurrences homologues et non homologues. Il convient donc de choisir une taille d'extension qui ne soit pas trop élevée pour pouvoir observer des similarités entre occurrences non homologues.

D'une manière générale, pour les trois motifs, notre métrique sur la topologie des contextes (la similarité contextuelle des k-extensions) est cohérente avec la métrique de similarité 3D des contextes (la RMSD des sous-structures 3D locales). Lorsque la RMSD est faible et indique donc que les contextes sont similaires en 3D, la similarité contextuelle est élevée et indique donc que les topologies de contexte sont également similaires.

Cependant, il existe des exceptions, en particulier pour le motif A-minor, qui peuvent être expliquées de plusieurs manières. Notre modèle de graphes

peut ne pas toujours refléter la topologie du contexte de manière adéquate. Certains de nos choix de représentation peuvent induire des incohérences par rapport au contexte 3D. De plus, ces exceptions peuvent montrer que la topologie du contexte ne détermine pas toujours le contexte 3D, et que d'autres éléments doivent être pris en compte, comme d'autres interactions que les seules interactions canoniques et non canoniques, et l'influence de la structure globale.

Une autre caractéristique de la corrélation entre les deux métriques est l'influence de l'homologie. La majorité des contextes structuraux trouvés similaires en 3D et en topologie sont des contextes d'occurrences homologues. Pour le motif trans WC/H et le motif G, on ne trouve quasiment pas de classes de contextes similaires qui contiennent des occurrences non homologues ou de familles d'ARN différentes. On en trouve davantage avec le motif A-minor, et l'une de ces classes possède à la fois des contextes 3D similaires et des topologies de contexte similaires.

Cela semble indiquer que, dans la majorité des cas, la topologie de contexte seule détermine suffisamment les similarités de contexte 3D. Etant donné que la topologie seule est bien plus facile à obtenir que les informations complètes du contexte 3D, elle pourrait être utilisée comme base de classification de nouvelles occurrences de motifs, ainsi que dans un but prédictif. C'est ce que nous verrons dans le dernier chapitre de ce manuscrit.

Dans le chapitre 4, nous allons d'abord effectuer une étude plus approfondie des classifications selon la RMSD du motif A-minor, pour déterminer quelles nouvelles informations biologiques peut apporter le contexte 3D de ce motif.

Chapitre 4

Analyse du motif A-minor selon la similarité 3D

4.1 Introduction

Dans le chapitre 3, nous avons étudié la relation entre le contexte topologique et le contexte 3D des occurrences de trois différents motifs d'ARN, et nous avons montré que les similarités de contexte topologique permettent d'expliquer généralement les similarités 3D et les relations d'homologie entre occurrences de motif.

Dans ce chapitre, nous allons nous intéresser en particulier à l'un de ces trois motifs : le motif A-minor. La comparaison des contextes structuraux des occurrences de ce motif dans le chapitre 3 laissait émerger des classes dans lesquelles se trouvaient des occurrences non homologues, qui possédaient donc des topologies de contexte similaires et/ou des contextes 3D similaires. Ces cas sont intéressants à étudier étant donné qu'ils peuvent témoigner d'un phénomène de convergence évolutive. De plus, des travaux [66, 63] ont montré l'importance du motif A-minor d'un point de vue biologique, mais peu d'études ont cherché à classifier ce motif selon son contexte 3D [101]. Ces raisons nous ont incités à approfondir l'étude de ce motif. Nous allons ainsi présenter dans ce chapitre plusieurs classifications exhaustives des occurrences du motif A-minor selon différentes définitions de contexte 3D.

Nous allons tout d'abord, dans la section 4.2, détailler la classification présentée dans le chapitre 3. Puis, dans la section 4.3, nous définirons des sous-contextes 3D, qui donneront lieu à d'autres classifications. Nous comparerons chacune de ces classifications à des caractéristiques structurales déjà connues sur le motif A-minor, comme le type d'élément de structure secondaire impliqué. Ensuite, dans la section 4.4, nous étudierons brièvement les sous-contextes d'un point de vue topologique, dans le but de déterminer si l'observation du chapitre 3 à propos de l'influence de la topologie sur le contexte 3D pour le contexte structural dans sa totalité, peut s'appliquer également aux sous-contextes. Enfin, dans la section 4.5, nous comparerons l'une de nos classifications à une autre classification des occurrences du motif

A-minor, établie dans [101].

4.2 Classification à 4 branches

Dans cette partie, nous allons étudier plus en détail la classification des occurrences de motif A-minor selon la similarité 3D, définie dans le chapitre 3 (section 3.5, Figure 3.16). Nous l'appellerons ici *classification à 4 branches*. Rappelons que cette classification est calculée dans un graphe dans lequel les sommets correspondent aux occurrences de motif A-minor et leur contexte, et une arête est présente entre deux sommets si la RMSD entre les deux sous-structures 3D locales associées est inférieure à un seuil de 2.5Å (chapitre 3, section 3.7.2). L'algorithme de classification utilisé permet d'obtenir une couverture de ce graphe (chapitre 3, section 3.5).

Nous avons montré dans le chapitre 3 que cette classification était cohérente avec l'homologie. En effet, les paires d'occurrences homologues sont en grande majorité regroupées dans une même classe. C'est un résultat prévisible étant donné que les structures 3D sont généralement conservées entre molécules homologues. Dans cette partie, nous allons alors nous intéresser aux classes composées d'occurrences non homologues, qui partagent des structure 3D locales similaires, ainsi qu'aux quelques exemples d'occurrences homologues qui n'appartiennent pas à la même classe.

4.2.1 Cas des classes d'occurrences non homologues

La classification à 4 branches est constituée de 11 classes composées d'occurrences non homologues (Figure 4.1), sur un total de 79 classes. Dans ces classes se trouvent 106 occurrences sur un total de 391 occurrences de motif A-minor dans notre jeu de données, ce qui équivaut à une moyenne de 9,63 occurrences par classe de non homologues.

La Figure 4.1 présente les classes de la classification à 4 branches qui contiennent des occurrences non homologues, ainsi que leurs alignements 3D. La Table 4.1 décrit des caractéristiques de ces classes, comme les familles d'ARN et les organismes dans lesquels sont trouvés les motifs, ou encore la densité du sous-graphe du graphe de RMSD induit par les occurrences de motif de la classe. On remarque tout d'abord que la plupart de ces classes regroupent des occurrences de motif A-minor appartenant à des familles d'ARN diverses et d'organimes divers. On retrouve souvent des ARN ribosomiques, qui sont les plus représentés dans notre jeu de données (87%), mais certaines classes regroupent à la fois des occurrences de motif d'ARN ribosomiques et des occcurrences de motif d'autres molécules, comme les introns, les ribozymes ou les riboswitchs. La majorité de ces classes possèdent également des occurrences de motif provenant d'au moins deux règnes du Vivant (Bactéries, Archées, ou Eucaryotes).

De plus, il peut être noté que les occurrences de motif d'une même classe possèdent le même type de boucle, pour presque toutes les classes (sauf les

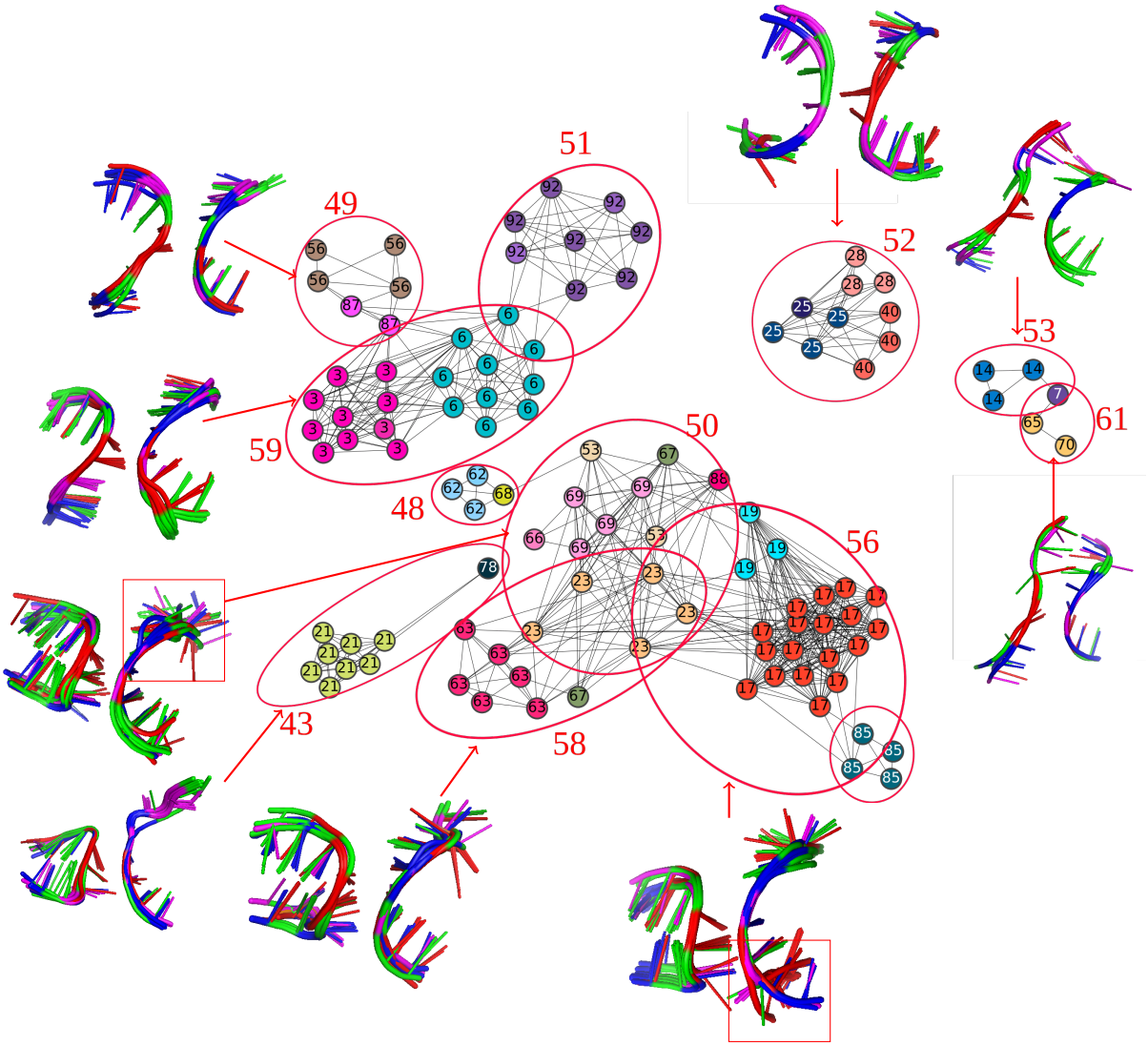
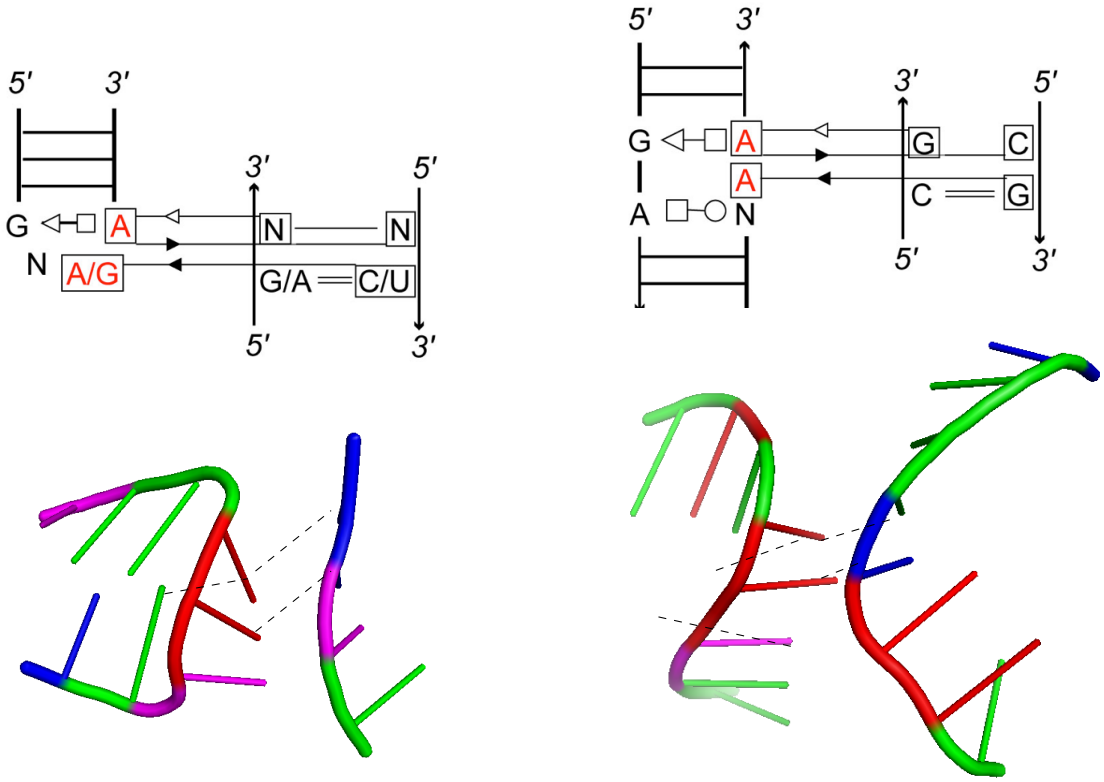


Figure 4.1 – Graphe de RMSD du motif A-minor de la classification à 4 branches (seuil de RMSD = 2,5Å), induit par les occurrences appartenant à des classes contenant des non homologues. La classification est indiquée en rouge. Les alignements 3D de la plupart des classes sont présentés. La partie encadrée en rouge dans certains alignements correspond à la partie la moins bien alignée entre les sous-structures locales.

Numéro de classe	Nombre d'occurrences	Nombre de groupes d'homologie	Densité de la classe	Type de boucle	Famille(s) d'ARN	Règnes des organismes
43	9	2	0,83	GNRA	ARNr 23S, ARNr 25S, Ribonucléase P'	Bactéries, Archées, Eucaryotes et non spécifié
48	5	3	0,70	GNRA, tetraloop	ARNr 16S, ARNr 18S, ARNr 25S	Bactérie, Eucaryotes
49	6	2	0,73	Boucle interne	ARNr 16S, Intron de groupe IIC	Bactéries, Eucaryotes
50	14	6	0,71	GNRA, tetraloop	ARNr 23S, ARNr 25S, ARNr 28S, ARNr 16S, ARNr 18S, Intron de groupe IIC	Bactéries, Eucaryotes
51	10	2	0,76	Multi boucle, A-rich loop	ARNr 16S, ARNr 23S	Bactéries
52	10	3	0,80	Boucle interne	ARNr 23S	Bactéries, Archées
53	4	2	0,67	Boucle interne, Multi boucle	ARNr 23S	Bactéries, Eucaryotes
56	26	4	0,70	GNRA	ARNr 23S, ARNr 28S, ARNr 25S, c-di-GMP Riboswitch	Bactéries, Archées, Eucaryotes, non spécifié
58	12	3	0,67	GNRA, tetraloop	ARNr 16S, ARNr 18S, ARNr 23S	Bactéries, Eucaryotes
59	20	2	0,61	A-rich Loop, boucle interne	ARNr 23S, ARNr 25S, ARNr 28S	Bactéries, Archées, Eucaryotes
61	3	3	0,67	Multi boucle	ARNr 23S, ARNr 16S	Bactéries

Table 4.1 – Description des classes composées d'occurrences non homologues, de la classification à 4 branches. La densité de la classe indique la densité du sous-graphe du graphe de RMSD (Figure 4.1) induit par les occurrences de la classe. Le type de boucle indique le type d'élément de structure secondaire du motif A-minor.

classes numéro 51 et 53) (Table 4.1). Comme expliqué dans la présentation du motif A-minor de type I/II faite dans le chapitre 3 (section 3.2.1), les interactions non canoniques du motif A-minor relient deux éléments de structure secondaire, une hélice et une boucle (voir exemple dans la Figure 4.2). Différents types de boucle sont alors observés : des boucles terminales, des boucles internes, des renflements et plus rarement des multi-boucles (voir des exemples de ces types de boucles dans le chapitre 1, section 1.5.3). Des études [66] ont alors mis en évidence la présence de boucles particulières dans les motifs A-minor, comme les *tetraloops*, qui sont des boucles terminales formées de 4 nucléotides. Parmi elles, on retrouve les boucles GNRA (Figure 4.2a), qui ont été évoquées comme motif structural local d'ARN, dans le chapitre 1, section 1.5.3. On trouve également des boucles internes particulières, que l'on nomme *A-rich loop* (Figure 4.2b). Ces deux types (GNRA et *A-rich loop*) ont également une topologie particulière, comme on peut le voir dans la Figure 4.2. Les structures 3D associées à deux types de boucle différents (par exemple, une boucle terminale et une boucle interne, ou une boucle interne et une multi-boucle, etc.) sont très différentes. Il est donc cohérent de retrouver rarement des occurrences de motif A-minor, possédant



(a) Motif A-minor avec GNRA (b) Motif A-minor avec boucle de type A-rich

Figure 4.2 – Motifs A-minor de type I/II, impliquant une boucle GNRA (a) ou une boucle interne de type A-rich (b). Les figures du haut sont issues de [66]. Les traits horizontaux correspondent à des interactions canoniques, et les interactions non canoniques sont représentées selon la nomenclature de Leontis-Westhof [62]. Les flèches et les annotations 5′ et 3′ indiquent le sens d'orientation de la séquence primaire. En bas, sont présentés des exemples de structures 3D d'occurrences de motif A-minor impliquant une GNRA ou une boucle A-rich. Les nucléotides sont colorés par type (A : rouge, C : bleu, G : vert, U : magenta), et les traits pointillés correspondent aux interactions non canoniques.

des types de boucle différents, dans une même classe.

D'un autre côté, il existe plusieurs classes dont les occurrences de motif sont toutes composées d'un même type de boucle. Il existe donc des différences de contexte 3D entre occurrences de motif composées du même type de boucle.

Ainsi, ces classes composées d'occurrences de motifs non homologues possèdent des contextes 3D similaires, comme en témoignent les alignements 3D présentés en Figure 4.1, bien que les occurrences soient issues de molécules parfois très différentes. Cette similarité n'est donc pas due à l'homologie, mais à un phénomène de convergence évolutive. Nous pouvons alors supposer que ces occurrences de motif A-minor sont impliquées dans des mécanismes cellulaires importants qui nécessitent que les structures soient restées similaires au cours de l'évolution.

4.2.2 Cas des occurrences homologues réparties en plusieurs classes

On trouve également deux classes différentes dont les occurrences sont homologues (classes 44 et 47 dans la Figure 4.3). Les contextes 3D des occurrences de ces deux classes ne sont donc pas similaires, malgré leur homologie, comme le montre l'alignement 3D. Cependant, il faut noter que les occurrences de la classe 44 sont issues d'organismes bactériens tandis que les occurrences de la classe 47 sont issues d'organismes eucaryotes ou d'archées. Cela peut donc expliquer le manque de similarité des contextes 3D. De plus, on remarque dans l'alignement 3D de ces deux classes (Figure 4.3) que la divergence se concentre en fait sur une des 4 parties des sous-structures 3D, tandis que les 3 autres sont très bien alignées.

Ce constat peut d'ailleurs se faire également sur les classes de non homologues. Les alignements 3D contiennent souvent une partie moins bien alignée que les autres (voir la partie encadrée en rouge dans les alignements de la Figure 4.1). Cela nous amène à penser que certaines parties du contexte 3D sont davantage conservées que d'autres, en particulier lorsqu'on considère des motifs non homologues. Pour étayer cette hypothèse, nous allons étudier d'autres classifications de motifs A-minor, à partir de définitions de contexte 3D différentes qui ne prendront en compte que certaines parties de ce contexte.

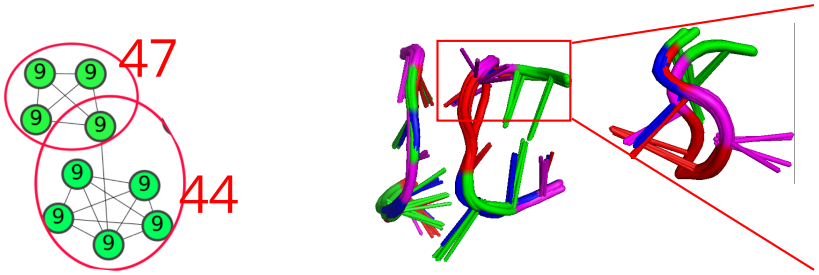


Figure 4.3 – Occurrences homologues n'appartenant pas à la même classe dans la classification à 4 branches. A gauche est présenté le sous-graphe du graphe de RMSD induit par ces occurrences, avec les classes indiquées en rouge. A droite est présenté l'alignement 3D du contexte de ces occurrences. La partie la moins bien alignée est entourée en rouge, et placée dans une autre orientation pour rendre visible les différences.

4.3 Autres classifications selon le contexte 3D

Dans cette partie, nous allons décrire des classifications qui ne prennent pas en compte la totalité du contexte 3D mais seulement une partie. Après avoir défini différents sous-contextes 3D, nous allons d'abord nous intéresser à deux classifications qui vont considérer 3 branches sur 4 du contexte 3D, que nous appellerons alors *classifications à 3 branches*, puis nous évoquerons deux

classifications considérant 2 branches sur 4 seulement, que nous appellerons alors *classifications à 2 branches*.

4.3.1 Définitions de sous-contextes 3D à 3 branches ou 2 branches

Comme définie dans le chapitre 3 (section 3.3), la sous-structure 3D locale d'un motif A-minor, que l'on appelle contexte 3D, est induite par les sommets de la k-extension correspondante. Les sommets de cette k-extension sont couverts par 6 ensembles de sommets que l'on appelle *branches* (voir chapitre 2, section 2.2.2). Nous définissons ici différents sous-contextes 3D d'une occurrence de motif A-minor à partir de ces branches :

- Le sous-contexte 3D induit par les sommets appartenant aux branches 1 et 3 et n'étant pas de type $\rho_O = \perp$ (chapitre 2, section 2.2.2). Ce sous-contexte 3D correspond au contexte 3D réduit aux branches de la boucle du motif A-minor : on l'appellera ainsi contexte de *boucle-1,3*.

- Le sous-contexte 3D induit par les sommets appartenant aux branches 2 et 4 et n'étant pas de type $\rho_O = \perp$ (chapitre 2, section 2.2.2). Ce sous-contexte 3D correspond au contexte 3D réduit aux branches de l'hélice du motif A-minr : on l'appellera ainsi contexte d'*hélice-2,4*.

- Le sous-contexte 3D induit par les sommets appartenant aux branches 1, 2 et 3, et n'étant pas de type $\rho_O = \perp$ (chapitre 2, section 2.2.2). Dans ce sous-contexte, on considère donc les deux branches de la boucle, ainsi que la branche de l'hélice numérotée 2 : on l'appellera ainsi contexte de *boucle-1,2,3*.

- Le sous-contexte 3D induit par les sommets appartenant aux branches 1, 3 et 4, et n'étant pas de type $\rho_O = \perp$ (chapitre 2, section 2.2.2). Dans ce sous-contexte, on considère donc les deux branches de la boucle, ainsi que la branche de l'hélice numérotée 4 : on l'appellera ainsi contexte de *boucle-1,3,4*.

Nous annoterons par *b-1,2,3* ou *b-1,3,4* les classes des classifications boucle-1,2,3 et boucle-1,3,4 respectivement, pour les différencier des classes de la classification à 4 branches (voir exemple en Figure 4.4).

En comparant les sous-contextes 3D d'un type donné pour toutes les k-extensions, on obtient différents graphes de RMSD, et différentes classifications. Nous noterons chacune de ces classifications de la même façon que les sous-contextes 3D qui permettent de les obtenir (classification *boucle-1,3*, classification *hélice-2,4*, classification *boucle-1,2,3*, classification *boucle-1,3,4*).

Nous allons à présent détailler chacune de ces classifications.

4.3.2 Classifications à 3 branches

Dans la classification à 4 branches décrite dans la section 4.2, nous avons remarqué que les types de boucle du motif A-minor avaient une influence sur les classes obtenues : les classes regroupent des occurrences de motif impliquées dans le même type de boucle de structure secondaire. C'est la raison pour laquelle nous nous intéressons en particulier aux sous-contextes 3D qui contiennent les deux branches correspondant à la boucle (numéro 1 et 3).

Les classes d'occurrences non homologues appartenant aux deux classifications à 3 branches, *boucle-1,2,3* et *boucle-1,3,4*, sont présentées dans la Figure 4.4. Nous avons choisi un seuil de RMSD de 2Å pour ces deux classifications, inférieure au seuil de RMSD de 2.5Å choisi pour la classification à 4 branches. Il n'existe pas de mesure objective permettant de choisir ces seuils sur la RMSD. Pour la classification à 4 branches, nous avons choisi un seuil supérieur au seuil défini pour l'homologie (2Å). Comme le nombre de nucléotides dans les sous-structures locales considérées est plus faible dans les classifications à 3 branches que dans la classification à 4 branches (en général 15 nucléotides), nous prenons un seuil plus faible pour les classifications à 3 branches. De plus, ce seuil nous permet d'avoir des alignements 3D de bonne qualité.

La classification boucle-1,2,3 contient 3 classes d'occurrences non homologues, regroupant 89 occurrences, sur un total de 86 classes. La classification boucle-1,3,4 contient de même 3 classes d'occurrences non homologues, regroupant 54 occurrences, sur un total de 85 classes. Le nombre d'occurrences dans ces classes composées d'occurrences non homologues est donc plus faible que dans la classification à 4 branches, et le nombre total de classes est équivalent, voire même légèrement plus élevé, car le seuil sur la RMSD considéré est plus faible que celui choisi pour la classification à 4 branches (2Å). Cependant, le nombre moyen d'occurrences dans une classe de non homologues est plus élevé pour ces deux classifications (respectivement 29,6 pour la classification boucle-1,2,3 et 18 pour la classification boucle-1,3,4).

Nous retrouvons dans ces deux classifications un certain nombre d'occurrences de motif déjà regroupées dans la classification à 4 branches. C'est le cas, par exemple, des occurrences appartenant aux groupes d'homologie 3 et 6 de la classe 50 b-1,2,3 (en haut dans la Figure 4.4), ou encore des occurrences des groupes d'homologie 23, 69 et 88 dans la classe 49 b-1,2,3 (en bas dans la Figure 4.4). Cependant, d'autres liens apparaissent entre occurrences non homologues. Comme on peut le voir dans la Table 4.2, les classes regroupent des occurrences de motif trouvées dans des familles d'ARN diverses, en particulier la classe 49 b-1,2,3 de la classification boucle-1,2,3 et la classe 51 b-1,3,4 de la classification boucle-1,3,4, qui sont composées de motifs issus d'ARN ribosomiques et d'autres familles d'ARN (ribonucléase et intron, ou riboswitch et intron). De plus, les occurrences au sein d'une classe sont toujours impliquées dans un même type de boucle (boucle interne ou boucle terminale), comme c'était le cas dans la classification à 4 branches.

En outre, on peut noter que certaines classes de la classification

boucle-1,2,3 se retrouvent partiellement dans des classes de la classification boucle-1,3,4, et inversement. Sur la Figure 4.4, les intersections de classes le mettent en évidence. Cela semble indiquer plusieurs niveaux de similarités entre les contextes 3D de motifs A-minor. Certaines occurrences possèdent des contextes 3D similaires dans leur totalité, tandis que d'autres occurrences possèdent des similarités de sous-contextes 3D seulement. Cela semble montrer que les occurrences du motif A-minor ne peuvent être classifiées de manière absolue selon leur contexte 3D, mais que plusieurs classifications doivent être prises en compte pour refléter les similarités de contexte.

Numéro de classe	Nombre d'occurrences	Nombre de groupes d'homologie	Densité de la classe	Type de boucle	Famille(s) d'ARN	Règnes des organismes
24 (b2)	14	3	0,60	Boucle interne, A-rich	ARNr 23S, ARNr 25S, ARNr 28S	Bactéries, Archées, Eucaryotes
50 (b2)	24	3	0,84	Boucle interne, A-rich	ARNr 23S, ARNr 25S, ARNr 28S	Bactéries, Archées, Eucaryote
49 (b2)	51	9	0,68	GNRA	ARNr 23S, ARNr 25S, ARNr 28S, ARNr 16S, ARNr 18S, Riboswitch c-di-GMP, Intron de groupe IIC	Bactéries, Archées, Eucaryotes

(a) Classification boucle-1,2,3

Numéro de classe	Nombre d'occurrences	Nombre de groupes d'homologie	Densité de la classe	Type de boucle	Famille(s) d'ARN	Règnes des organismes
30 (b4)	13	4	0,55	Boucle interne	ARNr 23S, ARNr 16S, Intron de groupe II	Bactéries, Archées
51 (b4)	37	9	0,73	GNRA	ARNr 16S, ARNr 18S, ARNr 23S, ARNr 25S, ARNr 28S, Ribonucléase P, Intron de groupe IIC	Bactéries, Archées, Eucaryotes, non spécifié
61 (b4)	4	3	0,67	GNRA	ARNr 16S, ARNr 25S, Ribonucléase P	Bactéries, Eucaryotes

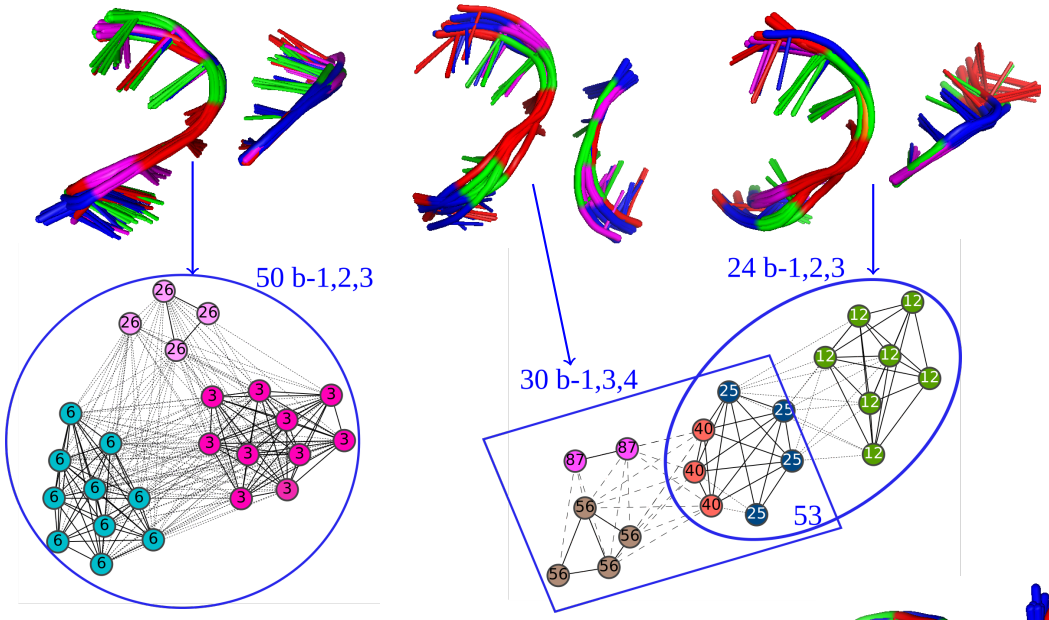
(b) Classification boucle-1,3,4

Table 4.2 – Description des classes composées d'occurrences non homologues, pour les classifications à 3 branches ((a) boucle-1,2,3, (b) boucle-1,3,4).

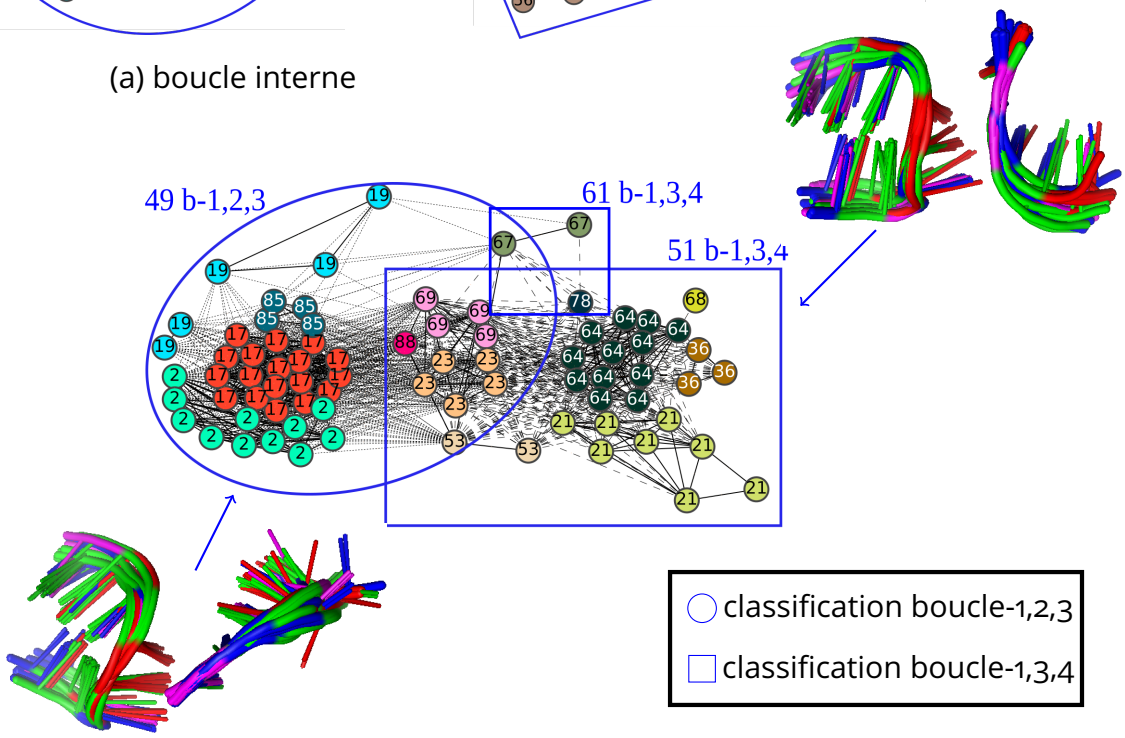
4.3.3 Classifications à 2 branches

Pour finir, nous allons étudier les classifications *boucle-1,3* et *hélice-2,4* à 2 branches (voir définitions en section 4.3.1). Nous avons conservé ici le seuil de 2Å sur les valeurs de RMSD, car diminuer légèrement ce seuil (par exemple à 1.5Å) ne change pas drastiquement les classifications obtenues.

La classification boucle-1,3 contient 6 classes (parfois recouvrantes) d'occurrences non homologues, qui regroupent 148 occurrences, sur un total de 81 classes.



(a) boucle interne



(b) tetraloop GNRA

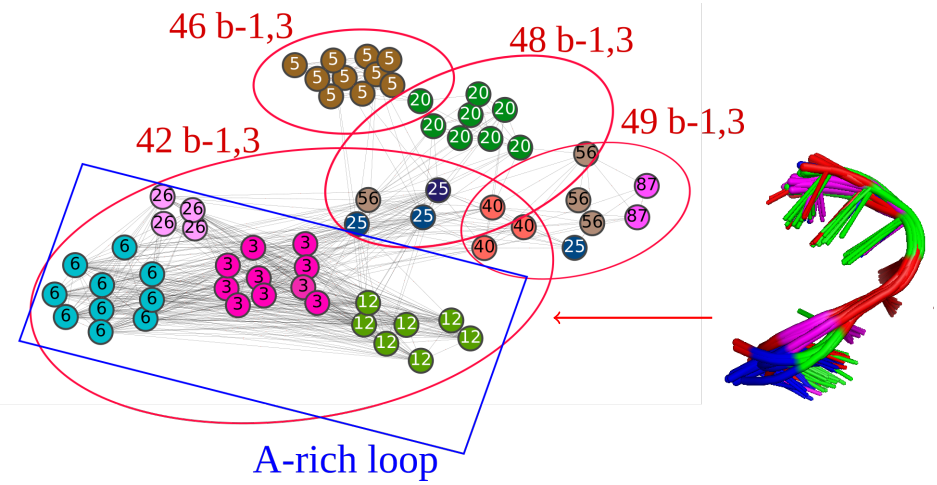
Figure 4.4 – Présentation des classes composées d'occurrences non homologues dans les classifications *boucle-1,2,3* et *boucle-1,3,4*. Dans les graphes, les arêtes en pointillés indiquent que la RMSD entre les deux sous-contextes 3D boucle-1,2,3 est en-dessous du seuil de 2 Å, contrairement à la RMSD entre les deux sous-contextes 3D boucle-1,3,4. Les arêtes en tirets indiquent que la RMSD entre les deux sous-contextes 3D boucle-1,3,4 est en-dessous du seuil de 2 Å, contrairement à la RMSD entre les deux sous-contextes 3D boucle-1,2,3. Les arêtes en trait plein indiquent que la RMSD pour les deux sous-contextes est inférieure au seuil de 2Å.

Les occurrences de motif A-minor appartenant aux classes des classifications à 3 branches numérotées 49 *b-1,2,3*, 51 *b-1,3,4* et 61 *b-1,3,4* (Figure 4.4) forment une seule classe dans la classification *boucle-1,3* à 2 branches, numérotée 37 *b-1,3* dans la Figure 4.5b. Il s'agit des motifs impliqués dans une boucle GNRA. Les contextes 3D des branches 1 et 3 pour ces motifs sont donc très similaires. On peut noter que cette classe regroupe la totalité des occurrences de motif A-minor de notre jeu de données qui impliquent une boucle GNRA. D'autres classes composées d'occurrences homologues regroupent des occurrences de motifs A-minor impliquant une boucle de 4 nucléotides (*tetraloop*), mais le contexte 3D ainsi que la signature de séquence sont différents de ceux d'une boucle GNRA.

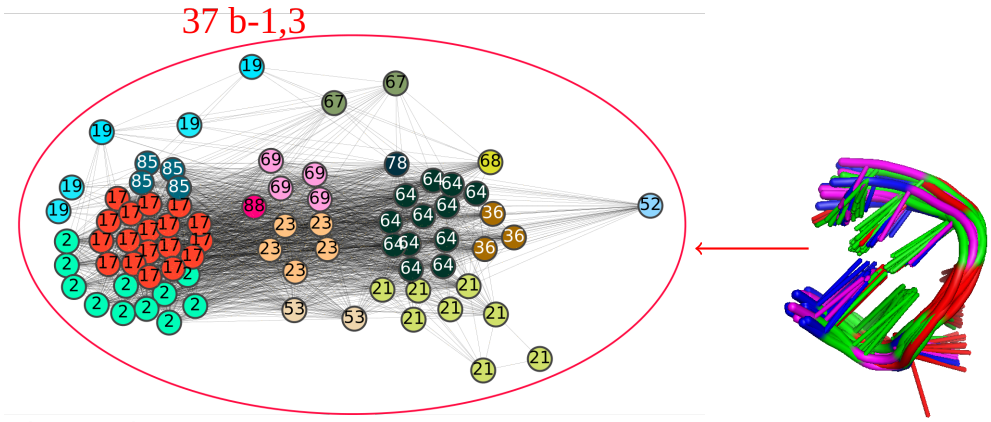
Nous avons également comparé ces résultats aux résultats obtenus par la méthode RNAMotifContrast [48]. Dans leur étude, les auteurs établissent des sous-familles de motifs locaux d'ARN, en fonction de leur similarité de structures 3D. Les motifs étudiés sont donc des motifs apparaissant au sein d'un élément de structure secondaire d'ARN, comme les boucles GNRA par exemple. Notons que les auteurs ont considéré toutes les boucles GNRA, apparaissant dans des structures non redondantes de la PDB. Par conséquent, nombre de ces boucles ne sont pas impliquées dans un motif A-minor. Les auteurs obtiennent alors plusieurs sous-familles de boucles GNRA, dont l'une contient la majorité des occurrences. Les boucles GNRA de notre jeu de données se trouvent toutes dans cette famille prépondérante. Les autres boucles GNRA considérées avec RNAMotifContrast sont des boucles qui ne sont pas impliquées dans un motif A-minor. Nos résultats sont donc cohérents avec ceux de cette étude, puisque nous regroupons l'ensemble des occurrences de motif A-minor impliquant une boucle GNRA dans une seule et même classe avec la classification boucle-1,3. Nous pouvons également en déduire que les boucles GNRA dans les motifs A-minor sont des boucles GNRA relativement classiques.

Pour les classes impliquant des boucles internes, de nouvelles sous-classes apparaissent avec la classification boucle-1,3 à 2 branches. La classe 42 *b-1,3* (Figure 4.5a) est la plus fournie et contient la classe 50 *b-1,2,3* (Figure 4.4) ainsi qu'une grande partie de la classe 24 *b-1,2,3* (Figure 4.4) de la classification *boucle-1,2,3*. Dans cette classe 42 *b-1,3*, se trouvent toutes les occurrences de motif A-minor impliquées dans une boucle interne de type A-rich, et quelques autres motifs, qui possèdent un contexte 3D similaire, mais à qui il manque les interactions non canoniques qui feraient d'eux des boucles de type A-rich (voir Figure 4.2b).

On peut également noter que les sous-graphes induits par les occurrences de motif de ces deux classes principales (classes 37 *b-1,3* et 42 *b-1,3* de la classification *boucle-1,3*) a une densité très faible (0.20 et 0.24 voir Table 4.3). Cela signifie que certaines paires de sous-contextes 3D ne sont pas aussi similaires que d'autres. C'est ce qu'on remarque aussi en observant les alignements 3D (voir Figure 4.5). Le clustering permet cependant de regrouper ces occurrences de motif, malgré leurs légères disparités de contextes 3D, ce qui est cohérent avec les sous-familles de motifs A-minor déjà connues.



(a) boucle interne



(b) tetraloop GNRA

Figure 4.5 – Présentation des classes de non homologues dans la classification *boucle-1,3*, à 2 branches. L'alignement 3D présenté en (a) correspond à l'alignement 3D des occurrences de la classe 42 *b-1,3*. Les occurrences de la classe 42 *b-1,3* dans le rectangle bleu possèdent une *A-rich loop*. La classe 37 *b-1,3* (en (b)) contient les occurrences de motif des classes 49 *b-1,2,3*, 51 *b-1,3,4* et 61 *b-1,3,4* des classifications à 3 branches (Figure 4.4).

Nous avons également souhaité étudier la classification *hélice-2,4*, qui prend en compte uniquement les branches 2 et 4, soit les branches de l'hélice de structure secondaire. Cette représentation ne permet cependant pas de faire émerger des classes d'occurrences non homologues. Cette observation tend à indiquer que le brin de la boucle dans le contexte du motif A-minor porte davantage d'informations structurales que le brin de l'hélice.

Considérer le brin de l'hélice uniquement permet seulement de rapprocher les occurrences homologues qui n'appartenaient pas à la même classe dans la classification à 4 branches (voir Figure 4.3, classes 44 et 47). En effet, les différences observées dans l'alignement 3D entre ces occurrences de motif se trouvent dans la partie de la boucle.

Numéro de classe	Nombre d'occurrences	Nombre de groupes d'homologie	Densité de classe	Type de boucle	Famille(s) d'ARN	Règnes des organismes
40 (b)	10	2	0,64	Boucle interne	ARNr 23S	Bactéries, Archées
42 (b)	38	7	0,20	Boucle interne, A-rich	ARNr 23S, ARNr 25S, ARNr 28S, ARNr 16S	Bactéries, Archées, Eucaryotes
46 (b)	11	2	0,82	Boucle interne, boucle externe	ARNr 23S, ARNr 25S, ARNr 28S	Bactéries, Archées, Eucaryotes
48 (b)	13	4	0,38	Boucle interne	ARNr 23S, ARNr 16S	Bactéries, Archées
49 (b)	9	4	0,61	Boucle interne	ARNr 16S, ARNr 23S, Intron de groupe IIC	Bactéries, Archées
37 (b)	77	15	0,24	GNRA	ARNr 23S, ARNr 25S, ARNr 28S, ARNr 16S, ARNr 18S, Intron de groupe IIC, Riboswitch c-di-GMP	Bactéries, Archées, Eucaryotes, non spécifié

Table 4.3 – Description des classes composées d'occurrences non homologues, pour la classification à 2 branches (boucle-1,3)

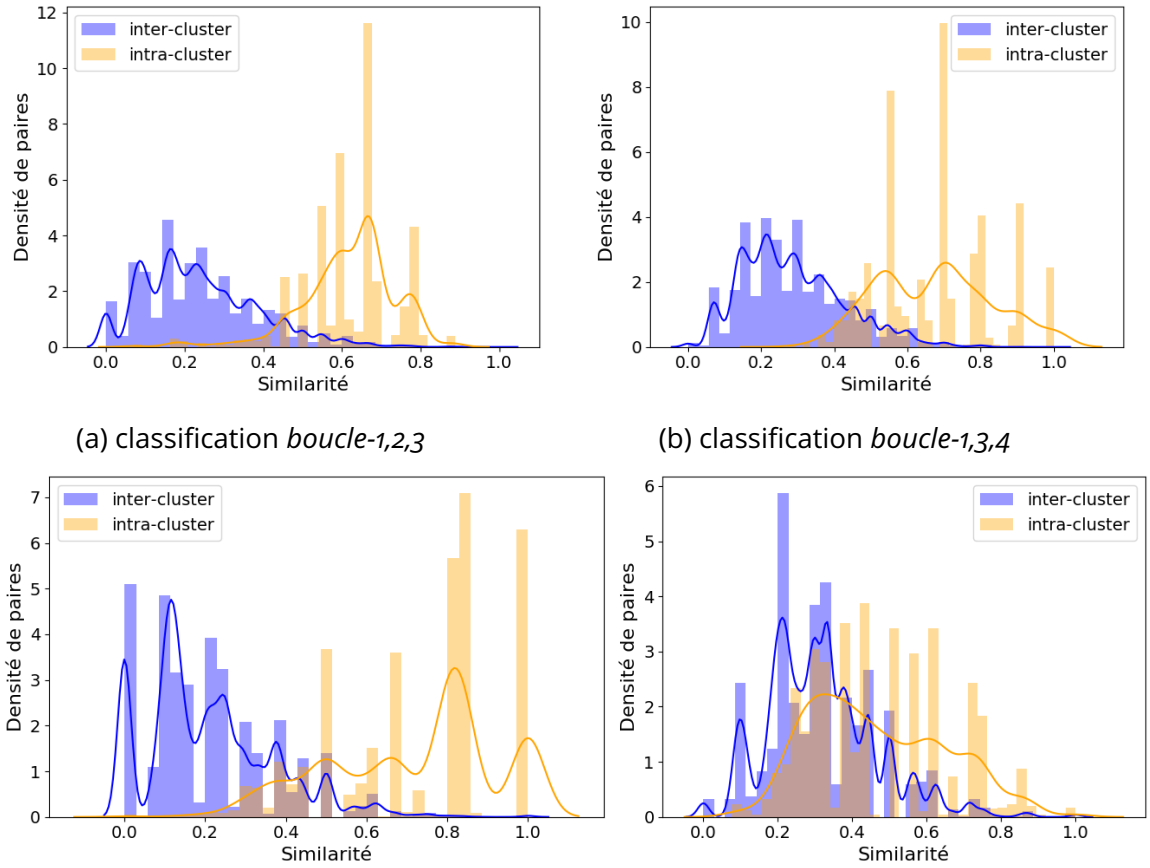
4.4 Retour à la similarité contextuelle

Dans le chapitre 3, nous avons comparé les résultats obtenus en classifiant les occurrences de motif A-minor selon leur contexte 3D et selon leur contexte topologique. Puis, ici, nous avons défini plusieurs sous-contextes 3D. Nous pourrions alors nous demander si la corrélation entre le contexte topologique et le contexte 3D peut également être observée pour ces sous-contextes.

Pour le savoir, nous considérons les k-extensions contractées induites par les mêmes branches que les sous-contextes définis dans la partie 4.3.1, et les comparons deux à deux en calculant la similarité contextuelle associée. Nous étudions ensuite la distribution des valeurs de similarité contextuelle au sein des classes définies selon le sous-contexte 3D correspondant (Figure 4.6).

Nous observons alors que d'une manière générale, les paires d'occurrences de motif appartenant à une même classe possèdent des similarités contextuelles plus élevées que les paires appartenant à des classes différentes, pour les classifications boucle-1,3, boucle-1,2,3 et boucle-1,3,4. En revanche, ce n'est pas le cas pour la classification hélice-2,4 (Figure 4.6d).

Ainsi, les métriques de similarité de sous-contextes topologiques et de sous-contextes 3D sont corrélées, sauf pour le sous-contexte composé des branches de l'hélice du motif uniquement. Cette observation est cohérente avec les conclusions faites sur les classifications : observer le contexte du motif A-minor uniquement sur les branches de l'hélice apporte peu d'informations.



(a) classification *boucle-1,2,3*

(b) classification *boucle-1,3,4*

(c) classification *boucle-1,3*

(d) classification *hélice-2,4*

Figure 4.6 – Distribution des paires d'occurrences de motif A-minor, en densité de paires (section 3.7), appartenant ou non à la même classe dans les différentes classifications selon la RMSD, en fonction des valeurs de similarité contextuelle.

4.5 Comparaison avec une autre classification de motifs A-minor

Comme évoqué dans l'introduction, peu d'études ont cherché à classifier les motifs A-minor de manière exhaustive selon des critères de similarité de topologie ou de 3D. Une étude de 2021 a cependant classifié les motifs A-minor selon des considérations de structure secondaire et de séquence [101].

Les auteurs de [101] classent les motifs A-minor en fonction du type d'éléments de structure secondaire impliqué dans les occurrences de motif, et en fonction de leur position relative les uns par rapport aux autres (interactions locales ou de longue distance). Dans la suite de cette section, nous nommerons cette classification *classification selon les SSE (éléments de structure secondaire)*.

Le jeu de données utilisé par les auteurs est légèrement différent du nôtre, car ils utilisent une méthode différente pour annoter les motifs A-minor (le programme DSSR [69]), et considèrent les classes de structures équivalentes

du groupe BGSU RNA (Bowling Green State University) [64] pour ne conserver que des occurrences non redondantes. Le groupe BGSU a en effet construit des classes de structures PDB équivalentes, telles qu'une classe de structures équivalentes regroupe toutes les structures PDB de la même molécule dans le même organisme. Une structure représentante pour chaque classe de structures équivalentes est sélectionnée selon différents critères de qualité d'une structure 3D : la résolution de la structure, la fraction de la molécule observée, et d'autres métriques liées à la méthode d'obtention de la structure (cristallographie, cryoEM, etc.). Les auteurs de la classification selon les SSE ont ainsi pris en compte les occurrences de motif A-minor se trouvant dans ces structures représentantes, pour s'assurer de ne pas avoir d'occurrences redondantes. Nous n'avons pas utilisé cette méthode.

De plus, nous ne considérons que les motifs A-minor de type I/II, et eux considèrent toutes les interactions de type A-minor (voir chapitre 1, section 1.5.3). Les deux jeux de données ont ainsi 297 occurrences d'A-minor en commun (sur 391 occurrences de motifs dans notre jeu de données).

Parmi les classifications que nous avons définies, la plus similaire à la classification selon les SSE est la classification *boucle-1,3*, à 2 branches. Nous allons ici comparer la classification selon les SSE, à la classification boucle-1,3, présentée en section 4.3.3, qui regroupe des occurrences de motif A-minor possédant des contextes 3D similaires, et à la classification qu'on obtient à partir des k-extensions correspondantes, regroupant donc des occurrences possédant des topologies de contexte similaires. Plus précisément, comme précédemment, nous considérons les k-extensions contractées induites par les sommets des branches 1 et 3, calculons les similarités contextuelles de chaque paire de sous-k-extensions, et en déduisons une classification des occurrences selon la similarité de topologie. Pour différencier nos deux classifications, nous les nommerons *classification boucle-1,3 avec la RMSD* et *classification boucle-1,3 avec la similarité contextuelle*. Nous pouvons noter que ces deux classifications possèdent des différences (elles ont un indice de Jaccard de 0.75, voir chapitre 3, section 3.6.3 pour la définition) mais sont cohérentes l'une avec l'autre comme le montre la Figure 4.6c.

Est présentée en Table 4.4 la comparaison de nos classifications boucle-1,3 avec la classification selon les SSE, en termes de nombre de paires d'occurrences groupées dans une même classe. Nous observons ainsi qu'environ 80% des paires d'occurrences appartenant à une même classe, dans notre classification utilisant la RMSD comme dans notre classification utilisant la similarité contextuelle, sont également groupées au sein d'une même classe dans la classification selon les SSE (2774/3488 et 2213/2809). D'un autre côté, moins du tiers des paires d'occurrences appartenant à une même classe dans la classification selon les SSE sont regroupées dans une même classe dans nos classifications (2774/9718 et 2213/9718). D'une manière générale, les classes de la classification selon les SSE sont en fait plus grandes et regroupent plusieurs de nos classes.

Ainsi, les deux classifications que nous définissons semblent être un raffinement des classes définies par [101].

	Comparaison classification *boucle-1,3* avec la RMSD et classification selon les SSE	Comparaison classification *boucle-1,3* avec la similarité contextuelle et classification selon les SSE
Nombre de paires d'occurrences classifiées dans la même classe dans les deux classifications considérées	2774	2213
Nombre de paires d'occurrences classifiées dans la même classe dans la classification *boucle* (RMSD ou similarité contextuelle)	3488	2809
Nombre de paires d'occurrences classifiées dans la même classe dans la classification selon les SSE	9718	9718

Table 4.4 – Comparaison de la classification selon les SSE avec deux des classifications que nous définissons pour les occurrences de motif A-minor : les classifications *boucle-1,3* sur les sous-structures 3D locales (RMSD), ou sur les 4-extensions (similarité contextuelle). Cette comparaison se fait en termes de nombre d'occurrences classifiées ou non dans une même classe dans les deux classifications comparées.

Par exemple, les deux classes les plus conséquentes de la classification boucle-1,3 avec la RMSD (classes 37 *b-1,3* et 42 *b-1,3*, voir Figure 4.5) possèdent chacune presque uniquement des occurrences appartenant à la même classe dans la classification selon les SSE. Pour la classe 37 *b-1,3*, la classe correspondante dans la classification selon les SSE contient des occurrences de motif, formées d'une boucle terminale et d'une hélice, reliées entre elles par des interactions non canoniques longue distance. D'autres classes parmi les classes de la classification boucle-1,3 avec la RMSD entrent dans cette catégorie également, et appartiennent donc à la même classe dans la classification selon les SSE. De la même façon, les occurrences de motif de la classe 42 *b-1,3* appartiennent presque toutes à la classe de la classification selon les SSE, dans laquelle les occurrences de motif sont formées d'une boucle interne reliée à une hélice par des interactions non canoniques longue distance.

La différence principale entre la classification selon les SSE et nos classifications réside dans la définition du contexte structural. Nous considérons le contexte structural à très courte distance sur la séquence, soit

par la position dans l'espace des nucléotides de ce contexte, soit par les interactions canoniques et non canoniques y apparaissant. Nous n'utilisons pas les informations de structure secondaire, ni la séquence. Quant à eux, les auteurs de cette étude [101] ne prennent pas en compte les interactions non canoniques, et s'intéressent aux éléments de structures secondaires et à leur position relative. Cela peut expliquer les différences observées entre les résultats de classification. Mais nous pouvons tout de même remarquer la cohérence des deux classifications.

4.6 Conclusion

Dans ce chapitre, nous avons donc montré que la classification à 4 branches que l'on définit selon la similarité du contexte 3D permet de détecter un petit nombre d'occurrences non homologues de motif A-minor possédant des caractéristiques structurales communes. Il semble cependant que certaines branches du contexte 3D présentent davantage de similarités que d'autres. C'est en particulier le cas des branches du brin de la boucle de structure secondaire. On voit ainsi d'autres similarités apparaître en ne considérant que 3 branches sur 4 (les deux branches de la boucle et une des branches de l'hélice), et d'autant plus en ne considérant que les deux branches de la boucle (formant le brin de la boucle), qui permettent alors d'obtenir une classification selon le type de boucle.

Ces observations tendent à indiquer que le contexte 3D du brin de la boucle dans le motif A-minor est mieux conservé que celui du brin de l'hélice dans les occurrences de motif, et a peut-être alors davantage d'importance dans la formation du motif.

Nous avons également montré que ces classifications de sous-contextes 3D sont en corrélation avec la topologie de ces contextes, comme c'est aussi le cas pour le contexte 3D dans sa totalité.

Enfin, nous avons montré que nos classifications selon des similarités de contexte 3D ou de topologie de contexte, sur le brin de la boucle uniquement, sont cohérentes avec une classification des occurrences de motif A-minor établie récemment (2021) [101], par une approche différente s'intéressant à la structure secondaire locale et à la séquence.

Ainsi, après avoir montré la corrélation entre des similarités de contexte 3D et des similarités de topologie de contexte (chapitre 3) entre occurrences de motif A-minor, et après avoir observé la présence de classes de contextes 3D similaires transcendant les relations d'homologie (chapitre 4), nous allons tenter de découvrir, dans le dernier chapitre de cette thèse, si certaines de ces similarités de contexte 3D peuvent être déduites des similarités de topologie de contexte et de séquence, dans un but prédictif.

Chapitre 5

Vers la prédiction du motif A-minor

5.1 Introduction

Dans les chapitres précédents, nous nous sommes intéressés au contexte structural de différents motifs d'ARN, et en particulier celui du motif A-minor, à deux échelles différentes, que sont la topologie du contexte et le contexte 3D, ce dernier contenant à la fois la topologie et la position des atomes dans l'espace. Le contexte 3D portant davantage d'information, nous l'avons utilisé pour classifier les motifs A-minor de différentes manières, en faisant le lien avec des classifications déjà existantes (chapitre 4). Cependant, nous avons également montré dans le chapitre 3 que même si le contexte 3D porte davantage d'information, dans la plupart des cas, la topologie du contexte structural est suffisante pour expliquer les similarités de contexte 3D ainsi que les relations d'homologie entre occurrences de motif.

Dans ce chapitre, nous allons étudier la capacité de prédiction de cette topologie de contexte. Nous considérons des classes d'occurrences de motif A-minor, regroupant des occurrences partageant des contextes 3D similaires, et nous représentons ces classes par leur topologie commune, ainsi que par une notion de séquence commune, pour étudier également la capacité de prédiction de la séquence.

Nous présentons alors une méthode permettant d'étudier la prédictibilité de ces représentants, c'est-à-dire la fréquence d'apparition de ces représentants dans des séquences et graphes d'ARN quelconques. Moins ils sont fréquents en l'absence de motifs A-minor, plus ils sont de bons critères de prédiction.

Nous présenterons d'abord, dans la section 5.2, la méthode permettant de calculer les représentants et de rechercher des occurrences de ces représentants. Nous aborderons d'abord les jeux de données utilisés, comportant un jeu de validation et un jeu de test. Puis, nous définirons précisément les représentants de classe, et la façon de rechercher des occurrences de ces représentants dans les séquences et les graphes d'ARN des jeux de données. Nous décrirons également les mesures statistiques que l'on utilisera pour évaluer les résultats.

La section 5.3 enfin portera sur l'analyse des résultats de prédictibilité, en fonction des différentes mesures d'évaluation.

5.2 Calcul et recherche de représentants à une classe de motifs A-minor

5.2.1 Présentation des données utilisées

Dans cette section 5.2.1, nous allons décrire les classes d'occurrences de motif A-minor que nous considérons, ainsi que les jeux de données de molécules d'ARN que nous utilisons.

Nous considérons les classes de plus de 3 occurrences, et contenant des occurrences de motifs intramoléculaires, dans notre classification à 4 branches selon la RMSD, présentée dans les chapitres 3 (section 3.7.2) et 4 (section 4.2).

Cette classification est obtenue selon des similarités de contexte 3D entre occurrences de motif A-minor. Parmi les différentes classifications utilisant la similarité de contexte 3D que nous avons présentées dans le chapitre précédent, nous avons choisi la classification à 4 branches, car elle regroupe des occurrences de motif A-minor ayant des topologies de contexte suffisamment similaires pour pouvoir étudier la capacité de prédiction de la topologie de contexte.

Nous pouvons également noter que nous ne considérons dans notre étude que les classes comprenant plus de 3 occurrences de motifs A-minor, car les autres possèdent des topologies trop particulières.

Nous recherchons des représentants de ces classes, que nous allons définir dans la partie suivante, dans deux types de jeux de données, que nous appellerons jeu de validation et jeu de test. Chacun d'entre eux est constitué d'un ensemble de séquences et de graphes d'ARN, extraits des structures de la PDB. Les graphes d'ARN en particulier sont obtenus à partir des structures 3D, à l'aide du programme FR3D [98], comme expliqué dans le chapitre 3 (section 3.2.2).

Nous considérons les structures PDB contenant au moins une occurrence de motif A-minor appartenant à une des classes de 3 occurrences ou plus de la classification à 4 branches. Les séquences et graphes d'ARN que l'on obtient à partir de ces structures constituent le jeu de données de validation. Ce jeu de données contient ainsi 136 séquences et graphes d'ARN (voir Table 5.2a), comprenant 374 occurrences de motifs A-minor intramoléculaires.

Nous allons également rechercher les représentants de ces mêmes classes dans toutes les autres structures non redondantes de la PDB, formant le jeu de données de test. Pour constituer cet ensemble non redondant de structures, nous avons utilisé les structures représentantes définies par le groupe BGSU RNA [64] (voir chapitre 4, section 4.5).

Certaines des structures du jeu de données de test contiennent des occurrences de motif A-minor. Ces structures (récupérées en juin 2020) ont été

ajoutées à la PDB après la première extraction d'occurrences de motifs A-minor que nous avions faite (en 2019, voir chapitre 3, section 3.2.2).

Le jeu de données de test contient ainsi 1446 séquences et graphes d'ARN, comprenant 98 occurrences de motifs A-minor (voir Tables 5.1 et 5.2b). Ces occurrences de motif A-minor sont en majorité trouvées dans des ARN ribosomiques (85%) comme pour le jeu de données de 2019 que nous avons utilisé pour construire nos classes.

Famille d'ARN / Organisme	ARNr 12S	ARNr 28S	ARNr 23S	ARNr 16S	ARNr alpha	ARNr beta	ARNr gamma	ARNt	anti-toxine	Ribo-zyme	Ribo-switch	Autres	**Total**
S. scrofa (E)	6												**6**
O. cuniculus (E)		9											**9**
A. baumannii (B)			15	7									**22**
P. aeruginosa (B)			18	6									**24**
T. celer (A)				6									**6**
H. sapiens (E)	1												**1**
T. cruzi (E)					4	1							**5**
L. donovani (E)					5	5	1						**11**
T. thermophilus (B)								1				1	**2**
E. coli (B)								2				2	**4**
S. enterica (B)								1					**1**
Hepatovirus A												1	**1**
C. Pelagibacter (B)											1		**1**
Geobacter (B)											1		**1**
Non spécifié									1	2	1		**4**
Total	**7**	**9**	**33**	**19**	**9**	**6**	**1**	**4**	**1**	**2**	**3**	**4**	**98**

Table 5.1 – Nombre d'occurrences de motifs A-minor du jeu de données de test, triées par organismes et familles d'ARN. Pour chaque organisme, il est indiqué entre parenthèses à quel règne du Vivant il appartient (Bactéries (B), Archées (A) ou Eucaryotes (E))

5.2.2 Définitions des représentants de classes et méthode de recherche

Nous allons nous intéresser à trois types de représentants de classes, que nous allons détailler dans cette partie. Le premier permettra de représenter les classes d'occurrences par des signatures de séquence et les deuxième et troisième par des topologies communes.

Nous expliquerons également comment rechercher les occurrences de ces représentants dans les séquences et graphes d'ARN des jeux de données définis plus haut.

Dans cette partie, nous noterons $\mathscr{E}_i = \{\tilde{G}_{i,1}, \tilde{G}_{i,2}, ..., \tilde{G}_{i,n_i}\}$ un sous-ensemble de k-extensions contractées (voir chapitre 2 définition 2.2.8) formant une classe d'occurrences de motifs A-minor de taille n_i dans la classification à 4 branches.

Nous considérerons également $G_{\mathscr{E}_i}$ le sous-graphe commun maximum (voir chapitre 2, définition 2.5) à cette classe $\mathscr{E}_i$.

Famille d'ARN	Nombre de structures PDB
ARNr 23S	62
ARNr 16S	29
Riboswitch	18
ARNr 18S	8
Ribozyme	8
ARNr 25S	5
Intron	5
ARNr 28S	1
Total	**136**

(a) Jeu de validation

Famille d'ARN	Nombre de structures PDB
ARNt	66
Riboswitch	39
Ribozyme	20
ARNr 5S	19
snRNA (*small nuclear*)	14
crRNA (crispr)	12
ARNr 16S	9
virus	9
ARNr 23S	8
ARNm	8
sgRNA (synthetic guide)	7
SRP	6
ARNr 12S	3
ARNr 5.8S	3
ARNr beta	3
ARNr 28S	2
ARNr alpha	2
ARNr gamma	2
ARNr epsilon	2
ARNr zeta	2
antitoxine	2
Intron	1
ARNr delta	1
Autres (taille de séquence supérieure à 50)	86
Autres (taille de séquence inférieure à 50)	1120
Total	**1446**

(b) Jeu de test

Table 5.2 – Nombre de structures PDB par famille d'ARN dans les deux jeux de données (validation (a) et test (b)).

Calcul de représentants de séquence à l'aide d'expressions régulières

Chaque classe sera caractérisée par deux ensembles d'expressions régulières. Rappelons qu'une expression régulière [27] est une chaîne de caractères permettant de décrire un langage dans un alphabet $\mathscr{A}$ donné. L'alphabet que l'on considère ici est $\mathscr{A} = \{A, C, G, U\}$, comprenant donc les lettres représentant les différents types de nucléotides dans les molécules d'ARN.

Nous considérons pour chaque k-extension contractée $\tilde{G}_{i,j}$, la k-extension non contractée $G_{i,j}$ dont elle est issue par les opérations de contraction définie dans le chapitre 2 (section 2.2.3). A chaque sommet u de la k-extension non contractée $G_{i,j}$, est associé le type de nucléotides correspondant dans la séquence d'ARN, que l'on notera $nt(u)$.

C'est à partir de ces k-extensions non contractées que nous allons définir les ensembles d'expressions régulières à une k-extension de la classe $\mathscr{E}_i$.

Nous définirons d'abord des chemins dans un sous-graphe particulier de k-extension non contractée. Nous définirons également d'autres sous-graphes

d'une k-extension non contractée, obtenus à partir du graphe décontracté du sous-graphe commun maximum à la classe. Puis à partir de ces deux éléments, nous définirons deux ensembles d'expressions régulières pour chaque k-extension contractée d'une classe $\mathscr{E}_i$. Ensuite, nous définirons des classes d'équivalence partitionnant les expressions régulières de toutes les k-extensions contractées de la classe. Et enfin, à partir de ces classes d'équivalence, nous définirons deux ensembles d'expressions régulières pour la classe $\mathscr{E}_i$.

Ainsi, nous définissons deux chemins dans un sous-graphe de k-extension non contractée. Un exemple est présenté en Figure 5.1.

Définition 5.2.1 ***Sous-graphe de séquence***

Soit $\tilde{G}_{i,j}$ une k-extension contractée de la classe $\mathscr{E}_i$.

Soit $G_{i,j}$ la k-extension non contractée dont est issue la k-extension contractée $\tilde{G}_{i,j}$.

***Le sous-graphe de séquence** de $G_{i,j}$, noté $\mathbf{G_{i,j}^{A_P}}$, est le sous-graphe couvrant de $G_{i,j}$ ne contenant que les arcs correspondant aux liaisons covalentes de la séquence primaire (noté A_P dans les graphes d'ARN, voir chapitre 2, définition 2.2.1).*

Dans ce sous-graphe, il existe en particulier deux chemins, l'un contenant les sommets 1 et 3 de l'occurrence de motif et l'autre contenant les sommets 2 et 4 de l'occurrence de motif (voir définition du motif A-minor, chapitre 2, définition 2.2.2, et voir un exemple dans la Figure 5.1). Nous définissons formellement ces chemins de la manière suivante.

Définition 5.2.2 ***Chemins $\mathbf{W_{B,j}}$ et $\mathbf{W_{H,j}}$***

Soit $G_{i,j}$ la k-extension non contractée dont est issue la k-extension contractée $\tilde{G}_{i,j}$.

Soit $G_{i,j}^{A_P}$ le sous-graphe de séquence de la k-extension non contractée $G_{i,j}$.

Nous appelons $\mathbf{W_{B,j}}$ (resp. $\mathbf{W_{H,j}}$) le chemin le plus long de $G_{i,j}^{A_P}$ contenant les sommets 1 et 3 de l'occurrence de motif et uniquement des sommets des branches 1 et 3 (resp. les sommets 2 et 4 de l'occurrence de motif et uniquement des sommets des branches 2 et 4).

Les branches des k-extensions sont définies dans le chapitre 2, section 2.2.2.

Ces deux chemins sont en général sommets-disjoints. Dans de rares cas, ils peuvent posséder des sommets en commun.

L'ensemble des sommets couverts par ces chemins est ainsi le même ensemble que celui que nous avions considéré pour définir les sous-structures 3D locales à partir des k-extensions (voir chapitre 3, section 3.3). Ce sont les seuls sommets correspondant nécessairement à des nucléotides consécutifs sur la séquence primaire, et qui peuvent donc être utilisés pour construire les expressions régulières.

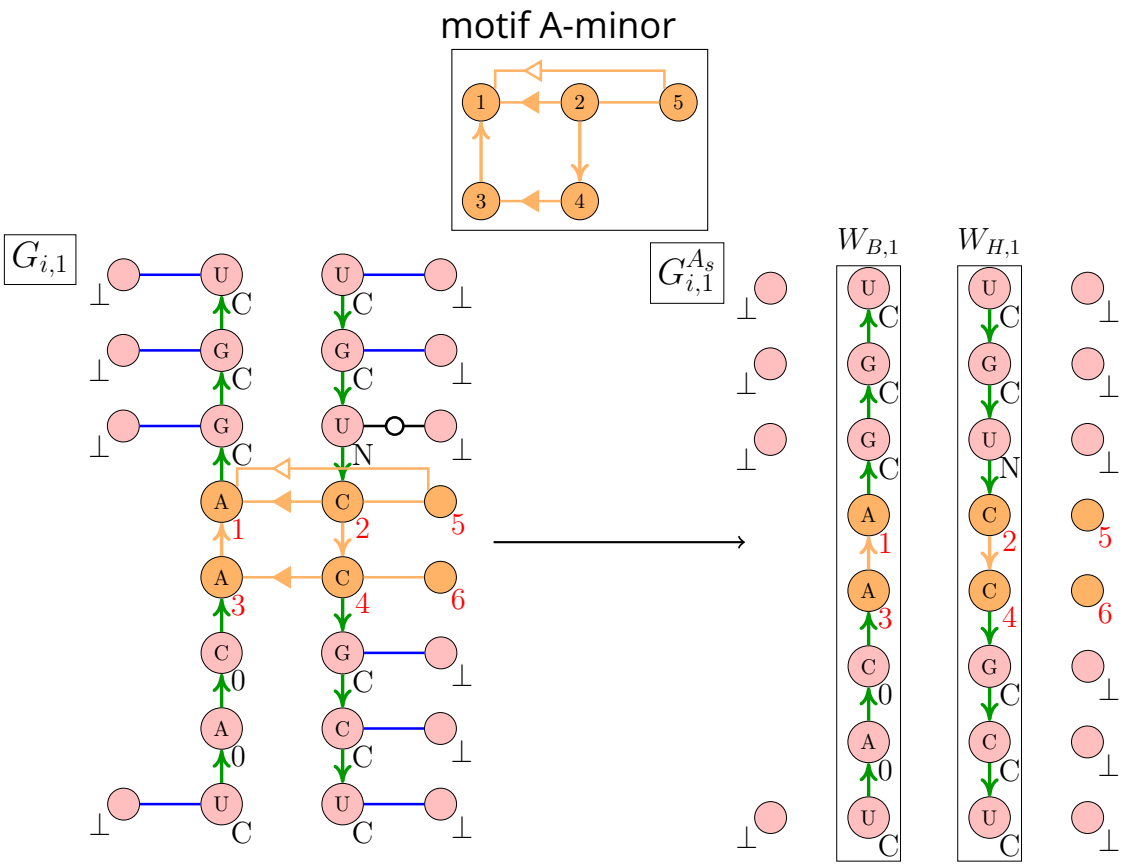


Figure 5.1 – Obtention du sous-graphe de séquence $G^{A_P}_{i,1}$ d'une 4-extension non contractée $G_{i,1}$. Le sous-graphe $G^{A_P}_{i,1}$ contient les chemins $W_{B,1}$ et $W_{H,1}$, encadrés en noir. Le chemin $W_{B,1}$ contient les sommets 1 et 3 de l'occurrence de motif, dont les numéros sont rappelés dans le motif A-minor présenté dans l'encadré noir au-dessus. Le chemin $W_{H,1}$ contient les sommets 2 et 4 de l'occurrence de motif. Chaque sommet des deux graphes est annoté par son type ρ_O.

Rappelons de plus que chaque k-extension contractée $\tilde{G}_{i,j}$ d'une classe $\mathscr{E}_i$ contient un sous-graphe partiel $\tilde{G'}_{i,j}$, correspondant au (π_s, πæ)-isomorphisme, qui définit le sous-graphe commun maximum $G_{\mathscr{E}_i}$ à la classe $\mathscr{E}_i$ (voir chapitre 2, section 2.3.1).

Nous définissons un graphe dérivé de celui-ci.

Définition 5.2.3 ***Graphe*** $\tilde{\mathbf{G}}''_{\mathbf{i,j}}$

Soit $\tilde{G'}_{i,j}$ le sous-graphe de la k-extension contractée $\tilde{G}_{i,j}$, selon le (π_s, π_{ae})-isomorphisme définissant le sous-graphe commun maximum $G_{\mathscr{E}_i}$ à la classe $\mathscr{E}_i$.

Le ***graphe*** $\tilde{\mathbf{G}}''_{\mathbf{i,j}}$ *est le graphe dérivé de $\tilde{G'}_{i,j}$, contenant tous les sommets et tous les arcs de $\tilde{G'}_{i,j}$. Chaque arc de $\tilde{G''}_{i,j}$ possède le même type que l'arc correspondant dans $\tilde{G'}_{i,j}$. Chaque sommet u dans $\tilde{G'}_{i,j}$ possède le même type que le sommet correspondant dans $\tilde{G'}_{i,j}$, et possède un poids égal au poids minimum p_{min} des*

sommets équivalents à u au sens de l'isomorphisme dans tous les graphes $\tilde{G'}_{i,l}$ (l étant un entier compris entre 1 et n_i) de la classe $\mathscr{E}_i$.

Nous décontractons ensuite ce graphe.

Définition 5.2.4 ***Graphe décontracté de*** $\tilde{\mathbf{G}''}_{\mathbf{i,j}}$

Le graphe décontracté de $\tilde{G''}_{i,j}$ est le graphe qui, par les opérations de contraction définies dans le chapitre 2, section 2.2.3, permet d'obtenir le graphe $\tilde{G''}_{i,j}$.

Ce graphe est unique, car il n'y a qu'une seule manière de contracter les chemins contractables d'un graphe d'ARN, comme précisé dans le chapitre 2.

A partir de ce graphe décontracté, nous pouvons définir un ensemble de graphes, dans lesquels nous définirons les expressions régulières à une k-extension contractée donnée. Ces ensembles de graphes sont dessinés pour un exemple de classe dans la Figure 5.2.

Définition 5.2.5 ***Ensemble de sous-graphes*** $\mathbf{\Gamma_{i,j}}$

Soit $G_{i,j}$ la k-extension non contractée dont est issue la k-extension contractée $\tilde{G}_{i,j}$.

Soit $\tilde{G'}_{i,j}$ le sous-graphe de la k-extension contractée $\tilde{G}_{i,j}$, selon le (π_s, $\pi_{æ}$)-isomorphisme définissant le sous-graphe commun maximum $G_{\mathscr{E}_i}$ à la classe $\mathscr{E}_i$.

Soit $\tilde{G''}_{i,j}$ le graphe dérivé de $\tilde{G'}_{i,j}$, dans lequel chaque sommet possède un poids égal au poids minimum des sommets qui lui sont équivalents au sens de l'isomorphisme, dans les autres k-extensions contractées de la classe

L'ensemble de sous-graphes $\mathbf{\Gamma_{i,j}}$ *est l'ensemble des sous-graphes de $G_{i,j}$ qui sont (π_s, $\pi_{æ}$)-isomorphes au graphe décontracté de $\tilde{G''}_{i,j}$.*

Trouver tous les sous-graphes d'un graphe G quelconque qui sont isomorphes à un graphe H quelconque est un problème NP-difficile [], mais certaines particularités de nos k-extensions et la faible valeur de k rendent la recherche exhaustive de tous les sous-graphes possible en un temps raisonnable, si on considère chaque branche indépendamment. En effet, l'occurrence de motif sera présente dans tous les sous-graphes, et pour la valeur de k considérée ici (k = 4), le nombre de sommets dans une branche ne dépasse pas 8 sommets.

Notons également que le (π_s, $\pi_{æ}$)-isomorphisme de graphes que l'on définit ne tient pas compte des types de nucléotides. Cependant, comme les sous-graphes de $\Gamma_{i,j}$ sont des sous-graphes de $G_{i,j}$, qui est une k-extension non contractée dont les sommets sont étiquetés par le type de nucléotide correspondant, alors les sommets des graphes de $\Gamma_{i,j}$ sont également étiquetés par le type de nucléotide correspondant dans la séquence primaire.

A partir de ces différentes informations, on définit alors deux ensembles d'expressions régulières pour chaque k-extension contractée d'une classe. La Figure 5.2 présente l'obtention de ces expressions régulières pour les k-extensions contractées de la classe donnée en exemple.

Définition 5.2.6 ***Expressions régulières*** $\mathbf{S_{Bj}}$ ***et*** $\mathbf{S_{Hj}}$ ***d'une k-extension contractée*** $\tilde{\mathbf{G}}_{\mathbf{i,j}}$

Soit $G_{i,j}^{A_P}$ le sous-graphe de séquence de la k-extension non contractée $G_{i,j}$, dont est issue la k-extension contractée $\tilde{G}_{i,j}$.

Soient $W_{B,j}$ et $W_{H,j}$ les deux chemins de $G_{i,j}^{A_P}$ contenant respectivement les sommets 1 et 3 de l'occurrence de motif et les sommets 2 et 4 de l'occurrence de motif.

Soit $\tilde{G'}_{i,j}$ le sous-graphe de la k-extension contractée $\tilde{G}_{i,j}$, selon le (π_s, $\pi_{œ}$)-isomorphisme définissant le sous-graphe commun maximum $G_{\mathscr{E}_i}$ à la classe $\mathscr{E}_i$.

Soit $\tilde{G''}_{i,j}$ le graphe dérivé de $\tilde{G'}_{i,j}$, dans lequel chaque sommet possède un poids égal au poids minimum des sommets équivalents au sens de l'isomorphisme dans les autres k-extensions contractées de la classe.

Soit $\Gamma_{i,j}$ l'ensemble des sous-graphes de la k-extension non contractée $G_{i,j}$ qui sont (π_s, $\pi_{œ}$)-isomorphes au graphe décontracté de $\tilde{G''}_{i,j}$. Dans chacun de ces sous-graphes, les sommets sont étiquetés par le type de nucléotide correspondant.

L'ensemble d'expressions régulières $\mathbf{S_{Bj}}$ = $(s_{Bj,1}, s_{Bj,2}, ..., s_{Bj,m})$ *(avec m égal au nombre de graphes dans $\Gamma_{i,j}$) de la k-extension contractée $\tilde{G}_{i,j}$ est défini comme suit :*

A chaque graphe g_s de $\Gamma_{i,j}$ correspond une expression régulière $s_{Bj,s}$ de l'ensemble S_{Bj}, qui est construite ainsi :

Pour chaque sommet u_p en position p dans la séquence de sommets reliés par les arcs du chemin $W_{B,j}$:

- $s_{Bj,s}[p] = nt(u_p)$ *si u_p appartient au graphe g_s*
- $s_{Bj,s}[p] = N$ *sinon, c'est-à-dire que toutes les lettres de l'alphabet $\mathscr{A}$ sont autorisées pour cette position ($N = (A|G|C|U)$)*

L'ensemble d'expressions régulières $\mathbf{S_{Hj}}$ *est définie de la même façon à partir du chemin $W_{H,j}$.*

Notons ainsi que, parmi les sommets induisant les chemins $W_{B,j}$ et $W_{H,j}$, nous ne prenons en compte les types de nucléotides que pour les sommets appartenant au sous-graphe $\tilde{G'}_{i,j}$ de la k-extension contractée. Autrement dit, nous ne considérons que les types de nucléotides appartenant à la topologie commune à la classe.

A présent, pour définir les expressions régulières à la classe $\mathscr{E}_i$, nous allons définir des classes d'équivalence permettant de partitionner les expressions régulières des k-extensions contractées de la classe.

Définition 5.2.7 ***Expressions régulières compatibles***

Soient $S_{B,\mathscr{E}_i} = [S_{B1}, S_{B2}, ..., S_{Bn_i}]$ et $S_{H,\mathscr{E}_i} = [S_{H1}, S_{H2}, ..., S_{Hn_i}]$ les ensembles d'expressions régulières de toutes les k-extensions contractées de la classe $\mathscr{E}_i$.

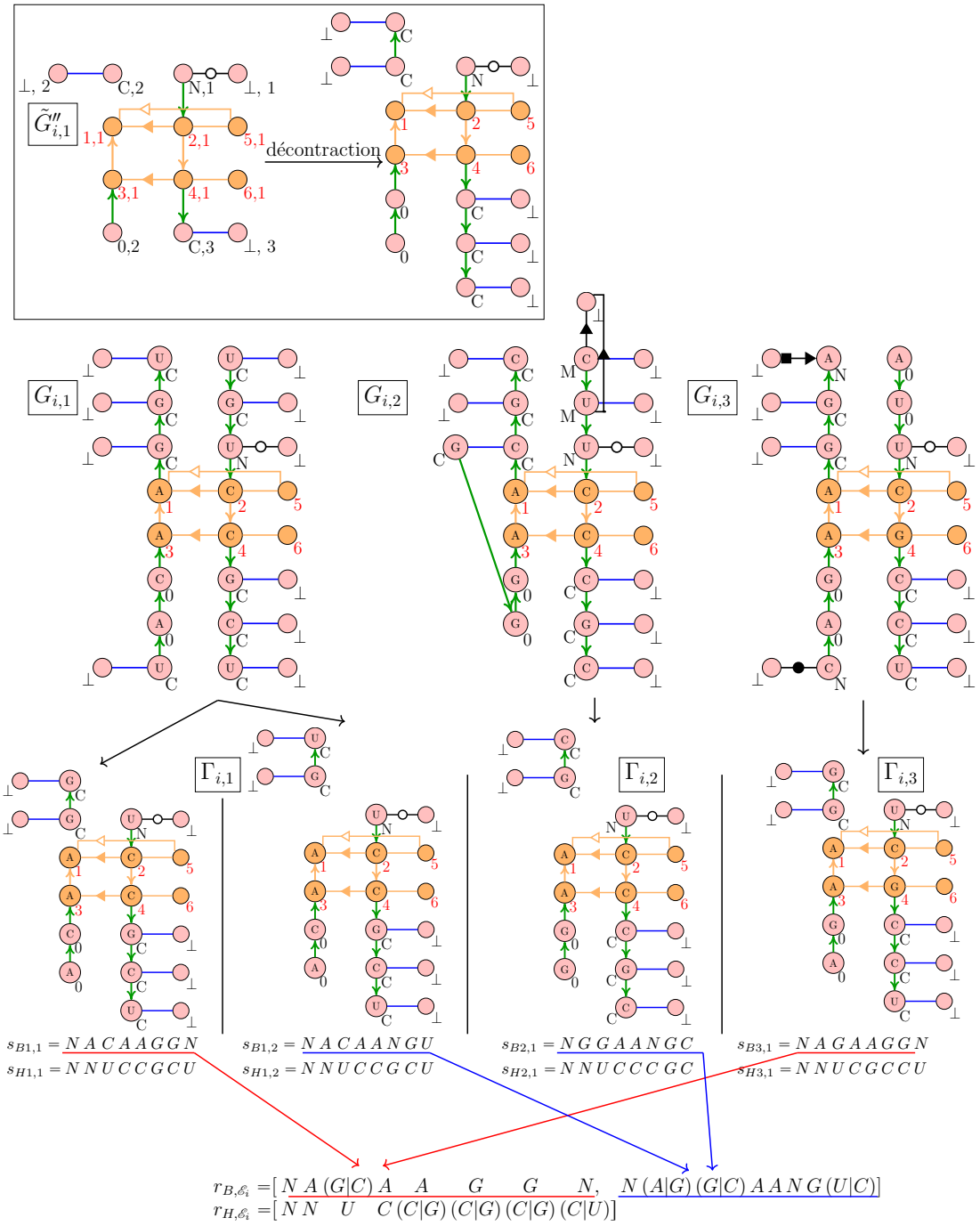


Figure 5.2 – Exemple d'une classe de trois occurrences de motif A-minor. Dans la partie droite de l'encadré noir en haut, est présenté le graphe décontracté du sous-graphe $\tilde{G}''_{i,1}$ de la 4-extension $\tilde{G}_{i,1}$. Puis, en-dessous de l'encadré noir, sont présentées les 4-extensions non contractées des occurrences de la classe : $G_{i,1}$, $G_{i,2}$ et $G_{i,3}$. Au milieu sont présentés les ensembles $\Gamma_{i,j}$ de sous-graphes de chaque $G_{i,j}$ (j étant ici compris entre 1 et 3). Les types de nucléotides correspondant aux sommets sont indiqués pour chaque graphe et chaque sommet des chemins $W_{B,j}$ et $W_{H,j}$. En-dessous de chaque graphe des ensembles $\Gamma_{i,j}$ sont indiqués les expressions régulières $s_{Bj,s}$ et $s_{Hj,s}$ correspondantes. Enfin, en bas de la figure sont indiquées les ensembles d'expressions régulières de la classe considérée. L'ensemble $r_{B,\mathscr{E}_i}$ contient deux expressions régulières, obtenues par l'union des expressions régulières à l'origine des flèches rouges et bleues respectivement, formant deux classes d'équivalence. L'ensemble $r_{H,\mathscr{E}_i}$ contient une seule expression régulière obtenue par l'union de toutes les expressions régulières $s_{Hj,s}$ de la classe, qui forment une seule classe d'équivalence.

*Deux expressions régulières s_1 et s_2, sur l'alphabet $\mathscr{A}$, appartenant toutes deux à $S_{B,\mathscr{E}_i}$ ou toutes deux à $S_{H,\mathscr{E}_i}$ seront dites **compatibles** si et seulement si s_1 et s_2 contiennent le même nombre de $N = (A|C|G|U)$ et en des positions identiques.*

Par exemple, dans la Figure 5.2, les expressions régulières $s_{B1,1}$ et $s_{B3,1}$ (soulignées en rouge) sont compatibles, comme les expressions régulières $s_{B1,2}$ et $s_{B2,1}$ (soulignées en bleu).

Pour une classe $\mathscr{E}_i$ donnée, un ensemble d'expressions régulières de $S_{B,\mathscr{E}_i}$ ou de $S_{H,\mathscr{E}_i}$, qui sont compatibles deux à deux (par exemple $s_{B1,1}$ et $s_{B3,1}$ dans la Figure 5.2), forme une *classe d'équivalence*.

A partir de cette dernière notion, on définit finalement comme suit les deux ensembles d'expressions régulières associées à la classe $\mathscr{E}_i$ (exemple en Figure 5.2).

Définition 5.2.8 ***Expressions régulières*** $\mathbf{r_{B,\mathscr{E}_i}}$ ***et*** $\mathbf{r_{H,\mathscr{E}_i}}$

Soient $S_{B,\mathscr{E}_i} = [S_{B1}, S_{B2}, ..., S_{Bn_i}]$ et $S_{H,\mathscr{E}_i} = [S_{H1}, S_{H2}, ..., S_{Hn_i}]$ les ensembles d'expressions régulières de toutes les k-extensions contractées de la classe $\mathscr{E}_i$.

Soient $\mathscr{C}_{B,\mathscr{E}_i}$ et $\mathscr{C}_{H,\mathscr{E}_i}$ les ensembles de classes d'équivalence, contenant respectivement les expressions régulières de $S_{B,\mathscr{E}_i}$ et de $S_{H,\mathscr{E}_i}$.

L'ensemble $\mathbf{r_{B,\mathscr{E}_i}}$ *est constitué d'un ensemble d'expressions régulières tel que chaque expression régulière de cet ensemble soit l'union des expressions régulières d'une classe d'équivalence de $\mathscr{C}_{B,\mathscr{E}_i}$.*

De la même façon, ***L'ensemble*** $\mathbf{r_{H,\mathscr{E}_i}}$ *est constitué d'un ensemble d'expressions régulières tel que chaque expression régulière de cet ensemble soit l'union des expressions régulières d'une classe d'équivalence de $\mathscr{C}_{H,\mathscr{E}_i}$.*

Ainsi, ces expressions régulières permettent de considérer les signatures de séquence induites par le sous-graphe commun maximum à la classe. Le premier ensemble d'expressions $r_{B,\mathscr{E}_i}$ décrit ainsi les séquences du brin de la boucle dans le motif A-minor et le deuxième ensemble d'expressions $r_{H,\mathscr{E}_i}$ décrit les séquences du brin de l'hélice.

Nous rechercherons alors toutes les occurrences de ces expressions régulières, dans les séquences d'ARN des deux jeux de données définis dans la section 5.2.1.

Calcul de sous-graphes communs comme représentants de classe

Pour tester la capacité de prédiction des classes à partir de leurs topologies communes, nous allons également considérer deux graphes, obtenus à partir du sous-graphe commun maximum à une classe, que nous avons défini dans le chapitre 2 section 2.5.

Dans cette partie, nous allons définir chacun de ces deux graphes, l'un permettant de représenter dans sa totalité la topologie de contexte commune entre occurrences de motif A-minor d'une même classe, et l'autre permettant de représenter uniquement la topologie commune réduite aux interactions de structures secondaires sans pseudonoeud. Les nucléotides non appariés

feront également partie de cette sous-topologie de contexte. Nous définirons ces notions plus formellement dans les graphes que l'on utilise.

Puis, nous définirons une distance minimale et une distance maximale entre certains couples de sommets de ces graphes, que l'on utilisera ensuite dans la recherche d'isomorphisme de graphes, présentée à la fin de cette section.

Définitions des sous-graphes communs à rechercher Nous considérons le sous-graphe commun maximum $G_{\mathscr{E}_i}$ à une classe $\mathscr{E}_i$, privé des arcs non canoniques de l'occurrence de motif (voir chapitre 2, définition 2.2.3) que nous noterons alors $\bar{G}_{\mathscr{E}_i}$. Nous voulons rechercher la présence d'un motif A-minor à partir de la topologie de son contexte structural, c'est pourquoi nous devons considérer cette topologie sans les interactions du motif lui-même.

Nous séparons l'ensemble des arcs canoniques de $G_{\mathscr{E}_i}$ en deux ensembles : l'un regroupant les arcs appartenant à une structure secondaire sans pseudonoeuds, qu'on dira alors de courte distance, et l'autre regroupant les arcs correspondant à un pseudonoeud dans la structure secondaire considérée, qu'on dira alors de longue distance.

Nous définissons alors le sous-graphe partiel de $\bar{G}_{\mathscr{E}_i}$, ne contenant que les arcs covalents, les arcs canoniques courte distance et les arcs fictifs (représentant les nucléotides non appariés, voir chapitre 2, section 2.2.8) dans $\bar{G}_{\mathscr{E}_i}$, et ne contenant pas les sommets impliqués uniquement dans des arcs non canoniques ou dans des arcs canoniques longue distance. Nous le noterons $\bar{G}^S_{\mathscr{E}_i}$.

Des exemples de ces deux sous-graphes sont présentés en Figure 5.3.

Définitions de distance minimale et maximale entre sommets des sous-graphes communs à une classe Le sous-graphe commun maximum à une classe n'est pas nécessairement connexe. Cependant, nous pouvons définir, dans les graphes $\bar{G}_{\mathscr{E}_i}$ et $\bar{G}^S_{\mathscr{E}_i}$, des valeurs minimale et maximale de distances autorisées entre sommets appartenant à des composantes connexes différentes. Ces distances sont induites par les distances entre les sommets correspondants dans les k-extensions contractées de la classe. Des exemples sont présentés en Figure 5.3.

Définition 5.2.9 ***Distance minimale et maximale entre sommets de*** $\bar{\mathbf{G}}_{\mathscr{E}_\mathbf{i}}$ ***et de*** $\bar{\mathbf{G}}^{\mathbf{S}}_{\mathscr{E}_\mathbf{i}}$

Soit $\mathscr{E}_i = \{\tilde{G}_{i,1}, \tilde{G}_{i,2}, ..., \tilde{G}_{i,n_i}\}$ *une classe de k-extensions contractées.*

Soit $\bar{G}_{\mathscr{E}_i}$ *et* $\bar{G}^S_{\mathscr{E}_i}$ *les deux sous-graphes communs à la classe* $\mathscr{E}_i$ *tels que définis plus haut.*

On considère tout couple de sommets $\langle u, v\rangle$ *dans* $G_{\mathscr{E}_i}$ *tel que :*

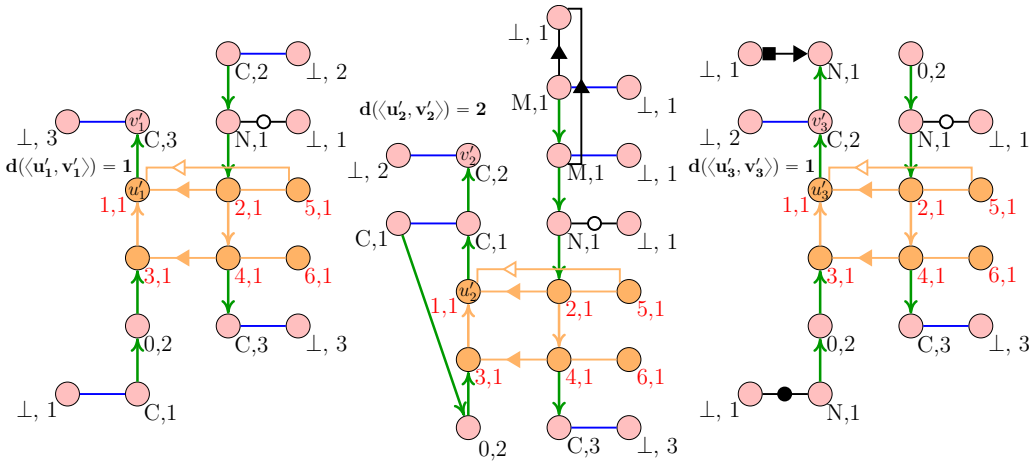
- *le type de* u *et le type de* v *soient différents de* $\perp$*,*
- u *et* v *appartiennent à deux composantes connexes différentes dans* $G_{\mathscr{E}_i}$*,*
- *dans chaque k-extension* $\tilde{G}_{i,j}$ *(où* j *est un entier compris entre 1 et* n_i*) de la classe* $\mathscr{E}_i$*, il y ait un chemin du sommet* u' *équivalent au sens de l'isomorphisme à* u*, jusqu'au sommet* v' *équivalent au sens de l'isomorphisme à* v*, composé uniquement d'arcs covalents, et tel que chaque sommet interne de ce chemin n'appartienne pas au sous-graphe de* $\tilde{G}_{i,j}$ *qui est isomorphe à* $G_{\mathscr{E}_i}$*, selon l'isomorphisme considéré. On note* $d(\langle u', v'\rangle)$ *la longueur de ce chemin.*

Pour tout couple de sommets $\langle u, v\rangle$ *dans* $G_{\mathscr{E}_i}$ *ainsi défini, on appelle* $\mathbf{d_{min}}(\langle \mathbf{u}, \mathbf{v}\rangle)$ *(resp.* $\mathbf{d_{max}}(\langle \mathbf{u}, \mathbf{v}\rangle)$*) la longueur minimum (resp. maximum) de tous les chemins entre les deux sommets équivalents au sens de l'isomorphisme à* u *et à* v *dans les k-extensions contractées de la classe* $\mathscr{E}_i$*. Autrement dit, la valeur* $d_{min}(\langle u, v\rangle)$ *(resp.* $d_{max}(\langle u, v\rangle)$*) est le minimum (resp. maximum) des valeurs* $d(\langle u', v'\rangle)$*, où* u' *et* v' *sont les sommets équivalents au sens de l'isomorphisme à* u *et à* v *respectivement, dans chaque k-extension contractée de la classe.*

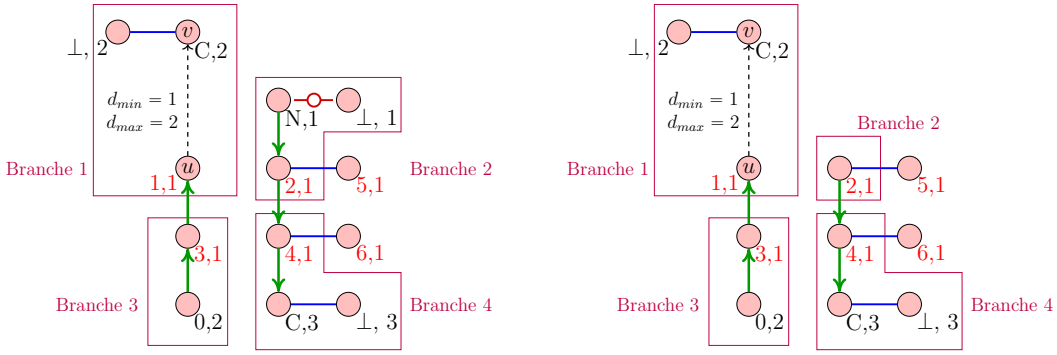
Comme la topologie commune que nous avons calculée est une topologie de contexte *locale* des occurrences de motif A-minor, nous ne devons pas autoriser une distance trop grande entre les composantes connexes des sous-graphes communs qui contiennent des sommets de même branche. De plus, ces valeurs de distances minimale et maximale nous permettent de diminuer le temps d'exécution, lors de la recherche d'isomorphisme des sous-graphes communs à une classe (voir paragraphe suivant).

Dans l'exemple de deux sous-graphes communs à une classe, en Figure 5.3, il existe des chemins de longueur 1 ou de longueur 2 dans les k-extensions contractées de la classe présentée (Figure 5.2), du sommet équivalent au sens de l'isomorphisme à u, au sommet équivalent au sens de l'isomorphisme à v.

Recherche d'isomorphisme de graphes dans les graphes d'ARN. Dans un graphe d'ARN d'un des jeux de données définis dans la section 5.2.1, nous recherchons tous les graphes (π_s, $\pi_{æ}$)-isomorphes (chapitre 2, définition 2.3.1) à $\bar{G}_{\mathscr{E}_i}$ ou à $\bar{G}^S_{\mathscr{E}_i}$, avec une contrainte supplémentaire. Pour qu'un graphe G, qui est (π_s, $\pi_{æ}$)-isomorphe à $\bar{G}_{\mathscr{E}_i}$ ou à $\bar{G}^S_{\mathscr{E}_i}$, soit une occurrence de l'un de nos représentants, il faut en effet que :



(a) k-extensions contractées de la classe



(b) sous-graphes communs de la classe ($\bar{G}_{\mathscr{E}_i}$ à gauche et $\bar{G}^S_{\mathscr{E}_i}$ à droite)

Figure 5.3 – k-extensions contractées de la classe présentée dans la Figure 5.2 (a) et sous-graphes communs à la classe $\bar{G}_{\mathscr{E}_i}$ et $\bar{G}^S_{\mathscr{E}_i}$ (b). Ces deux graphes possèdent trois composantes connexes. Les distances d_{min} et d_{max} entre les sommets u et v sont indiquées, et correspondent aux valeurs minimum et maximum des valeurs $d(\langle u'_1, v'_1\rangle)$, $d(\langle u'_2, v'_2\rangle)$ et $d(\langle u'_3, v'_3\rangle)$ dans les k-extensions contractées du dessus. On cherchera alors des graphes isomorphes à ces graphes pour lesquels la distance entre u et v est d'au plus 2. Les sous-ensembles de sommets, appelés branches, sont indiqués. Dans les sous-graphes communs, les arcs canoniques sont en bleu et les arcs non canoniques sont en rouge.

- pour tout couple de sommets $\langle u', v' \rangle$ dans G, équivalents au sens de l'isomorphisme à un couple de sommets $\langle u, v \rangle$ dans $\bar{G}_{E_i}$ ou $\bar{G}^S_{E_i}$, qui respecte les conditions de la définition 5.2.9, il existe un chemin formé uniquement d'arcs covalents du sommet u' au sommet v' dans G, et

- que la longueur de ce chemin soit comprise entre $d_{min}(\langle u, v \rangle)$ et $d_{max}(\langle u, v \rangle)$.

Par exemple, dans la Figure 5.3, les sommets u et v respectent les conditions définies dans la définition 5.2.9, et les graphes que l'on recherche devront donc respecter les distances minimales et maximales entre u et v.

La Table 5.3 récapitule les caractéristiques des graphes et expressions régulières définies dans la section 5.2.2.

5.2.3 Mesures de prédictibilité

La prédictibilité d'une classe $\mathscr{E}_i$ d'occurrences de motif A-minor correspond donc ici à la capacité à prédire la présence d'un motif A-minor par une approche de reconnaissance de sous-graphes et d'expressions régulières caractérisant cette classe.

La méthodologie que nous utilisons pour mesurer la prédictibilité d'une classe $\mathscr{E}_i$ est la suivante : nous représentons chaque classe par un (ou plusieurs) de ses représentants décrits plus haut (expressions régulières $r_{B,\mathscr{E}_i}$ et $r_{H,\mathscr{E}_i}$ et sous-graphes $\bar{G}_{\mathscr{E}_i}$ et $\bar{G}^S_{\mathscr{E}_i}$), puis nous recherchons dans les deux jeux de données (validation et test) les occurrences de ce(s) représentant(s).

Nous allons présenter dans cette partie les différentes mesures statistiques que l'on utilise pour évaluer les résultats.

La précision ou valeur prédictive positive modifiée (PPVm) définie par classe

La première mesure qui sera calculée est la proportion d'occurrences de motif A-minor parmi toutes les occurrences trouvées. Cette mesure est connue sous le nom de *précision* ou *valeur prédictive positive* (PPV) [34], et calcule le nombre de vrais positifs divisé par le nombre total de positifs. Les vrais positifs dans notre étude sont les occurrences de représentants trouvées qui correspondent à une occurrence de motif A-minor. Comme nous allons le voir dans les paragraphes suivants, le calcul du nombre total de positifs dans notre cas n'est pas trivial. Nous allons donc le décrire précisément, pour chaque type de représentants d'une classe, c'est-à-dire, pour les couples d'ensembles d'expressions régulières d'une part, et pour les sous-graphes communs à la classe d'autre part.

Cas des expressions régulières. Avec les expressions régulières, nous obtenons des occurrences pour le premier ensemble d'expressions régulières $r_{B,\mathscr{E}_i}$, correspondant au brin de la boucle dans le motif A-minor, et des

Nom de l'élément	Description
$G_{\mathscr{E}_i}$	sous-graphe commun maximum aux k-extensions contractées d'une classe $\mathscr{E}_i$
$G_{i,j}$	k-extension non contractée dont les sommets sont étiquetés par le type de nucléotide correspondant
$G_{i,j}^{A_P}$	sous-graphe couvrant d'une k-extension non contractée $G_{i,j}$, ne contenant que les arcs covalents
$W_{B,j}$	chemin le plus long dans le sous-graphe $G_{i,j}^{A_P}$ contenant les sommets 1 et 3 de l'occurrence de motif, et uniquement des sommets des branches 1 et 3
$W_{H,j}$	chemin le plus long dans le sous-graphe $G_{i,j}^{A_P}$ contenant les sommets 2 et 4 de l'occurrence de motif, et uniquement des sommets des branches 2 et 4
$\tilde{G'}_{i,j}$	sous-graphe d'une k-extension contractée $\tilde{G}_{i,j}$ correspondant au (π_s, πæ)-isomorphisme qui définit le sous-graphe commun maximum $G_{\mathscr{E}_i}$
$\tilde{G''}_{i,j}$	graphe dérivé de $\tilde{G'}_{i,j}$, dans lequel chaque sommet possède un poids égal au poids minimum des sommets équivalents au sens de l'isomorphisme dans les autres k-extensions contractées de la classe $\mathscr{E}_i$
$\Gamma_{i,j}$	ensemble des sous-graphes d'une k-extension non contractée $G_{i,j}$ qui sont (π_s, πæ)-isomorphes au graphe décontracté de $\tilde{G''}_{i,j}$
S_{Bj}	ensemble des expressions régulières associé à une k-extension contractée $\tilde{G}_{i,j}$, obtenu à partir du chemin $W_{B,j}$, correspondant au brin de la boucle du motif
S_{Hj}	ensemble des expressions régulières associé à une k-extension contractée $\tilde{G}_{i,j}$, obtenu à partir du chemin $W_{H,j}$, correspondant au brin de l'hélice du motif
$r_{B,\mathscr{E}_i}$	ensemble des expressions régulières associé à une classe $\mathscr{E}_i$, pour le brin de la boucle du motif
$r_{H,\mathscr{E}_i}$	ensemble des expressions régulières associé à une classe $\mathscr{E}_i$, pour le brin de l'hélice du motif
$\bar{G}_{\mathscr{E}_i}$	sous-graphe partiel du graphe $G_{\mathscr{E}_i}$ possédant tous les sommets et tous les arcs de $G_{\mathscr{E}_i}$ sauf les arcs non canoniques de l'occurrence de motif (voir chapitre 2, définition 2.2.3)
$\bar{G}^S_{\mathscr{E}_i}$	sous-graphe partiel du graphe $G_{\mathscr{E}_i}$ possédant tous les sommets et tous les arcs de $G_{\mathscr{E}_i}$ sauf les arcs non canoniques de l'occurrence de motif et sauf les arcs non canoniques et canoniques à longue distance

Table 5.3 – Récapitulatif des graphes et expressions régulières définis pour construire les représentants de classe

occurrences pour le deuxième ensemble d'expressions régulières $r_{H,\mathscr{E}_i}$, correspondant au brin de l'hélice dans le motif A-minor.

Pour définir le nombre total d'occurrences positives, nous devons donc prendre en compte les deux ensembles d'occurrences. Pour cela, nous définissons une PPV modifiée, notée PPVm :

$$PPVm(\mathscr{E}_i) = \frac{\#occurrences\ de\ motif\ A{-}minor\ trouvés}{Min(\#occurrences\ boucle, \#occurrences\ hélice)} \quad (5.1)$$

où $\mathscr{E}_i$ est la classe considérée, et $boucle$ et $hélice$ sont les ensembles d'occurrences des expressions régulières $r_{B,\mathscr{E}_i}$ et $r_{H,\mathscr{E}_i}$, associées à $\mathscr{E}_i$.

Le nombre de vrais positifs dans cette mesure correspond au nombre de paires d'occurrences formant un motif A-minor, et le nombre total de positifs correspond au nombre maximum de motifs A-minor qu'on pourrait former avec les paires d'occurrences trouvées. Autrement dit, le nombre total de positifs correspond au nombre d'occurrences minimum entre les occurrences de $r_{B,\mathscr{E}_i}$ et les occurrences de $r_{H,\mathscr{E}_i}$. Un exemple de calcul de PPVm est présenté en Figure 5.4.

Plus la PPVm est élevée, plus la proportion de vrais positifs est élevée et donc plus la classe aura une forte capacité de prédiction.

Cas des sous-graphes communs. Pour les graphes représentants $\bar{G}_{\mathscr{E}_i}$ et $\bar{G}^S_{\mathscr{E}_i}$, nous recherchons donc tous les sous-graphes partiels des graphes d'ARN, qui respectent l'isomorphisme décrit à la fin de la section 5.2.2.

Rappelons que les sommets des k-extensions, et donc également les sommets des graphes représentants peuvent être couverts par des sous-ensembles, appelés branches, au nombre de 6 pour le motif A-minor (voir section 2.2.2 du chapitre 2, et des exemples en Figure 5.3).

Il est à noter que les graphes $\bar{G}_{\mathscr{E}_i}$ et $\bar{G}^S_{\mathscr{E}_i}$ ne contiennent souvent aucun arc entre les branches 1 et 3 d'une part et les branches 2 et 4 d'autre part, car le sous-graphe commun maximum dont ils sont issus ne contient que les arcs de l'occurrence de motif entre les branches 1 et 3 et les branches 2 et 4. C'est par exemple le cas des graphes présentés en Figure 5.3.

Dans ce cas, on recherche séparément d'une part, des sous-graphes isomorphes au sous-graphe de $\bar{G}_{\mathscr{E}_i}$ induit par les sommets des branches 1 et 3, et d'autre part des sous-graphes isomorphes au sous-graphe de $\bar{G}_{\mathscr{E}_i}$ induit par les sommets des branches 2 et 4. Le même raisonnement s'applique au graphe $\bar{G}^S_{\mathscr{E}_i}$.

Ici, de la même façon que pour les expressions régulières, on obtient donc deux ensembles d'occurrences pour un graphe représentant donné : l'un correspondant au brin de la boucle du motif A-minor (comme les expressions régulières $r_{B,\mathscr{E}_i}$), l'autre correspondant au brin de l'hélice du motif A-minor (comme les expressions régulières $r_{H,\mathscr{E}_i}$).

Si au contraire, des arcs existent entre les branches 1 et 3 et les branches 2 et 4 des graphes représentants, alors on recherche des sous-graphes (π_s, πæ)-isomorphes au graphe $\bar{G}_{\mathscr{E}_i}$ ou au graphe $\bar{G}^S_{\mathscr{E}_i}$, et on obtient un seul ensemble d'occurrences.

Dans le premier cas, on peut donc utiliser le même calcul de PPVm, et dans le deuxième cas, on calcule le nombre d'occurrences formant un motif A-minor divisé par le nombre total d'occurrences trouvées.

Cette mesure ne prend donc pas en compte le nombre de combinaisons possibles de toutes les paires d'occurrences d'un ensemble avec toutes les paires d'occurrences de l'autre ensemble. Cela nous permet de mettre autant d'importance sur les faux positifs que sur les vrais positifs, puisque chaque paire d'occurrences va compter pour un dans notre mesure.

Il est cependant à noter que cette mesure ne permet pas de distinguer deux cas différents : le cas où l'un des ensembles possède un nombre beaucoup plus élevé d'occurrences que l'autre, par rapport au cas où les deux ensembles possèdent un nombre équivalent et peu élevé d'occurrences. Mais, sur les jeux de données que nous avons étudiés, ces cas n'arrivent que pour des classes qui ont une faible PPVm. On peut de plus noter que l'on obtient des résultats équivalents en considérant une PPV par brin, calculant d'une part la proportion de vrais positifs du brin de la boucle, et d'autre part la proportion de vrais positifs du brin de l'hélice.

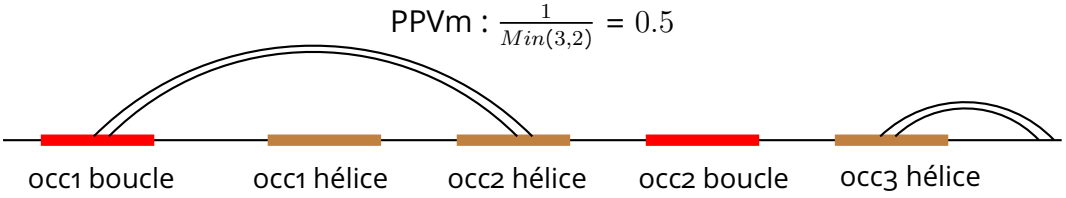


Figure 5.4 – Exemple de calcul de la PPVm pour un cas où deux occurrences sont trouvées pour le premier ensemble que l'on appelle boucle, et trois occurrences pour le deuxième que l'on appelle hélice. La ligne horizontale représente une séquence primaire d'ARN. Les lignes noires courbées représentent des interactions de motif A-minor, et les boîtes rouges et marron représentent les occurrences trouvées. Les deux extrémités de l'occurrence du motif A-minor de gauche sont trouvées (occ1 boucle, occ2 hélice) et nous avons donc une occurrence de motif A-minor trouvée. Une seule des deux extrémités de l'occurrence du motif A-minor de droite est trouvée (occ3 hélice). On ne considère donc pas cette occurrence comme un vrai positif.

Spécificité des représentants de chaque classe

Nous nous intéressons également à l'ensemble des vrais positifs trouvés pour chaque classe $\mathscr{E}_i$. Ces vrais positifs peuvent être des occurrences de motif A-minor de la classe $\mathscr{E}_i$, ou bien des occurrences de motif A-minor d'autres classes.

Nous calculons la proportion d'occurrences de motif A-minor appartenant à la classe $\mathscr{E}_i$ par rapport à l'ensemble des vrais positifs pour la classe $\mathscr{E}_i$, pour un représentant donné. Nous nommerons cette mesure la *spécificité* du représentant pour la classe donnée.

Plus cette proportion est élevée, mieux le représentant en question permet de caractériser la classe.

Sensibilité sur l'ensemble des classes

Nous calculons également la sensibilité des résultats pour chaque représentant ou ensemble de représentants. La sensibilité nous indique la probabilité de prédire la présence d'un motif A-minor en une certaine position, sachant qu'il y en a effectivement un. D'une manière générale, cette probabilité est égale au nombre de vrais positifs, divisé par la somme du nombre de vrais positifs et du nombre de faux négatifs [34]. Dans notre étude, il s'agit donc du nombre d'occurrences de motif A-minor trouvées par rapport au nombre total d'occurrences de motif A-minor se trouvant dans le jeu de données considéré.

Combinaison de représentants

Pour chaque classe $\mathscr{E}_i$, nous recherchons chaque représentant indépendamment, et nous recherchons également des occurrences dont les séquences appartiennent aux langages définis par les expressions régulières $r_{B,\mathscr{E}_i}$ et $r_{H,\mathscr{E}_i}$, et dont le graphe est (π_s, πæ)-isomorphe à l'un ou l'autre des graphes représentants $\bar{G}_{\mathscr{E}_i}$ et $\bar{G}^S_{\mathscr{E}_i}$.

La prédiction peut ainsi se baser sur différentes informations (que nous appellerons par la suite *catégorie de prédictibilité*) :

- *Séquence* : on recherche uniquement les expressions régulières $r_{B,\mathscr{E}_i}$ et $r_{H,\mathscr{E}_i}$
- *Toutes interactions* : on recherche uniquement le graphe représentant $\bar{G}_{\mathscr{E}_i}$
- *Séquence - Interactions canoniques courte distance* : on recherche les expressions régulières $r_{B,\mathscr{E}_i}$ et $r_{H,\mathscr{E}_i}$ et le graphe représentant $\bar{G}^S_{\mathscr{E}_i}$
- *Séquence - Toutes interactions* : on recherche les expressions régulières $r_{B,\mathscr{E}_i}$ et $r_{H,\mathscr{E}_i}$ et le graphe représentant $\bar{G}_{\mathscr{E}_i}$

5.3 Résultats de prédictibilité

Dans cette section 5.3, nous présentons tous les résultats de mesures de prédictibilité à partir des classes de la classification à 4 branches et sur les jeux de données définis en section 5.2, ainsi que les conclusions que l'on peut en tirer.

5.3.1 Résultats de PPVm sur les deux jeux de données

Nous allons ici décrire les résultats de PPVm obtenus avec les deux jeux de données. Nous commencerons par des observations générales avant de décrire les résultats obtenus avec chaque catégorie de prédictibilité.

Nous présentons en Table 5.4 les valeurs moyennes de PPVm obtenues pour chaque catégorie de prédictibilité, sur le jeu de données de validation, obtenues pour l'ensemble des classes de plus de 3 occurrences de motif A-minor de la classification à 4 branches.

Catégories de prédictiblité	PPVm moyenne
Séquence	0,20
Toutes interactions	0,38
Séquence - Interactions canoniques courte distance	0,36
Séquence - Toutes interactions	0,63

Table 5.4 – Moyenne des valeurs de PPVm, pour chaque catégorie de prédictibilité, sur le jeu de données de validation, pour l'ensemble des classes contenant 3 occurrences ou plus de la classification à 4 branches.

Pour pouvoir comparer les valeurs obtenues sur les deux jeux de données, nous avons par ailleurs considéré uniquement un sous-ensemble de classes. Ce sous-ensemble comprend chaque classe, pour laquelle il existe au moins une occurrence de motif A-minor du jeu de test, partageant une sous-structure 3D locale similaire avec au moins une occurrence de motif A-minor de la classe, c'est-à-dire pour laquelle la RMSD entre une occurrence de motif A-minor du jeu de test et une des occurrences de motif A-minor de la classe est inférieure à 2,5Å. Les valeurs sont présentées en Table 5.5.

Catégorie de prédictibilité	PPVm moyenne jeu de validation	PPVm moyenne jeu de test
Séquence	0,14	0,03
Toutes interactions	0,31	0,23
Sequence - Interactions canoniques courte distance	0,27	0,12
Sequence - Toutes interactions	0,56	0,38

Table 5.5 – Moyenne des valeurs de PPVm, pour chaque catégorie de prédictibilité, avec le jeu de données de validation et avec le jeu de données de test, pour un sous-ensemble de classes comprenant chaque classe pour laquelle il existe au moins une occurrence de motif A-minor du jeu de test, partageant une sous-structure 3D locale similaire avec au moins une occurrence de motif A-minor de la classe.

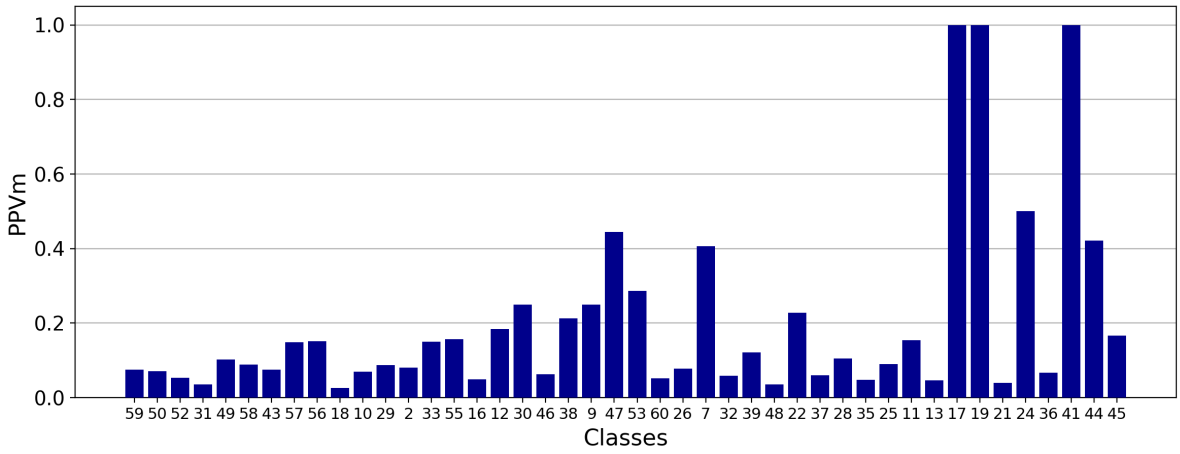
Pour les deux jeux de données, les valeurs les plus élevées de PPVm sont obtenues avec la combinaison de la séquence et de toutes les interactions (0.63 pour le jeu de validation sur toutes les classes (Table 5.4), 0,56 pour le jeu de validation sur le sous-ensemble de classes, et 0,38 pour le jeu de test sur le sous-ensemble de classes (Table 5.5). C'est un résultat cohérent étant donné que cette catégorie est celle utilisant le plus d'informations (séquence et topologie).

On remarque également que les valeurs de PPVm sont toujours inférieures pour le jeu de test par rapport au jeu de validation, quelle que soit la catégorie de prédictibilité. Cependant, il est à noter que les valeurs de PPVm sur les deux jeux de données ne peuvent être comparées directement, étant donné que les jeux de données ont des propriétés et des tailles différentes.

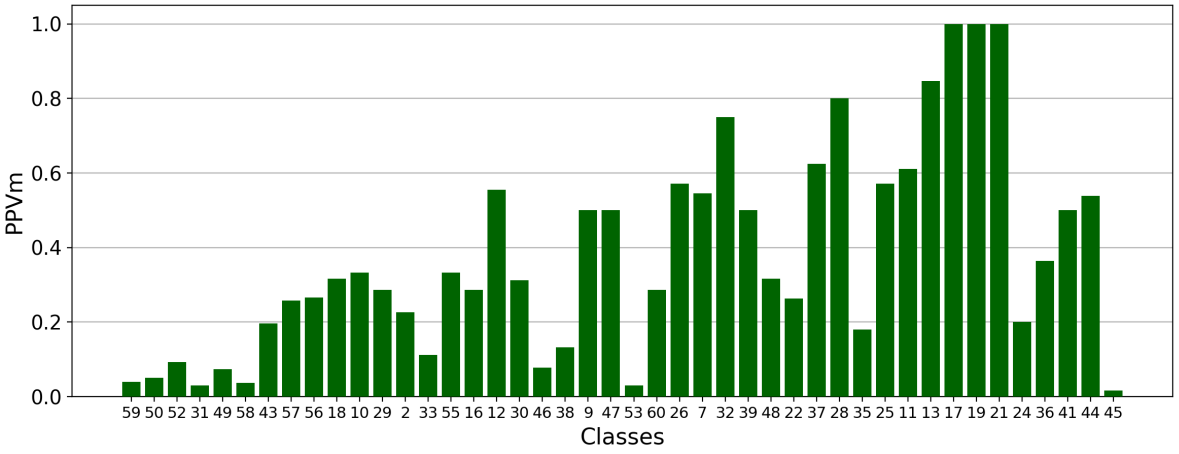
Les valeurs de PPVm pour chaque classe sont présentées en Figures 5.5 pour le jeu de validation sur toutes les classes et en Figure 5.6 pour la comparaison des deux jeux de données. Sur la Figure 5.6, les valeurs de PPVm sur le jeu de test y sont représentées par les barres bleues. Les barres bleues hachurées correspondent aux valeurs de PPVm égales à 0, mais pour lesquelles aucune occurrence (ni vrai positif, ni faux positif) n'a été trouvée. On les différencie ainsi des valeurs de PPVm égale à 0, pour lesquelles aucun vrai positif n'a été trouvé mais pour lesquelles des faux positifs ont été trouvés.

Prédictibilité à partir de la séquence seule

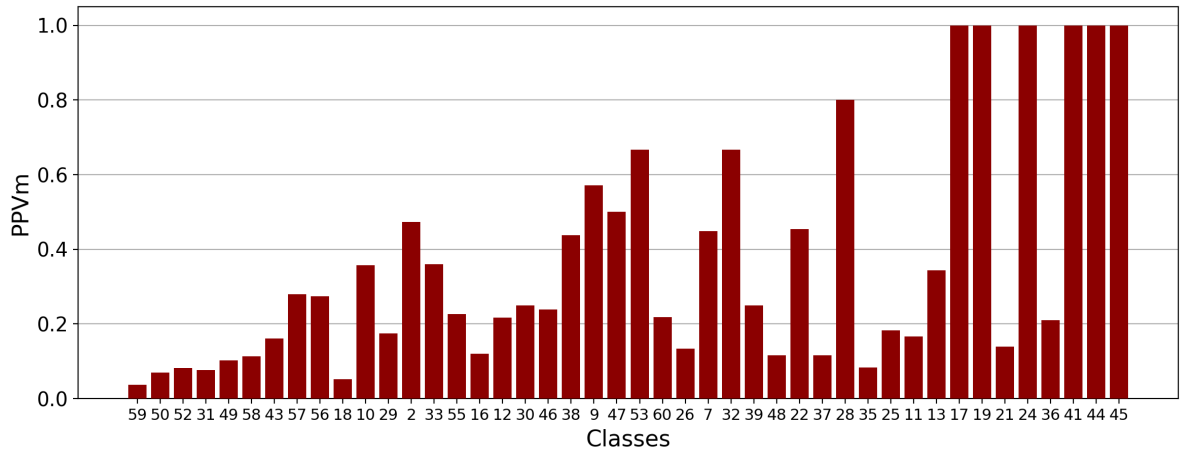
Avec les expressions régulières uniquement (catégorie séquence), quelques classes ont une PPVm égale à 1 pour le jeu de validation (classes numéro 17, 19,



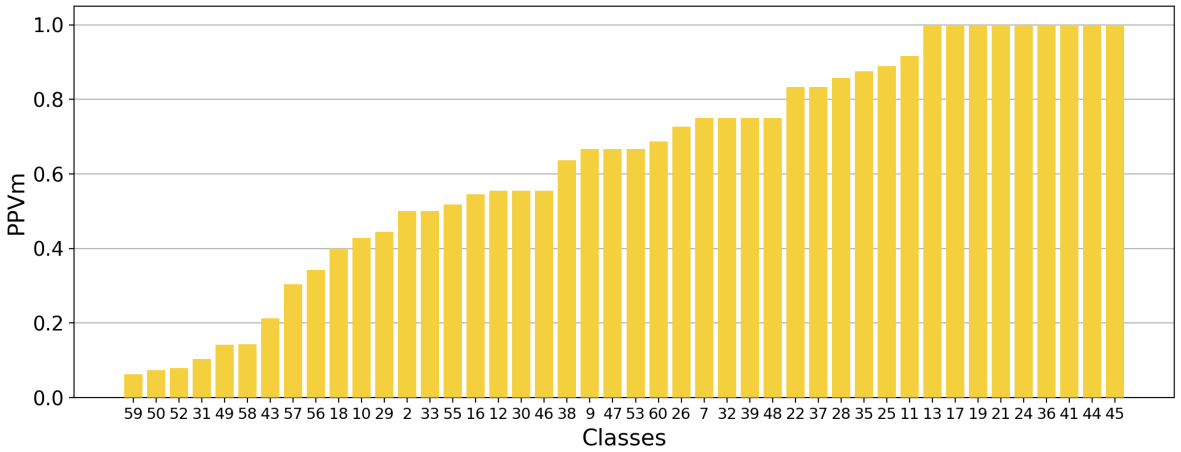
(a) Séquence



(b) Toutes interactions



(c) Séquence - Interactions canoniques courte distance



(d) Séquence - Toutes interactions

Figure 5.5 – Résultats de PPVm sur le jeu de données de validation, pour chaque catégorie de prédictibilité et chaque classe de la classification à 4 branches, représentée par leur numéro dans la classification. Les classes sont ordonnées par valeur de PPVm croissante sur la catégorie Séquence - Toutes interactions.

41 dans la Figure 5.5a). Ce sont de petites classes de 3 occurrences de motif A-minor chacune. Elles possèdent des séquences très similaires (pas plus d'une position où plusieurs types de nucléotides sont autorisés) et des topologies de contexte également très similaires (les sous-graphes communs maximum $\bar{G}_{\mathscr{E}_i}$ ont entre 10 et 17 arcs non covalents pondérés par le poids des sommets, pour une moyenne de 9,36 arcs non covalents sur toutes les classes). Le sous-graphe commun maximum et les expressions régulières de la classe 17 sont présentés en Figure 5.7.

Ces classes regroupent en fait plusieurs annotations de la même molécule dans le même organisme. Nous pouvons également noter que ces classes ne font pas partie du sous-ensemble de classes étudiées avec le jeu de test, ce qui signifie qu'aucune occurrence de motif A-minor du jeu de test ne possède une sous-structure 3D locale similaire aux occurrences de motif de ces 3 classes. Ces classes possèdent donc des caractéristiques structurales bien particulières.

Toutes les autres classes ont une PPVm inférieure à 0,5 avec le jeu de validation, et inférieure à 0,2 avec le jeu de test (Figure 5.6a).

Cette observation démontre que la séquence seule n'est pas suffisante pour prédire la position d'un motif A-minor.

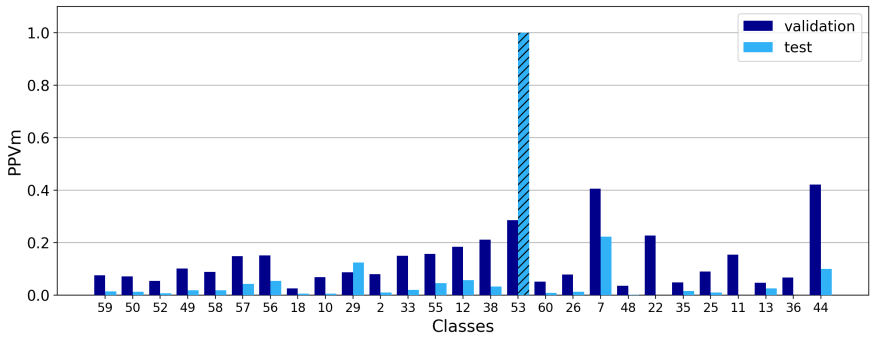
Prédictibilité à partir de la topologie seule

A partir des interactions du sous-graphe commun maximum (catégorie toutes interactions), 4 classes ont une PPVm supérieure à 0,75 sur le jeu de validation (numérotées 13, 21, 28, 32) (voir Figure 5.5b). Avec le jeu de test (Figure 5.6b), la classe 13 possède également une PPVm supérieure à 0,7, et les 3 autres classes ne font pas partie du sous-ensemble de classes considérées pour le jeu de test. Une autre classe (la classe 25) possède une PPVm supérieure à 0,7 sur le jeu de test (la PPVm de cette classe sur le jeu de validation est d'un peu plus de 0,55).

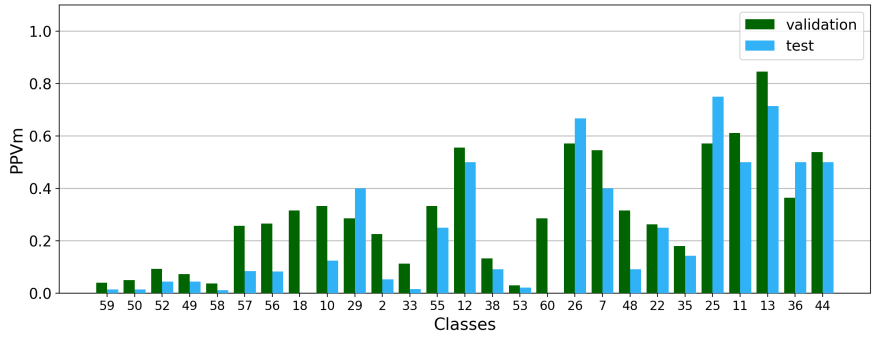
Le sous-graphe commun maximum $\bar{G}_{\mathscr{E}_i}$ de ces classes est conséquent, avec plus de 10 arcs non covalents pondérés par le poids des sommets (sur une moyenne de 9,36 sur toutes les classes) et une forte proportion d'arcs non canoniques (plus de la moitié). Ainsi, les topologies de contexte structural au sein de ces classes sont très similaires, ce qui explique qu'on les trouve préférentiellement lorsqu'il y a une occurrence de motif A-minor. Les sous-graphes communs des classes 21 et 32 ont également la particularité de posséder au moins un arc non canonique entre les nucléotides du motif, et qui n'appartient pas au motif (voir le sous-graphe commun de la classe 21 en Figure 5.7). Cela peut grandement diminuer le nombre de faux positifs, étant donné le lien entre les deux brins du motif.

Séquence vs toutes interactions

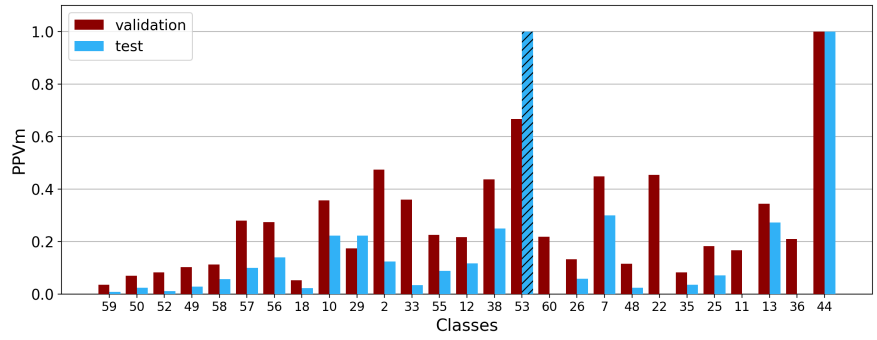
Un petit nombre de classes ont une meilleure PPVm avec la séquence seule qu'avec toutes les interactions. Cela concerne 11 classes sur le jeu de validation (Figure 5.5, classes 24, 31, 33, 38, 41, 45, 49, 50, 53, 58, 59), et 2 classes (classes 58 et 60) sur le jeu de test (Figure 5.6). La classe 58 est ainsi la seule à vérifier cette propriété sur les deux jeux de données. Les occurrences de motif A-minor au sein de ces classes se trouvent principalement dans un ou deux des règnes du Vivant : un seul règne pour 5 classes, deux règnes (Bactéries et Eucaryotes) pour 5 classes. Il y a une exception : les occurrences de motif de la classe 59 se retrouvent dans les trois règnes du Vivant. Notons cependant que, pour cette classe, la PPVm avec la séquence sur le jeu de validation est égal à 0,08 (0,01 sur le jeu de test). Concernant les caractéristiques topologiques, les sous-graphes communs maximums de ces 11 classes possèdent une faible proportion d'arcs non canoniques, et pour 7 de ces classes, le sous-graphe commun maximum est très réduit, contenant moins de 8 arcs canoniques et non canoniques. A part deux classes (24 et 41) qui ont des séquences très similaires, les 9 autres classes ne montrent que peu de similarité de séquence en réalité. Les valeurs de PPVm obtenues avec la séquence pour ces classes



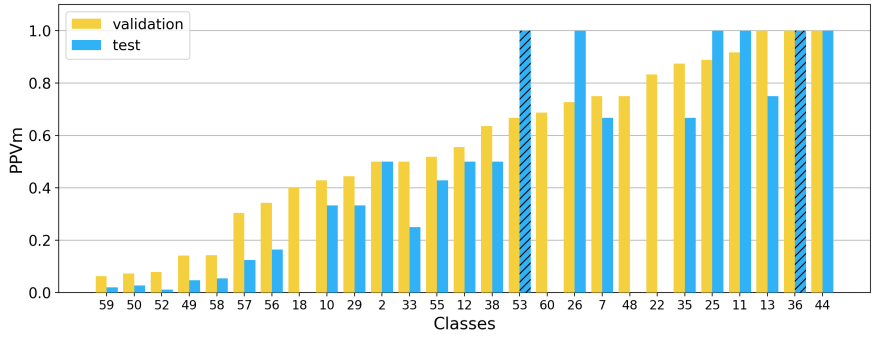
(a) Séquence



(b) Toutes interactions



(c) Séquence - Interactions canoniques courte distance



(d) Séquence - Toutes interactions

Figure 5.6 – Résultats de PPVm pour chaque catégorie de prédictibilité, sur le jeu de données de test (en bleu sur chaque figure), et sur le jeu de données de validation. Les barres en bleu hachurées correspondant aux cas où aucune occurrence n'est trouvée, pour les différencier du cas où aucune occurrence correspondant à un motif A-minor n'est trouvée mais où des faux positifs sont trouvés. Un sous-ensemble de classes est considéré : chaque classe pour laquelle il existe au moins une occurrence de motif A-minor du jeu de test, partageant une sous-structure 3D locale similaire avec une occurrence de motif A-minor de la classe.

sont très faibles (inférieur à 0,27 avec le jeu de validation et encore inférieur avec le jeu de test), bien qu'elles soient meilleures que les valeurs de PPVm obtenues avec toutes les interactions. Il s'agit donc peut-être davantage d'un biais sur les données.

Davantage de classes ont au contraire une meilleure PPVm avec les interactions seules qu'avec la séquence seule (33 classes sur 44 pour le jeu de validation et 25 sur 27 pour le jeu de test). En particulier, le ratio entre la valeur de PPVm pour toutes les interactions et la valeur de PPVm pour la séquence seule est supérieur à 2 sur les deux jeux de données, pour 12 classes (classes 2, 10, 11, 12, 13, 25, 26, 29, 35, 36, 48, 55). Parmi ces 12 classes, 11 contiennent des occurrences de motif A-minor appartenant à des molécules d'au moins deux règnes du Vivant : 6 classes contiennent des occurrences trouvées dans les 3 règnes, et 5 classes contiennent des occurrences trouvées dans 2 règnes sur 3. La classe 36 est particulière car elle contient des occurrences de motif A-minor trouvés chez les Bactéries et chez les chloroplastes. Le sous-graphe commun maximum $\bar{G}_{\mathscr{E}_i}$ de 7 des 12 classes contient plus de 9 interactions (canoniques et non canoniques), pour une moyenne à 9,36 arcs. Les 5 autres classes possèdent 7 ou 8 interactions, dont plus de la moitié sont des interactions non canoniques, et parfois même des interactions assez rares (exemple *trans* Watson Crick/ Hoogsteen dans la classe 2). En revanche, les 12 classes sont toutes caractérisées par des séquences variées avec plus de 30 et parfois des centaines de séquences théoriques possibles pour une classe, c'est-à-dire les séquences possibles que l'on peut obtenir à partir des expressions régulières.

Le nombre de classes ayant une meilleure PPVm avec les interactions plutôt qu'avec la séquence est ainsi bien plus élevé que le nombre de classes ayant une meilleure PPVm avec la séquence plutôt qu'avec les interactions. Cela semble ainsi démontrer que la topologie du contexte structural est mieux conservée et donc plus informative que la séquence associée.

Prédictibilité de séquence et de topologie

Avec la séquence et les interactions canoniques courte distance, on remarque une augmentation de la PPVm, par rapport à la séquence seule, pour la majorité des classes, sur les deux jeux de données. C'est en particulier le cas pour plusieurs classes (classes 28, 24, 44, 45, 53 pour le jeu de validation en Figures 5.5c, et 44 et 53 pour le jeu de test en Figures 5.6c). Les motifs A-minor des classes 28, 45 et 53 partagent en fait une particularité dans leur contexte structural qui peut l'expliquer : des interactions canoniques sont présentes entre les deux brins du motif A-minor (entre les branches 1 et 3 et les branches 2 et 4). Le sous-graphe commun et les expressions régulières de la classe 53 sont présentés en Figure 5.7, pour exemple.

Les classes numéro 24 et 44 ont, quant à elles, un sous-graphe commun maximum conséquent (plus de 10 arcs non covalents pondérés par le poids de sommets), comprenant une grande proportion d'arcs canoniques (5 et 8). Leurs expressions régulières induisent également peu de séquences possibles.

Nous pouvons noter cependant que la séquence et les interactions canoniques courte distance ne donnent globalement pas de meilleurs résultats de PPVm que toutes les interactions (voir Figure 5.5b et Figure 5.5c). Pour le jeu de test, les résultats sont même globalement moins bons (voir Figure 5.6b et Figure 5.6c). C'est une observation qui tend à montrer l'importance des interactions non canoniques.

Avec la séquence et toutes les interactions, une grande partie des classes ont une PPVm supérieure à 0,5 sur le jeu de validation (32 sur 44, voir Figure 5.5d). C'est également le cas d'un peu moins de la moitié des classes du jeu de test (12 sur 27, voir Figure 5.6). Dans certaines classes, le sous-graphe commun maximum contient des arcs non covalents entre les sommets du motif, autres que ceux du motif, comme par exemple, les classes 30 et 35. Pour d'autres classes, le nombre d'arcs non covalents dans le sous-graphe commun maximum est important (par exemple, les classes 13 et 26), ou bien les expressions régulières induisent peu de séquences possibles, au moins pour l'un des ensembles d'expressions (par exemple, les classes 7, 9 ou 22). L'exemple de la classe 22 est présenté en Figure 5.7.

Ainsi, pour un grand nombre des classes considérées, les informations de séquence et de topologie combinées semblent constituer un bon signal de prédiction.

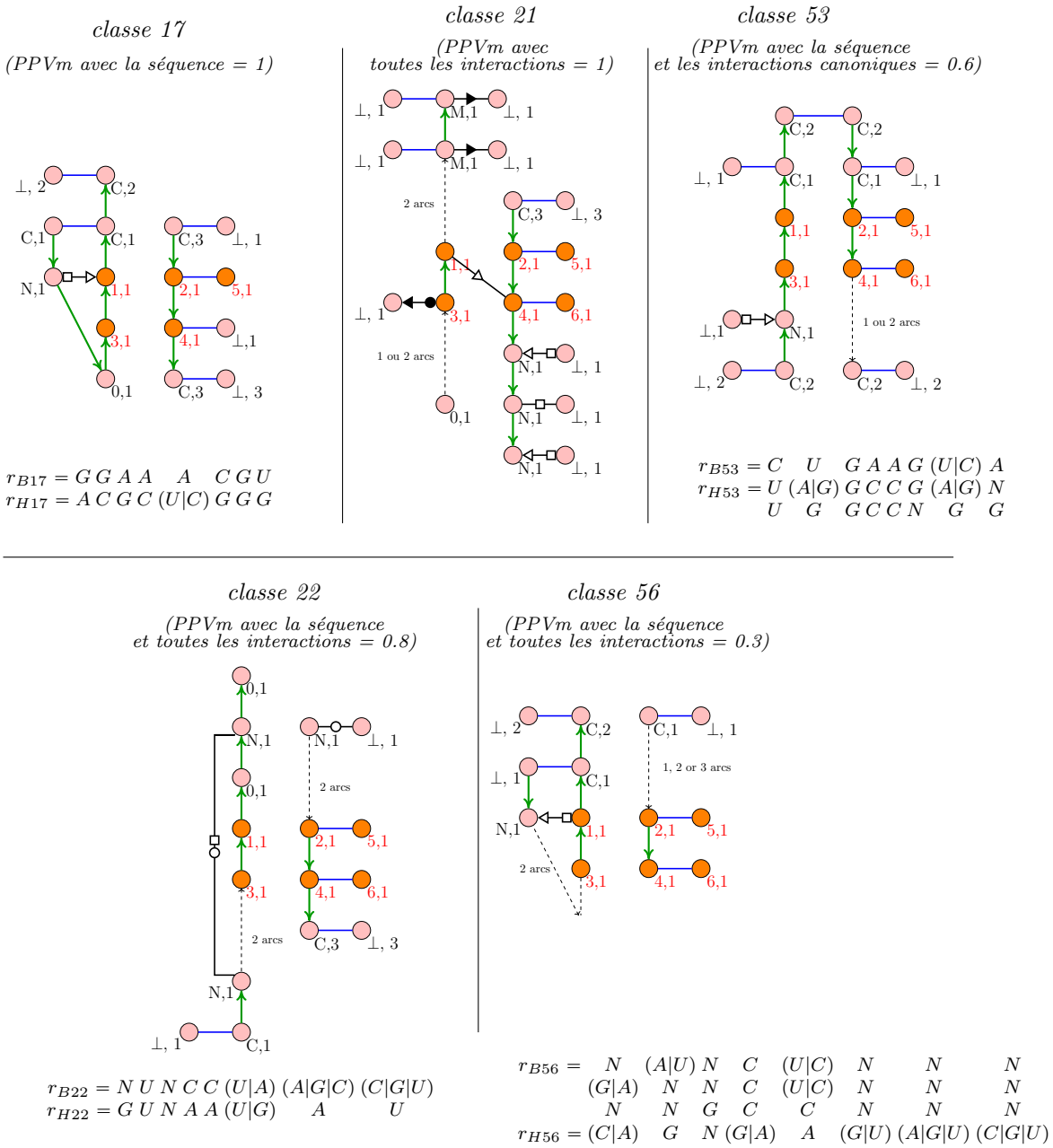
Classes possédant une faible PPVm quelque soit la catégorie de prédictibilité

Sur le jeu de validation, une dizaine de classes ont une PPVm très faible (inférieure à 0,4) quelle que soit la catégorie de prédictibilité étudiée (Figure 5.5). On retrouve 8 de ces mêmes classes dans le jeu de test, ainsi que 3 autres dont la PPVm est égale à 0 sur le jeu de test et supérieure à 0,6 sur le jeu de validation (classes 60, 48 et 22). Notons cependant pour ce dernier point que le nombre total de positifs est faible pour ces 3 classes sur le jeu de test, comme sur le jeu de validation. Ces 3 classes possèdent donc des représentants retrouvés peu fréquemment dans les séquences et les graphes d'ARN.

Parmi les classes ayant une faible PPVm sur un jeu de données au moins, on retrouve la quasi-totalité des classes comportant des occurrences non homologues (classes 43, 49, 50, 52, 56, 58, 59, voir Figure 5.4d). L'exemple de la classe 56 est présenté en Figure 5.7. Les classes 48, 53, 61, également composées d'occurrences non homologues (voir chapitre 4, section 4.2), ont de meilleures valeurs de PPVm mais sont constituées d'une seule occurrence non homologue aux autres occurrences, ce qui explique ce résultat.

Toutes les classes ayant une faible valeur de PPVm quelle que soit la catégorie de prédictibilité possèdent donc une topologie commune de trop petite taille, et des séquences trop variées pour être prédites.

Pour les classes d'occurrences non homologues en particulier, ce résultat est cohérent avec les analyses des chapitres précédents, dans lesquelles nous



$r_{B17} = G\ G\ A\ A\quad A\quad C\ G\ U$
$r_{H17} = A\ C\ G\ C\ (U|C)\ G\ G\ G$

$r_{B53} = C\quad U\quad G\ A\ A\ G\ (U|C)\ A$
$r_{H53} = U\ (A|G)\ G\ C\ C\ G\ (A|G)\ N$
$\quad\quad U\quad G\quad G\ C\ C\ N\quad G\quad G$

$r_{B22} = N\ U\ N\ C\ C\ (U|A)\ (A|G|C)\ (C|G|U)$
$r_{H22} = G\ U\ N\ A\ A\ (U|G)\quad A\quad\quad U$

$r_{B56} = N\quad (A|U)\ N\quad C\quad (U|C)\quad N\quad\quad N\quad\quad N$
$\quad\quad (G|A)\quad N\quad N\quad C\quad (U|C)\quad N\quad\quad N\quad\quad N$
$\quad\quad N\quad\quad N\quad G\quad C\quad\quad C\quad\quad N\quad\quad N\quad\quad N$
$r_{H56} = (C|A)\quad G\quad N\ (G|A)\quad A\quad (G|U)\ (A|G|U)\ (C|G|U)$

Figure 5.7 – Exemples de classes possédant des particularités, et possédant des valeurs de PPVm élevées pour certaines catégories. Les PPVm indiquées sont celles du jeu de validation. Pour chaque classe $\mathscr{E}_i$ sont représentés les sous-graphes communs $\bar{G}_{\mathscr{E}_i}$ et les expressions régulières $r_{B,\mathscr{E}_i}$ et $r_{H,\mathscr{E}_i}$ lorsqu'elles sont utilisées (pour toutes les catégories, sauf la catégorie toutes interactions).

avions remarqué que les occurrences de motif au sein de ces classes avaient des contextes 3D similaires mais des topologies de contexte moins similaires.

Conclusion

L'étude des valeurs de PPVm sur le jeu de validation révèle donc qu'utiliser la séquence seule semble ne pas être suffisant pour caractériser une occurrence de motif A-minor et prédire sa position. La topologie commune, quant à elle, paraît suffisante dans des cas particuliers : lorsque des arcs, autres que ceux du motif, sont présents entre les branches 1 et 3 et les branches 2 et 4, ou bien lorsque la proportion d'arcs non canoniques est importante. D'autres contextes particuliers, comme la présence d'arcs canoniques entre les branches 1 et 3 et les branches 2 et 4, donnent de bons résultats de PPVm, pour la catégorie comprenant la séquence et la topologie contenant uniquement les arcs canoniques de courte distance. Pour finir, la séquence et la topologie dans sa totalité donnent un bon signal pour la plupart des classes.

Nous avons également observé que la topologie dans sa totalité donne des résultats équivalents ou même parfois de meilleurs résultats que la séquence et la topologie réduite aux interactions canoniques courte distance, ce qui montre l'importance des interactions non canoniques dans le contexte topologique du motif A-minor.

Bien que les valeurs de PPVm soient globalement moins élevées sur le jeu de test que sur le jeu de validation, les ensembles de résultats sur ces deux jeux de données sont cohérents pour chaque classe étudiée. Ainsi, les résultats restent pertinents sur un jeu de données qui n'a pas été utilisé pour la classification.

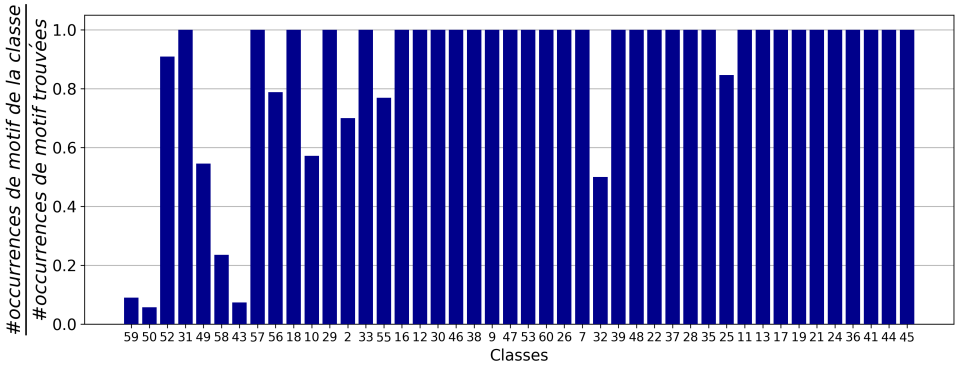
5.3.2 Spécificité des représentants

Comme expliqué dans la section 5.2.3, nous calculons également sur le jeu de validation, pour chaque classe $\mathscr{E}_i$, la proportion d'occurrences de motifs A-minor de la classe $\mathscr{E}_i$ par rapport au nombre total d'occurrences de motifs A-minor effectivement trouvées. Les résultats sont présentés dans la Figure 5.8.

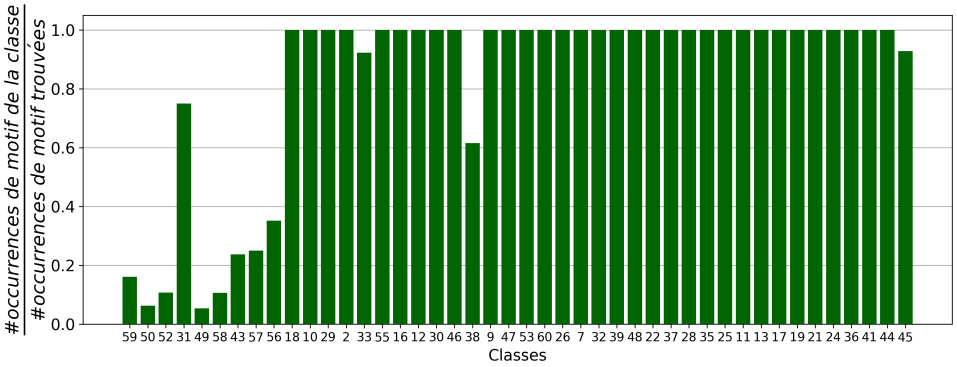
Cela nous indique si les représentants d'une classe permettent de la caractériser, c'est-à-dire si, lorsqu'on trouve une occurrence de motif A-minor à partir de ces représentants, alors on sait que cette occurrence possède une topologie de contexte et/ou une séquence similaires aux occurrences de la classe, et uniquement aux occurrences de cette classe. On peut alors considérer que cette occurrence appartient à la classe en question.

Nous dirons qu'un représentant, ou une combinaison de représentants, est *spécifique* de la classe lorsque la proportion d'occurrences de motif A-minor de la classe par rapport au nombre total d'occurrences de A-minor trouvées est égale à 1.

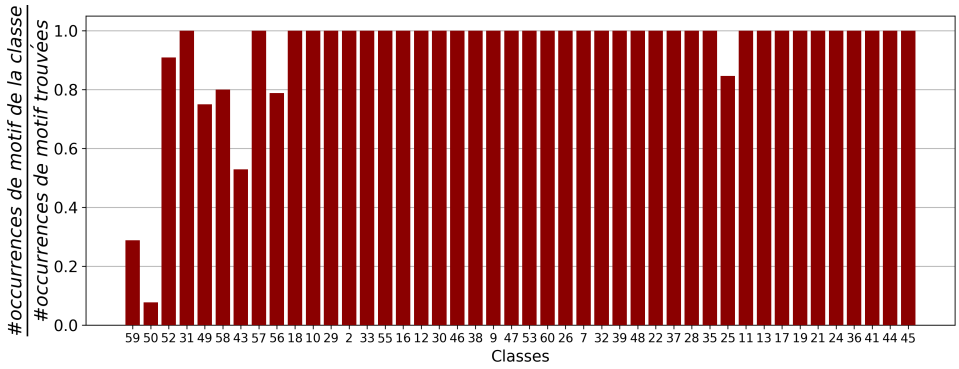
Pour la séquence, pour 31 classes sur 45, les vrais positifs ne correspondent qu'aux occurrences de la classe (voir dans la Figure 5.8a, les classes avec une



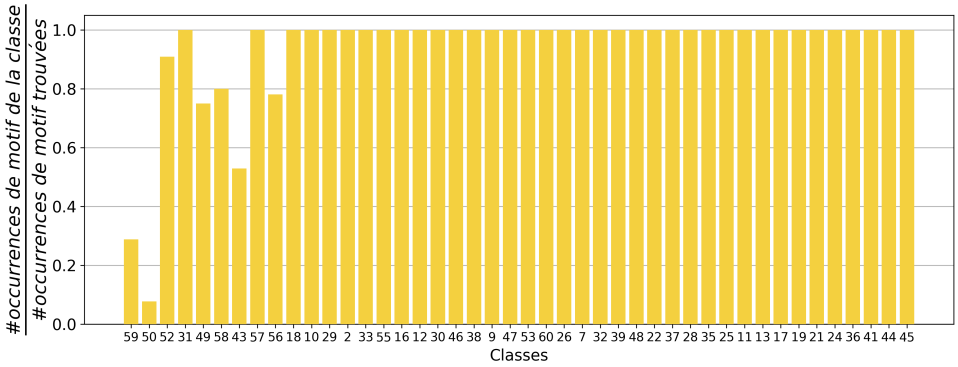
(a) Séquence



(b) Toutes interactions



(c) Séquence - Interactions canoniques courte distance



(d) Séquence - Toutes interactions

Figure 5.8 – Proportion d'occurrences de motif A-minor appartenant à la classe par rapport au nombre total de vrais positifs, dans le jeu de validation, pour chaque classe, pour chaque catégorie de prédictibilité.

proportion égale à 1). A partir des expressions régulières représentantes de la classe, on ne retrouve donc que la position des occurrences de motif A-minor de la classe en question. Les 14 autres classes possèdent des expressions régulières non spécifiques. Parmi elles, on trouve toutes les classes comportant des occurrences non homologues et quelques autres classes.

D'après la Figure 5.8b, 32 classes sur 45 possèdent une topologie de contexte qui leur est spécifique. Les autres classes possèdent donc des sous-graphes communs maximum qui sont inclus dans le sous-graphe commun maximum d'autres classes. C'est le cas des classes composées d'occurrences non homologues, et d'autres classes comme la classe 55 ou la classe 12, qui ont des sous-graphes communs maximum avec peu de sommets et peu d'arcs. Il y a également le cas de la classe 47, dont le sous-graphe commun maximum est inclus dans celui de la classe 44, car ces deux classes sont constituées d'occurrences homologues (voir chapitre 4, section 4.2). Dans la Figure 5.8, on remarque également que certaines classes, comme la classe 52 et la classe 31 possèdent des expressions régulières plus spécifiques que ne l'est leur sous-graphe commun maximum (proportion plus faible dans la Figure 5.8b que dans la Figure 5.8a).

Les deux autres catégories de prédictibilité en Figures 5.8c et 5.8d donnent des proportions plus élevées pour toutes les classes que les deux catégories précédentes. 36 classes sur 45 ont une proportion de 1 pour la catégorie de la séquence et des interactions canoniques courte distance (Figure 5.8c) et 37 classes sur 45 ont une proportion de 1 pour la catégorie de la séquence et de toutes les interactions (Figure 5.8d). Les classes n'ayant pas une proportion de 1, avec la séquence et toutes les interactions, sont les classes composées d'occurrences non homologues.

Ainsi, le représentant constitué de la topologie de contexte commune (sous-graphe commun maximum à la classe) et de la séquence (couple d'ensembles d'expressions régulières) est spécifique, au sens où nous l'avons défini ici, pour chaque classe composée d'occurrences homologues uniquement.

5.3.3 Sensibilité sur l'ensemble des classes

La mesure de sensibilité dans notre cas est donc le nombre d'occurrences de motifs A-minor trouvées avec l'ensembles des classes, par rapport au nombre total d'occurrences de motif A-minor se trouvant dans le jeu de données considéré.

Les résultats de sensibilité sont indiqués dans la Table 5.6, pour chaque catégorie de prédictibilité, en considérant les résultats de toutes les classes confondues.

On y remarque que la grande majorité des motifs A-minor sont trouvés dans le jeu de validation (plus de 90%), pour toutes les catégories de prédictibilité. Toutes les occurrences de motif A-minor appartenant aux classes comportant plus de 3 occurrences sont forcément trouvées, comme ce sont les classes dont nous avons calculé les représentants. Les occurrences de motif A-minor manquants dans ce jeu de données se trouvent dans des classes de

Jeu de données	Catégorie de prédictibilité	Proportion d' occurrences de A-minor trouvées sur le total
Validation	Séquence	$368/377 \approx 0.98$
	Toutes interactions	$353/377 \approx 0.94$
	Séquence - Interactions canoniques courte distance	$342/377 \approx 0.91$
	Séquence - Toutes interactions	$342/377 \approx 0.91$
Test	Séquence	$85/98 \approx 0.87$
	Toutes interactions	$75/98 \approx 0.77$
	Séquence - Interactions canoniques courte distance	$65/98 \approx 0.66$
	Séquence - Toutes interactions	$60/98 \approx 0.61$

Table 5.6 – Mesures de sensibilité pour chaque catégorie de prédictibilité, sur chaque jeu de données (validation et test).

taille 2 ou de taille 1 dans la classification, qui possèdent donc des contextes 3D peu similaires aux autres occurrences de motif.

Dans le jeu de test, plus de 60% des occurrences de motif sont trouvées avec la prédictibilité sur la séquence et la topologie. Cela signifie que plus de 60% des occurrences de motif A-minor de ce jeu de données partagent des topologies et des séquences similaires avec des occurrences de motif d'une des (ou de plusieurs) classes du jeu de validation.

On peut noter également que la dernière catégorie de prédictibilité (avec la séquence et toutes les interactions) possède la plus faible valeur de sensibilité pour les deux jeux de données. C'est en effet la catégorie la plus restrictive, puisque les occurrences de motif A-minor doivent posséder une topologie de contexte similaire et une séquence similaire pour être détectés.

5.4 Conclusion

Dans ce chapitre, nous avons étudié la capacité de prédiction de plusieurs représentants de classes d'occurrences de motif A-minor : les séquences communes, la topologie commune dans sa totalité ainsi que la topologie correspondant à une structure secondaire sans pseudonœud uniquement, et enfin des combinaisons de ces représentants.

Les résultats indiquent que la séquence seule n'est pas suffisante pour prédire la position d'un motif A-minor. Notons à ce sujet que les expressions régulières que nous considérons sont très sommaires. Elles ne prennent pas en compte la probabilité d'apparition d'un nucléotide par exemple. Des modèles plus complexes pour représenter une séquence commune ont déjà été utilisés, comme les réseaux Bayésiens [96], et seraient peut-être intéressants à étudier pour nos classes, bien que la quantité de données que nous avons soit très petite.

La topologie seule, quant à elle, est suffisante pour certaines classes possédant des topologies particulières, et la combinaison des deux donne un bon signal pour la majorité des classes.

D'autre part, la plupart des classes possèdent également des représentants qui leur sont spécifiques, c'est-à-dire que les occurrences de ces représentants ne sont trouvées que dans les contextes des occurrences de motif de la classe. Cela indique que ces représentants permettent de caractériser la classe.

Seule une dizaine de classes restent très peu prédictibles, en particulier les classes composées d'occurrences non homologues, partageant des structures 3D locales similaires, mais des topologies locales moins similaires.

Mais, d'une manière générale, cette étude confirme l'influence de la topologie du contexte local d'un motif A-minor sur sa formation, et définit des représentants de classe d'occurrences de motif A-minor apportant un bon signal pour une approche prédictive.

Conclusion

Au cours de cette thèse, nous avons étudié le contexte structural de motifs longue distance d'ARN, par une approche d'algorithmique de graphes. Le but de cette étude était de déterminer si le contexte structural topologique de motifs longue distance d'ARN, c'est-à-dire le contexte composé des interactions canoniques et non canoniques, est suffisant pour déterminer le contexte 3D, et pourrait alors permettre de faire des avancées dans la prédiction de ces motifs. En effet, la prédiction de ces motifs à longue distance reste difficile par les méthodes algorithmiques actuelles, avant tout en raison de la présence d'interactions non canoniques et de la distance sur la séquence séparant les différentes parties des motifs.

Nous avons tenté de répondre à cette question dans les chapitres 2 et 3.

Dans le chapitre 2, nous avons présenté la méthode que nous avons mise au point pour représenter et comparer les topologies de contexte, ainsi que les contextes 3D de motifs d'ARN. Nous avons modélisé le contexte structural topologique de motifs d'ARN par des graphes. Nous avons ainsi défini la k-extension d'une occurrence de motif, comme un graphe orienté dans lequel les sommets correspondent aux nucléotides et les arcs représentent les interactions covalentes ainsi que les interactions canoniques et non canoniques, apparaissant dans le contexte de taille k de cette occurrence. Une k-extension contient deux ensembles d'arcs : les arcs correspondant aux interactions covalentes, qui sont orientés dans le sens 5'-3' de la séquence primaire, et les arcs correspondant aux interactions canoniques et non canoniques. A chaque arc de ce dernier ensemble est associé un type, en fonction du type d'interactions selon la nomenclature de Leontis–Westhof [62], et lorsqu'il existe un arc (x, y) dans cet ensemble, l'arc (y, x) existe également. Les types de ces deux arcs (x, y) et (y, x) peuvent cependant être différents (exemple *cSH* et *cHS*). Nous avons également défini un deuxième type de graphe, que nous avons appelé k-extension contractée, permettant de représenter les hélices et les boucles de structure secondaire sans interactions non canoniques comme des blocs insécables, et ainsi s'affranchir des différences d'un ou deux nucléotides dans ces éléments, qui n'induisent en général pas de grandes différences dans les structures 3D. Dans le chapitre 3, nous avons montré qu'utiliser les k-extensions contractées permettait de réduire le temps d'exécution des algorithmes et de ne pas perdre d'information par rapport au contexte 3D.

Pour comparer les contextes topologiques de différentes occurrences de

motifs, nous avons recherché le sous-graphe commun à deux k-extensions, maximisant le nombre d'arcs non covalents, par une méthode exacte de résolution du problème MCES [90] et par une méthode heuristique que nous avons développée. Nous avons alors montré dans le chapitre 2 que la méthode heuristique permettait d'obtenir le sous-graphe commun maximum dans la plupart des cas.

Nous avons défini également dans le chapitre 2 le contexte 3D de taille k d'un motif d'ARN, comme les positions dans l'espace d'un ensemble de nucléotides se trouvant dans la k-extension correspondante. Pour comparer ces contextes, nous avons utilisé la RMSD, mesure de la qualité d'un alignement de deux structures 3D.

A partir de ces définitions, dans le chapitre 3, nous avons étudié la corrélation entre les similarités de contexte topologique et les similarités de contexte 3D pour les occurrences de trois motifs d'ARN à longue distance (motif A-minor, motif G, motif trans WC/H), choisis pour leur fréquence dans les structures 3D d'ARN, et l'implication d'interactions non canoniques dans ces motifs. D'une manière générale, nous avons observé que des occurrences de motif possédant des contextes 3D similaires partagent également des contextes topologiques similaires et inversement. Ce propos doit cependant être nuancé. La majorité des occurrences de motif partageant à la fois des contextes 3D similaires et des topologies de contexte similaires sont en fait des occurrences homologues. Ces occurrences sont donc dérivées d'un ancêtre commun, ce qui explique cette similarité, car des molécules homologues partagent souvent des structures similaires. De plus, une petite proportion, mais non négligeable, de paires d'occurrences partagent des contextes 3D similaires mais des topologies de contexte non similaires ou inversement. C'est en particulier le cas pour les occurrences de motif A-minor. Malgré ces deux constats, il est intéressant de remarquer qu'avec la topologie seule, nous retrouvons des similarités existant entre contextes 3D, alors que le contexte 3D contient davantage d'informations, incluant la topologie. Cela semble donc indiquer que la topologie de contexte de ces trois motifs permet de déterminer leur contexte 3D dans la plupart des cas, même si des exceptions existent.

Après cette première conclusion, nous nous sommes intéressés au motif A-minor en particulier dans les chapitres 4 et 5. Parmi les trois motifs étudiés, ce motif présente les moins bons résultats de corrélation entre les similarités de topologie de contexte et les similarités de contexte 3D. Cela peut indiquer que l'influence du contexte structural sur la formation de ce motif est la moins importante pour les trois motifs. Cependant, davantage d'occurrences non homologues de ce motif, par rapport aux autres motifs, possèdent des contextes 3D et/ou des topologies de contexte similaires. Cette raison, ainsi que l'importance du motif A-minor dans le repliement et la fonction de certaines molécules d'ARN [66], nous ont décidés à approfondir l'étude du contexte structural de ce motif en particulier.

Dans le chapitre 4, nous avons ainsi étudié plusieurs classifications

recouvrantes des occurrences de ce motif selon le contexte 3D, pour déterminer si les similarités de contexte 3D parmi ces occurrences pouvaient apporter des informations nouvelles sur ce motif. En comparant des sous-contextes 3D, c'est-à-dire des parties du contexte 3D seulement, nous avons remarqué que des similarités apparaissaient entre occurrences non homologues, qui étaient moins visibles en comparant les contextes 3D dans leur totalité. Cela peut indiquer que certaines parties du contexte 3D, en particulier le brin de la boucle, sont mieux conservées que d'autres, parmi les occurrences de motif A-minor. Nous pourrions alors formuler l'hypothèse que le contexte du brin de la boucle, mieux conservé, est plus important pour la structure 3D que celui du brin de l'hélice.

Les contextes de certaines occurrences ainsi classifiées dans une même classe selon leur sous-contexte 3D, partagent des motifs locaux déjà caractérisés d'un point de vue de la topologie et de la séquence, comme les boucles GNRA, ou les boucles A-rich. Ces classifications selon les sous-contextes 3D sont d'ailleurs cohérentes avec les classifications obtenues selon la topologie de ces mêmes sous-contextes, en utilisant notre représentation de k-extensions contractées. Cela confirme que la topologie de contexte permet de déterminer le contexte 3D pour le motif A-minor dans la plupart des cas.

Dans une perspective de prédiction, nous avons finalement souhaité déterminer, dans le chapitre 5, dans quelle mesure il était possible d'utiliser la topologie de contexte pour prédire la présence d'un motif A-minor. Nous avons utilisé la classification des occurrences de motif A-minor selon le contexte 3D total (classification à 4 branches). Nous avons défini la topologie de contexte commune à une classe donnée comme le sous-graphe commun à cette classe, et nous avons également défini la séquence commune associée à ce contexte, à l'aide d'expressions régulières. Puis, nous avons recherché des occurrences de ces représentants de classe ainsi définis dans deux ensembles de séquences et de graphes d'ARN. Plusieurs informations principales peuvent être déduites de cette étude. Premièrement, la séquence seule associée au contexte structural n'est pas suffisante pour prédire la présence d'un motif A-minor. Deuxièmement, considérer la topologie seule fournit, pour la plupart des classes, de meilleurs résultats que considérer la séquence seule, et que considérer la séquence avec la topologie réduite aux interactions canoniques appartenant à une structure secondaire. Cette observation indique que la topologie est mieux conservée que la séquence, et que les interactions non canoniques ont de l'importance. Enfin, le signal fourni par la topologie et la séquence combinées est un bon signal pour la quasi-totalité des classes ne contenant que des occurrences homologues, même pour des classes contenant des occurrences issues de molécules certes homologues mais très éloignées d'un point de vue évolutif (des molécules des trois règnes du Vivant). Les motifs A-minor de ces dernières classes sont peut-être particulièrement importants pour la structure de la molécule, étant donné qu'ils sont conservés dans des molécules des trois règnes du Vivant. Ces résultats sur le motif

A-minor montrent que le contexte structural très local de ce motif a de l'importance pour sa formation dans de nombreux cas, et laissent ainsi espérer que la prédiction de certaines occurrences de motif A-minor soit possible.

Nous avons ainsi développé dans cette thèse une méthode permettant de modéliser et comparer des contextes structuraux de motifs longue distance d'ARN par des algorithmes de graphes et de clustering. Cette méthode permet l'obtention de classifications automatiques et exhaustives d'occurrences de motif selon la similarité de leur contexte topologique, ou selon la similarité de leur contexte 3D. Cette méthode a été appliquée sur trois motifs d'ARN, mais peut être utilisée sur d'autres motifs, de manière non entièrement automatique cependant à l'heure de l'écriture de cette thèse. La métrique de similarité des contextes topologiques, appelée similarité contextuelle, peut être paramétrée pour considérer de manière plus ou moins identique des éléments de structure secondaire de taille différente.

Ces travaux ouvrent des perspectives aussi bien sur le plan de la modélisation et sur le plan technique que plus largement sur le problème de prédiction des motifs structuraux d'ARN. C'est ce que nous allons décrire à présent.

L'échelle que nous utilisons dans notre modélisation du contexte structural topologique est ainsi l'échelle du nucléotide ou de quelques nucléotides pour un sommet, dans le cas de contraction. Comme vu dans le chapitre 4, d'autres travaux [101] ont utilisé une échelle plus grande, de l'ordre de l'élément de structure secondaire, pour étudier le contexte structural du motif A-minor. Mais ces travaux ne prennent pas en compte les interactions non canoniques. Nous pourrions alors envisager une modélisation à gros grain, permettant de représenter les motifs locaux apparaissant à proximité du motif à longue distance que l'on étudie (par exemple, les boucles GNRA pour le motif A-minor). Comme certains de ces motifs locaux sont prédictibles par des méthodes algorithmiques, malgré la présence d'interactions non canoniques, les considérer dans une modélisation serait intéressant pour la prédiction des motifs à longue distance.

De plus, l'une des limites de notre modélisation, pouvant expliquer la non corrélation de la similarité du contexte topologique par rapport à la similarité du contexte 3D pour certaines occurrences de motif, est la non prise en compte de certaines interactions, comme les interactions d'empilements (*stacking*) que l'on sait aujourd'hui importantes pour le repliement d'une molécule d'ARN [117]. Les prendre en considération dans une modélisation du contexte structural pourrait ainsi permettre de se rapprocher de l'information fournie par le contexte 3D. Il est cependant à noter bien sûr, que les graphes étudiés seraient nécessairement de taille plus importante, et devraient peut-être être des multigraphes, car deux mêmes nucléotides peuvent interagir par une interaction de type canonique ou non canonique et par une interaction d'empilements.

Pour rendre nos résultats plus facilement accessibles, nous avons songé à créer une base de données pour stocker et manipuler les différentes classes de motif A-minor avec leurs caractéristiques communes. Le temps ayant manqué

pour cela au cours de cette thèse, nous pourrions l'envisager dans l'avenir.

Quant à la prédiction du motif A-minor, la topologie commune que nous construisons ainsi pour chaque classe d'occurrences pourrait être considérée comme un "méta-motif", c'est-à-dire un motif incluant le motif A-minor et les autres interactions de la topologie commune. Ce méta-motif, associé à une signature de séquence peut-être plus complexe que ce que l'on définit dans ce document, pourrait être utilisé en entrée des méthodes existantes de prédiction de motifs d'ARN les plus récentes [96], et serait peut-être plus facile à prédire que le motif A-minor seul.

Une perspective plus globale pourrait être d'utiliser les informations d'homologie que nous avons définies, dans des approches comparatives pour aider à la prédiction des motifs A-minor notamment. Nous avons vu dans le chapitre 5 que les occurrences homologues de motif A-minor partagent des séquences et des topologies de contexte similaires, qui sont retrouvées peu fréquemment dans les structures d'ARN. Ces informations pourraient constituer des modèles à rechercher dans les structures d'ARN.

Bibliographie

[1] F. N. Abu-Khzam, N. F. Samatova, M. A. Rizk, and M. A. Langston. The maximum common subgraph problem : Faster solutions via vertex cover. In *IEEE/ACS International Conference on Computer Systems and Applications*, pages 367–373, 2007. doi:10.1109/AICCSA.2007.370907.

[2] T. Akutsu. A polynomial time algorithm for finding a largest common subgraph of almost trees of bounded degree. *IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences*, E76-A(9) :1488–1493, Sept. 1993.

[3] T. Akutsu. Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discrete Applied Mathematics*, 104(1) :45–62, Aug. 2000. doi:10.1016/S0166-218X(00)00186-4.

[4] M. Andronescu, D. Dees, L. Slaybaugh, Y. Zhao, A. Condon, B. Cohen, and S. Skiena. Algorithms for testing that sets of DNA words concatenate without secondary structure. *Natural Computing*, 2(4) :391–415, Dec. 2003. doi:10.1023/B:NACO.0000006770.91995.ec.

[5] J. Aslam, K. Pelekhov, and D. Rus. Static and dynamic information organisation with star clusters. In *Association for Computing Machinery, New York*, pages 208–217, Nov. 1998.

[6] N. Ban, P. Nissen, J. Hansen, P. B. Moore, and T. A. Steitz. The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. *Science (New York, N.Y.)*, 289(5481) :905–920, Aug. 2000. doi:10.1126/science.289.5481.905.

[7] C. Berge. *Théorie des graphes et ses applications*. Paris, Dunod, 1958.

[8] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acid Research*, page 235–242, Jan. 2000. doi:10.1093/nar/28.1.235.

[9] G. A. Bermejo, G. M. Clore, and C. D. Schwieters. Improving NMR Structures of RNA. *Structure*, 24(5) :806–815, May 2016. doi:10.1016/j.str.2016.03.007.

[10] S. H. Bokhari. On the Mapping Problem. *IEEE Transactions on Computers*, C-30(3) :207–214, Mar. 1981. doi:10.1109/TC.1981.1675756.

[11] M. J. Boniecki, G. Lach, W. K. Dawson, K. Tomala, P. Lukasz, T. Soltysinski, K. M. Rother, and J. M. Bujnicki. SimRNA : a coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic Acids Research*, 44(7) :e63–e63, Apr. 2016. doi:10.1093/nar/gkv1479.

[12] A. Bonnet, A. R. Grosso, A. Elkaoutari, E. Coleno, A. Presle, S. C. Sridhara, G. Janbon, V. Géli, S. F. d. Almeida, and B. Palancade. Introns Protect Eukaryotic Genomes from Transcription-Associated Genetic Instability. *Molecular Cell*, 67(4) :608–621.e6, Aug. 2017. doi:10.1016/j.molcel.2017.07.002.

[13] M. Boudard, D. Barth, J. Bernauer, A. Denise, and J. Cohen. GARN2 : coarse-grained prediction of 3D structure of large RNA molecules by regret minimization. *Bioinformatics (Oxford, England)*, 33(16) :2479–2486, Aug. 2017. doi:10.1093/bioinformatics/btx175.

[14] M. Boudard, J. Bernauer, D. Barth, J. Cohen, and A. Denise. GARN : Sampling RNA 3D Structure Space with Game Theory and Knowledge-Based Scoring Strategies. *PLOS ONE*, 10(8) :e0136444, Aug. 2015. doi:10.1371/journal.pone.0136444.

[15] P. Brion and E. Westhof. Hierarchy and Dynamics of RNA Folding. *Annual Review of Biophysics and Biomolecular Structure*, 26(1) :113–137, 1997. doi:10.1146/annurev.biophys.26.1.113.

[16] R. Brüschweiler. Efficient RMSD measures for the comparison of two molecular ensembles. Root-mean-square deviation. *Proteins*, 50(1) :26–34, 2003. doi:10.1002/prot.10250.

[17] Z. C, T. H, and Z. S. RNAMotifScan : automatic identification of RNA structural motifs using secondary structural alignment. *Nucleic acids research*, 38(18), Oct. 2010. doi:10.1093/nar/gkq672.

[18] J. J. Cannone, S. Subramanian, M. N. Schnare, J. R. Collett, L. M. D'Souza, Y. Du, B. Feng, N. Lin, L. V. Madabusi, K. M. Müller, N. Pande, Z. Shang, N. Yu, and R. R. Gutell. The Comparative RNA Web (CRW) Site : an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*, 3(1) :2, Jan. 2002. doi:10.1186/1471-2105-3-2.

[19] S. Cao and S.-J. Chen. Physics-based de novo prediction of RNA 3D structures. *The Journal of Physical Chemistry. B*, 115(14) :4216–4226, Apr. 2011. doi:10.1021/jp112059y.

[20] E. Capriotti and M. A. Marti-Renom. RNA structure alignment by a unit-vector approach. *Bioinformatics*, 24(16) :i112–i118, 2008. doi:10.1093/bioinformatics/btn288.

[21] E. Charpentier. CRISPR-Cas9 : how research on a bacterial RNA-guided mechanism opened new perspectives in biotechnology and biomedicine. *EMBO Molecular Medicine*, 7(4) :363–365, Apr. 2015. doi:10.15252/emmm.201504847.

[22] E. Charpentier and J. A. Doudna. Rewriting a genome. *Nature*, 495(7439) :50–51, Mar. 2013. doi:10.1038/495050a.

[23] G. Chojnowski, T. Waleń, and J. M. Bujnicki. RNA Bricks—a database of RNA 3D motifs and their interactions. *Nucleic acids research*, 42(D1) :D123–D131, 2014.

[24] J. A. Cruz and E. Westhof. Sequence-based identification of 3D structural modules in RNA with RMDetect. *Nature Methods*, 8(6) :513–521, June 2011. doi:10.1038/nmeth.1603.

[25] K. Darty, A. Denise, and Y. Ponty. VARNA : Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, 25(15) :1974–1975, Aug. 2009. doi:10.1093/bioinformatics/btp250.

[26] R. Das and D. Baker. Automated de novo prediction of native-like RNA tertiary structures. *Proceedings of the National Academy of Sciences*, 104(37) :14664–14669, Sept. 2007. doi:10.1073/pnas.0703836104.

[27] B. Desgraupes. *Introduction aux expressions régulières - Bernard Desgraupes*. Paris, Vuibert, sept 2001.

[28] S. M. Dibrov, J. McLean, J. Parsons, and T. Hermann. Self-assembling RNA square. *Proceedings of the National Academy of Science*, 108 :6405–6408, Apr. 2011. doi:10.1073/pnas.1017999108.

[29] Y. Ding, C. Y. Chan, and C. E. Lawrence. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA (New York, N.Y.)*, 11(8) :1157–1166, Aug. 2005. doi:10.1261/rna.2500605.

[30] Y. Ding and C. E. Lawrence. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Research*, 31(24) :7280–7301, Dec. 2003. doi:10.1093/nar/gkg938.

[31] M. Djelloul and A. Denise. Automated motif extraction and classification in RNA tertiary structures. *RNA*, 14(12) :2489–2497, Dec. 2008. doi:10.1261/rna.1061108.

[32] J. A. Doudna and E. Charpentier. The new frontier of genome engineering with CRISPR-Cas9. *Science*, 346(6213) :1258096, Nov. 2014. doi:10.1126/science.1258096.

[33] J. Downward. RNA interference. *BMJ*, 328(7450) :1245–1248, May 2004. doi:10.1136/bmj.328.7450.1245.

[34] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8) :861–874, June 2006. doi:10.1016/j.patrec.2005.10.010.

[35] W. Fontana, D. A. Konings, P. F. Stadler, and P. Schuster. Statistics of RNA secondary structures. *Biopolymers*, 33(9) :1389–1404, Sept. 1993. doi:10.1002/bip.360330909.

[36] S. M. Freier, R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Neilson, and D. H. Turner. Improved free-energy parameters for predictions of RNA duplex stability. *Proceedings of the National Academy of Sciences of the United States of America*, 83(24) :9373–9377, Dec. 1986.

[37] H. H. Gan, S. Pasquali, and T. Schlick. Exploring the repertoire of RNA secondary motifs using graph theory ; implications for RNA design. *Nucleic Acids Research*, 31(11) :2926–2943, June 2003.

[38] M. R. Garey and D. S. Johnson. *Computers and intractability : a guide to the theory of NP-completeness*, volume 29. WH Freeman and Company, New York, 1979.

[39] E. F. Garman. Developments in x-ray crystallographic structure determination of biological macromolecules. *Science (New York, N.Y.)*, 343(6175) :1102–1108, Mar. 2014. doi:10.1126/science.1247829.

[40] C. Geary, A. Chworos, and L. Jaeger. Promoting RNA helical stacking via A-minor junctions. *Nucleic Acids Research*, 39(3) :1066–1080, Feb. 2011. doi:10.1093/nar/gkq748.

[41] C. Gianfrotta, V. Reinharz, D. Barth, and A. Denise. A Graph-Based Similarity Approach to Classify Recurrent Complex Motifs from Their Context in RNA Structures. In D. Coudert and E. Natale, editors, *19th International Symposium on Experimental Algorithms (SEA 2021)*, volume 190 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 19 :1–19 :18, Dagstuhl, Germany, 2021. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. doi:10.4230/LIPIcs.SEA.2021.19.

[42] D. P. Giedroc, C. A. Theimer, and P. L. Nixon. Structure, stability and function of RNA pseudoknots involved in stimulating ribosomal frameshifting. *Journal of Molecular Biology*, 298(2) :167–185, Apr. 2000. doi:10.1006/jmbi.2000.3668.

[43] R. Giegerich, B. Voss, and M. Rehmsmeier. Abstract shapes of RNA. *Nucleic Acids Research*, 32(16) :4843–4851, 2004. doi:10.1093/nar/gkh779.

[44] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S. R. Eddy. Rfam : an RNA family database. *Nucleic Acids Research*, 31(1) :439–441, Jan. 2003.

[45] T. M. Henkin. Riboswitch RNAs : using RNA to sense cellular metabolism. *Genes & Development*, 22(24) :3383–3390, Dec. 2008. doi:10.1101/gad.1747308.

[46] M. Hochsmann, B. Voss, and R. Giegerich. Pure multiple RNA secondary structure alignments : a progressive profile approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(1) :53–62, Jan. 2004. doi:10.1109/TCBB.2004.11.

[47] T. Ipoutcha, I. Tsarmpopoulos, V. Talenton, C. Gaspin, A. Moisan, C. A. Walker, J. Brownlie, A. Blanchard, P. Thebault, and P. Sirand-Pugnet. Multiple Origins and Specific Evolution of CRISPR/Cas9 Systems in Minimal Bacteria (Mollicutes). *Frontiers in Microbiology*, 10, 2019.

[48] S. Islam, M. M. Rahaman, and S. Zhang. RNAMotifContrast : a method to discover and visualize RNA structural motif subfamilies. *Nucleic Acids Research*, 49(11) :e61, June 2021. doi:10.1093/nar/gkab131.

[49] P. Jaccard. The Distribution of the Flora in the Alpine Zone. *New Phytologist*, 11(2) :37–50, 1912. doi:10.1111/j.1469-8137.1912.tb05611.x.

[50] M. Jonikas, R. Radmer, A. Laederach, R. Das, S. Pearlman, D. Herschlag, and R. Altman. Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA (New York, N.Y.)*, 15 :189–99, Mar. 2009. doi:10.1261/rna.1270809.

[51] A. Juanchich, P. Bardou, O. Rué, J.-C. Gabillard, C. Gaspin, J. Bobe, and Y. Guiguen. Characterization of an extensive rainbow trout miRNA transcriptome by next generation sequencing. *BMC Genomics*, 17(1) :164, Mar. 2016. doi:10.1186/s12864-016-2505-9.

[52] I. Kalvari, J. Argasinska, N. Quinones-Olvera, E. P. Nawrocki, E. Rivas, S. R. Eddy, A. Bateman, R. D. Finn, and A. I. Petrov. Rfam 13.0 : shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Research*, 46(D1) :D335–D342, Jan. 2018. doi:10.1093/nar/gkx1038.

[53] V. Kann. On the Approximability of NP-complete Optimization Problems. page 168, mai 1992. PhD thesis.

[54] P. Kerpedjiev, C. H. z. Siederdissen, and I. L. Hofacker. Predicting RNA 3D structure using a coarse-grain helix-centered model. *RNA*, 21(6) :1110–1121, Jan. 2015. doi:10.1261/rna.047522.114.

[55] N. Kim, M. Zahran, and T. Schlick. Computational prediction of riboswitch tertiary structures including pseudoknots by RAGTOP : a hierarchical graph sampling approach. *Methods in Enzymology*, 553 :115–135, 2015. doi:10.1016/bs.mie.2014.10.054.

[56] D. Klein, T. Schmeing, P. Moore, and T. Steitz. The kink-turn : a new RNA secondary structure motif. *The EMBO Journal*, 20(15) :4214–4221, Aug. 2001. doi:10.1093/emboj/20.15.4214.

[57] E. V. Koonin. RNA Worlds : From Life's Origins to Diversity in Gene Regulation edited by John F. Atkins, Raymond F. Gesteland, and Thomas

R. Cech. *The Quarterly Review of Biology*, 87(1) :66–66, Mar. 2012. doi:10.1086/663891.

[58] E. B. Krissinel and K. Henrick. Common subgraph isomorphism detection by backtracking search. *Software : Practice and Experience*, 34(6) :591–607, 2004. doi:10.1002/spe.588.

[59] S.-Y. Le, R. Nussinov, and J. V. Maizel. Tree graphs of RNA secondary structures and their comparisons. *Computers and Biomedical Research*, 22(5) :461–473, Oct. 1989. doi:10.1016/0010-4809(89)90039-6.

[60] A. Legendre, E. Angel, and F. Tahi. Bi-objective integer programming for RNA secondary structure prediction with pseudoknots. *BMC Bioinformatics*, 19(1) :13, Jan. 2018. doi:10.1186/s12859-018-2007-7.

[61] N. B. Leontis, J. Stombaugh, and E. Westhof. The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic Acids Research*, 30(16) :3497–3531, Aug. 2002. doi:10.1093/nar/gkf481.

[62] N. B. Leontis and E. Westhof. Geometric nomenclature and classification of RNA base pairs. *RNA*, 7(4) :499–512, Apr. 2001.

[63] N. B. Leontis and E. Westhof. Analysis of RNA motifs. *Current opinion in structural biology*, 13(3) :300–308, 2003.

[64] N. B. Leontis and C. L. Zirbel. Nonredundant 3D Structure Datasets for RNA Knowledge Extraction and Benchmarking. In N. Leontis and E. Westhof, editors, *RNA 3D Structure Analysis and Prediction*, Nucleic Acids and Molecular Biology, pages 281–298. Springer, Berlin, Heidelberg, 2012. doi:10.1007/978-3-642-25740-7_13.

[65] A. Lescoute, N. B. Leontis, C. Massire, and E. Westhof. Recurrent structural RNA motifs, Isostericity Matrices and sequence alignments. *Nucleic Acids Research*, 33(8) :2395–2409, 2005. doi:10.1093/nar/gki535.

[66] A. Lescoute and E. Westhof. The A-minor motifs in the decoding recognition process. *Biochimie*, 88(8) :993–999, Aug. 2006. doi:10.1016/j.biochi.2006.05.018.

[67] A. Liaw and M. Wiener. Classification and Regression by randomForest. 2 :5, 2002.

[68] R. Lorenz, S. H. Bernhart, C. Höner Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker. ViennaRNA Package 2.0. *Algorithms for molecular biology : AMB*, 6 :26, Nov. 2011. doi:10.1186/1748-7188-6-26.

[69] X.-J. Lu, H. J. Bussemaker, and W. K. Olson. DSSR : an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Research*, 43(21) :e142, Dec. 2015. doi:10.1093/nar/gkv716.

[70] H. Ma, X. Jia, K. Zhang, and Z. Su. Cryo-EM advances in RNA structure determination. *Signal Transduction and Targeted Therapy*, 7(1) :1–6, Feb. 2022. doi:10.1038/s41392-022-00916-0.

[71] M. Magnus, K. Kappel, R. Das, and J. M. Bujnicki. RNA 3D structure prediction guided by independent folding of homologous sequences. *BMC Bioinformatics*, 20(1) :512, 2019. doi:10.1186/s12859-0193120-y.

[72] D. H. Mathews and D. H. Turner. Dynalign : an algorithm for finding the secondary structure common to two RNA sequences. *Journal of Molecular Biology*, 317(2) :191–203, Mar. 2002. doi:10.1006/jmbi.2001.5351.

[73] J. S. Mattick and I. V. Makunin. Non-coding RNA. *Human Molecular Genetics*, 15(suppl_1) :R17–R29, Apr. 2006. doi:10.1093/hmg/ddl046.

[74] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7) :1105–1119, June 1990. doi:10.1002/bip.360290621.

[75] F. Michel and E. Westhof. Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *Journal of Molecular Biology*, 216(3) :585–610, Dec. 1990. doi:10.1016/0022-2836(90)90386-Z.

[76] S. Motameny, S. Wolters, P. Nürnberg, and B. Schumacher. Next Generation Sequencing of miRNAs – Strategies, Resources and Methods. *Genes*, 1(1) :70–84, June 2010. doi:10.3390/genes1010070.

[77] J. H. Nagel, A. P. Gultyaev, K. Gerdes, and C. W. Pleij. Metastable structures and refolding kinetics in hok mRNA of plasmid R1. *RNA (New York, N.Y.)*, 5(11) :1408–1418, Nov. 1999. doi:10.1017/s1355838299990805.

[78] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3) :443–453, Mar. 1970. doi:10.1016/0022-2836(70)90057-4.

[79] P. Nissen, J. A. Ippolito, N. Ban, P. B. Moore, and T. A. Steitz. RNA tertiary interactions in the large ribosomal subunit : The A-minor motif. *Proceedings of the National Academy of Sciences*, 98(9) :4899–4903, Apr. 2001. doi:10.1073/pnas.081082398.

[80] H. F. Noller, V. Hoffarth, and L. Zimniak. Unusual resistance of peptidyl transferase to protein extraction procedures. *Science (New York, N.Y.)*, 256(5062) :1416–1419, June 1992. doi:10.1126/science.1604315.

[81] R. Nussinov and A. B. Jacobson. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proceedings of the National Academy of Sciences*, 77(11), Nov. 1980. doi:10.1073/pnas.77.11.6309.

[82] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043) :814–818, June 2005. doi:10.1038/nature03607.

[83] M. Parisien and F. Major. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, 452 :51–5, Apr. 2008. doi:10.1038/nature06684.

[84] S. Pasquali and P. Derreumaux. HiRE-RNA : A High Resolution Coarse-Grained Energy Model for RNA. *The Journal of Physical Chemistry B*, 114(37) :11957–11966, Sept. 2010. doi:10.1021/jp102497y.

[85] S. Pasquali, H. H. Gan, and T. Schlick. Modular RNA architecture revealed by computational analysis of existing pseudoknots and ribosomal RNAs. *Nucleic Acids Research*, 33(4) :1384–1398, 2005. doi:10.1093/nar/gki267.

[86] W. R. Pearson. An introduction to sequence similarity ("homology") searching. *Current Protocols in Bioinformatics*, Chapter 3 :Unit3.1, June 2013. doi:10.1002/0471250953.bi0301s42.

[87] A. I. Petrov, C. L. Zirbel, and N. B. Leontis. Automated classification of RNA 3D motifs and the RNA 3D Motif Atlas. *Rna*, 19(10) :1327–1340, 2013.

[88] M. Popenda, M. Szachniuk, M. Antczak, K. J. Purzycka, P. Lukasiak, N. Bartol, J. Blazewicz, and R. W. Adamiak. Automated 3D structure composition for large RNAs. *Nucleic Acids Research*, 40(14) :e112, Aug. 2012. doi:10.1093/nar/gks339.

[89] A. Pérez-Suárez, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, and J. E. Medina-Pagola. OClustR : A new graph-based algorithm for overlapping clustering. *Neurocomputing*, 121 :234–247, Dec. 2013. doi:10.1016/j.neucom.2013.04.025.

[90] J. Raymond, E. Gardiner, and P. Willett. RASCAL : Calculation of Graph Similarity using Maximum Common Edge Subgraphs. *Computer Journal*, 45 :631–644, Apr. 2002.

[91] J. Reeder, P. Steffen, and R. Giegerich. pknotsRG : RNA pseudoknot folding including near-optimal structures and sliding windows. *Nucleic Acids Research*, 35(suppl_2) :W320–W324, July 2007. doi:10.1093/nar/gkm258.

[92] V. Reinharz, A. Soulé, E. Westhof, J. Waldispühl, and A. Denise. Mining for recurrent long-range interactions in RNA structures reveals embedded hierarchies in network families. *Nucleic Acids Research*, 46(8) :3841–3851, May 2018. doi:10.1093/nar/gky197.

[93] E. Rivas and S. R. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *Journal of Molecular Biology*, 285(5) :2053–2068, Feb. 1999. doi:10.1006/jmbi.1998.2436.

[94] M. Saman Booy, A. Ilin, and P. Orponen. RNA secondary structure prediction with convolutional neural networks. *BMC Bioinformatics*, 23(1) :58, Feb. 2022. doi:10.1186/s12859-021-04540-7.

[95] D. Sankoff. Simultaneous Solution of the RNA Folding, Alignment and Protosequence Problems. *SIAM Journal on Applied Mathematics*, 45(5) :810–825, Oct. 1985. doi:10.1137/0145048.

[96] R. Sarrazin-Gendron, V. Reinharz, C. G. Oliver, N. Moitessier, and J. Waldispühl. Automated, customizable and efficient identification of 3D base pair modules with BayesPairing. *Nucleic Acids Research*, 47(7) :3321–3332, Apr. 2019. doi:10.1093/nar/gkz102.

[97] R. Sarrazin-Gendron, H.-T. Yao, V. Reinharz, C. G. Oliver, Y. Ponty, and J. Waldispühl. Stochastic Sampling of Structural Contexts Improves the Scalability and Accuracy of RNA 3D Module Identification. In R. Schwartz, editor, *Research in Computational Molecular Biology*, Lecture Notes in Computer Science, pages 186–201, Cham, 2020. Springer International Publishing. doi:10.1007/978-3-030-45257-5_12.

[98] M. Sarver, C. L. Zirbel, J. Stombaugh, A. Mokdad, and N. B. Leontis. FR3D : finding local and composite recurrent structural motifs in RNA 3D structures. *Journal of mathematical biology*, 56(1-2) :215–252, Jan. 2008. doi:10.1007/s00285-007-0110-x.

[99] K. Sato, M. Akiyama, and Y. Sakakibara. RNA secondary structure prediction using deep learning with thermodynamic integration. *Nature Communications*, 12(1) :941, Feb. 2021. doi:10.1038/s41467-021-21194-4.

[100] K. Sato, Y. Kato, M. Hamada, T. Akutsu, and K. Asai. IPknot : fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics (Oxford, England)*, 27(13) :i85–93, July 2011. doi:10.1093/bioinformatics/btr215.

[101] A. A. Shalybkova, D. S. Mikhailova, I. V. Kulakovskiy, L. I. Fakhranurova, and E. F. Baulin. Annotation of the local context of the RNA secondary structure improves the classification and prediction of A-minors. *RNA*, page rna.078535.120, May 2021. doi:10.1261/rna.078535.120.

[102] B. A. Shapiro and K. Zhang. Comparing multiple RNA secondary structures using tree comparisons. *Bioinformatics*, 6(4) :309–318, Oct. 1990. doi:10.1093/bioinformatics/6.4.309.

[103] S. Sharma, F. Ding, and N. V. Dokholyan. iFoldRNA : three-dimensional RNA structure prediction and folding. *Bioinformatics (Oxford, England)*, 24(17) :1951–1952, Sept. 2008. doi:10.1093/bioinformatics/btn328.

[104] S. Sheikh, R. Backofen, and Y. Ponty. *Impact of the Energy Model on the Complexity of RNA Folding with Pseudoknots*. July 2012. doi:10.1007/978-3-642-31265-6_26.

[105] W. Shu, X. Bo, Z. Zheng, and S. Wang. A novel representation of RNA secondary structure based on element-contact graphs. *BMC Bioinformatics*, 9(1) :188, Apr. 2008. doi:10.1186/1471-2105-9-188.

[106] A. Soulé, V. Reinharz, R. Sarrazin-Gendron, A. Denise, and J. Waldispühl. Finding recurrent RNA structural networks with fast maximal common subgraphs of edge-colored graphs. *PLOS Computational Biology*, 17(5) :e1008990, May 2021. doi:10.1371/journal.pcbi.1008990.

[107] J. Spirollari, J. T. L. Wang, K. Zhang, V. Bellofatto, Y. Park, and B. A. Shapiro. Predicting consensus structures for RNA alignments via pseudo-energy minimization. *Bioinformatics and Biology Insights*, 3 :51–69, June 2009. doi:10.4137/bbi.s2578.

[108] A. P. Suárez, J. F. M. Trinidad, J. A. C. Ochoa, and J. E. Medina Pagola. A New Incremental Algorithm for Overlapped Clustering. In E. Bayro-Corrochano and J.-O. Eklundh, editors, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, Lecture Notes in Computer Science, pages 497–504, Berlin, Heidelberg, 2009. Springer. doi:10.1007/978-3-642-10268-4_58.

[109] R. K. Tan, A. S. Petrov, and S. C. Harvey. YUP : A Molecular Simulation Program for Coarse-Grained and Multi-Scaled Models. *Journal of chemical theory and computation*, 2(3) :529–540, May 2006. doi:10.1021/ct050323r.

[110] T. Tanimoto. *An elementary mathematical theory of classification and prediction*. New York, International Business Machines Corporation, 1958.

[111] D. Thirumalai and C. Hyeon. Theory of RNA Folding : From Hairpins to Ribozymes. In N. G. Walter, S. A. Woodson, and R. T. Batey, editors, *Non-Protein Coding RNAs*, Springer Series in Biophysics, pages 27–47. Springer, Berlin, Heidelberg, 2009. doi:10.1007/978-3-540-70840-7_2.

[112] I. Tinoco and C. Bustamante. How RNA folds. *Journal of Molecular Biology*, 293(2) :271–281, Oct. 1999. doi:10.1006/jmbi.1999.3001.

[113] H. Touzet and O. Perriquet. CARNAC : folding families of related RNAs. *Nucleic Acids Research*, 32(Web Server issue) :W142–W145, July 2004. doi:10.1093/nar/gkh415.

[114] D. H. Turner and D. H. Mathews. NNDB : the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Research*, 38(Database issue) :D280–D282, Jan. 2010. doi:10.1093/nar/gkp892.

[115] F. van Hemert and B. Berkhout. Nucleotide composition of the Zika virus RNA genome and its codon usage. *Virology Journal*, 13(1) :95, June 2016. doi:10.1186/s12985-016-0551-1.

[116] G. Varani and W. H. McClain. The G·U wobble base pair. *EMBO Reports*, 1(1) :18–23, July 2000. doi:10.1093/embo-reports/kvd001.

[117] Q. Vicens and J. S. Kieft. Thoughts on how to think (and talk) about RNA structure. *Proceedings of the National Academy of Sciences*, 119(17) :e2112677119, Apr. 2022. Publisher : Proceedings of the National Academy of Sciences. URL : https://www.pnas.org/doi/10.1073/pnas.2112677119, doi:10.1073/pnas.2112677119.

[118] J. Wang, P. Daldrop, L. Huang, and D. M. J. Lilley. The k-junction motif in RNA structure. *Nucleic Acids Research*, 42(8) :5322–5331, Apr. 2014. doi:10.1093/nar/gku144.

[119] J. Wang, K. Mao, Y. Zhao, C. Zeng, J. Xiang, Y. Zhang, and Y. Xiao. Optimization of RNA 3D structure prediction using evolutionary restraints of nucleotide-nucleotide interactions from direct coupling analysis. *Nucleic Acids Research*, 45(11) :6299–6309, June 2017. doi:10.1093/nar/gkx386.

[120] J. J. Whang, I. S. Dhillon, and D. F. Gleich. Non-exhaustive, Overlapping k-means. In *Proceedings of the 2015 SIAM International Conference on Data Mining (SDM)*, Proceedings, pages 936–944. Society for Industrial and Applied Mathematics, June 2015. doi:10.1137/1.9781611974010.105.

[121] S. Wuchty, W. Fontana, I. L. Hofacker, and P. Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49(2) :145–165, Feb. 1999. doi:10.1002/(SICI)1097-0282(199902)49:2<145::AID-BIP4>3.0.CO;2-G.

[122] A. Yamaguchi, K. F. Aoki, and H. Mamitsuka. Finding the maximum common subgraph of a partial k-tree and a graph with a polynomially bounded number of spanning trees. *Information Processing Letters*, 92(2) :57–63, 2004. URL : https://www.sciencedirect.com/science/article/pii/S0020019004002005, doi:https://doi.org/10.1016/j.ipl.2004.06.019.

[123] J. Yao, V. Reinharz, F. Major, and J. Waldispühl. RNA-MoIP : prediction of RNA secondary structure and local 3D motifs from sequence data. *Nucleic acids research*, 45(W1) :W440–W444, 2017.

[124] J. Yao, V. Reinharz, F. Major, and J. Waldispühl. RNA-MoIP : prediction of RNA secondary structure and local 3D motifs from sequence data. *Nucleic Acids Research*, 45(W1) :W440–W444, 2017. doi:10.1093/nar/gkx429.

[125] Y. Zhang, F. Abu-khzam, N. Baldwin, E. Chesler, M. Langston, and N. Samatova. Genome-Scale Computational Approaches to Memory-Intensive Applications in Systems Biology. volume 2005, page 12, Jan. 2005. doi:10.1109/SC.2005.29.

[126] C. L. Zirbel, J. Roll, B. A. Sweeney, A. I. Petrov, M. Pirrung, and N. B. Leontis. Identifying novel sequence variants of RNA 3D motifs. *Nucleic Acids Research*, 43(15) :7504–7520, Sept. 2015. doi:10.1093/nar/gkv651.

[127] M. Zuker. On finding all suboptimal foldings of an RNA molecule. *Science (New York, N.Y.)*, 244(4900) :48–52, Apr. 1989. doi:10.1126/science.2468181.

[128] M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31(13) :3406–3415, July 2003. doi:10.1093/nar/gkg595.

[129] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9(1) :133–148, Jan. 1981.