Modélisation 3D de complexes ARN-protéine par assemblage combinatoire de fragments structuraux

Antoine Moniot

UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

Modélisation 3D de complexes ARN-protéine par assemblage combinatoire de fragments structuraux

THÈSE

présentée et soutenue publiquement le 12 décembre 2022

pour l'obtention du

Doctorat de l'Université de Lorraine

(mention informatique)

par

Antoine Moniot

Composition du jury

Président :	Alexandre G. de Brevern
Rapporteurs :	Fabrice Leclerc
	Nicolas Wicker
Examinateurs :	Alexandre G. de Brevern
	Fariza Tahi
Encadrants :	Isaure Chauvot de Beauchêne
	Yann Guermeur

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Mis en page avec la classe thesul.

Remerciements

Je remercie tout d'abord les membres de mon jury de thèse **Fabrice Leclerc**, **Nicolas Wicker**, **Alexandre G. de Brevern** et **Fariza Tahi** pour avoir accepté d'évaluer mon travail de thèse.

Je souhaite ensuite remercier mes encadrants de thèse, **Dave Ritchie**, **Isaure Chauvot de Beauchêne** et **Yann Guermeur** pour l'accompagnement que vous m'avez donné pendant ces quatre années. Dave, je tiens à te remercier pour la rigueur scientifique que tu m'as montrée et que tu as essayé de m'inculquer. Je te remercie pour la patience que tu avais à mon égard et je mesure la chance que j'ai eu de pouvoir travailler avec toi. Ton départ m'a beaucoup marqué et je me souviendrai de toi jusqu'au bout. Isaure, je tiens à te remercier pour ton implication dans ma thèse, pour les discussions à bâtons rompus qui ont été les meilleurs moments de ma thèse et pour ta bonne humeur perpétuelle. Je te remercie enfin pour les très nombreuses relectures que tu as pu faire de ce manuscrit. Yann, je te remercie d'avoir pris ma thèse en main à un moment où j'en avais besoin. Tu m'as fait découvrir la complexité de la recherche en informatique, et je reste admiratif devant ta capacité à aller chercher les points de détail.

Je remercie l'ensemble de l'équipe CAPSID avec laquelle j'ai pu évoluer au cours de ma thèse. **Marie-Dominique** et **Malika**, je vous remercie pour votre point de vue critique toujours très pertinent sur ma thèse. Je tiens à remercier mes nombreux co-bureaux, **Anna**, **Athénaïs**, **Bishnu**, **Diego** et **Kévin** pour les moments de discussions, les moments drôles et les moments réconfortants. Merci **Dominique** pour les sursauts dus à tes arrivées surprises. Je remercie aussi **Hrishikesh** et **Kamrul** pour les nombreux repas à trois. Encore merci pour la bonne ambiance de l'équipe qui donne envie de venir au labo pour travailler avec vous tous.

Je remercie aussi l'ensemble du personnel du laboratoire, que ce soit l'accueil ou la restauration, merci pour votre bienveillance et votre enthousiasme à toute épreuve. Plus particulièrement, merci à **Florianne** et **Caro** pour les nombreux cafés réalisés et les discussions qui les accompagnent et **Isabelle** pour les supers desserts et ta bonne humeur quotidienne.

Je tiens à remercier mes amis **Claire**, **Emmanuel**, **Philippe**, **Max** et **Clémentine** pour les soirées au bar, les soirées jeux et globalement la majorité des moments de détente survenus au cours de ces quatre années. Merci aussi pour votre soutien sans faille et votre patience pendant mes moments de doutes.

Un remerciement à ma famille, sans qui je n'en serais pas là aujourd'hui. **Papa**, **Maman**, **Manon** et **Susie**, merci pour votre soutien pendant ces quatre années, mais aussi pendant mes neuf longues années d'études !

Enfin, un immense merci à **Méline** sans qui je ne serais jamais arrivé au bout de cette thèse. Merci pour ta joie, ton engagement et tous les moments passés ensemble. Merci à ta famille, **Marie**, **Didier**, **Élodie** et **Johan** pour votre accueil pendant les confinements, ça restera des moments inoubliables.

Sommaire

Introduction générale

1 Cadre de l'étude

La connaissance de la structure des complexes ARN-protéine est d'un intérêt majeur en biologie. Du point de vue de la connaissance fondamentale, elle permet de mieux comprendre les types d'interactions entre protéines et ARN ainsi que les processus biologiques dans lesquels ces interactions interviennent. D'un point de vue plus pratique, l'étude des complexes macromoléculaires impliqués dans des maladies peut permettre de trouver des traitements, la Covid-19 et le complexe spike-ACE fournissant un exemple d'actualité. Comme souvent en biologie structurale, deux approches sont envisageables pour atteindre l'objectif : la détermination expérimentale et la modélisation, les deux méthodes étant complémentaires. L'expérimentation *in vivo* et *in vitro* permet d'identifier des protéines et ARN interagissant. Une structure 3D de la protéine et la structure secondaire de l'ARN dans l'état non lié peuvent être connues expérimentalement (par cristallographie, RMN, etc.) et/ou prédites par les outils computationnels existants. L'objet de cette thèse est donc de développer une méthode permettant de prédire la structure 3D d'un complexe ARN-protéine sous les hypothèses minimales suivantes : une structure 3D de la protéine est connue, ainsi que la séquence et la structure secondaire de la partie de l'ARN en interaction avec la protéine.

De manière générale, le cas soulevant le plus de problèmes pour la modélisation d'un tel complexe est celui où l'ARN est simple brin dans sa région d'interaction. La difficulté découle de la flexibilité de ces chaînes de nucléotides non appariés, et de l'immense espace de recherche conformationnel associé. On recense déjà plusieurs méthodes ayant pour objectif de la surmonter. Parmi celles-ci, l'approche utilisant les hypothèses les plus facilement satisfaites (se rapprochant des hypothèses minimales énoncées plus haut) est la méthode ssRN'ATTRACT, développée actuellement dans l'équipe CAPSID. Elle repose sur l'utilisation de bibliothèques de fragments structuraux d'ARN, des tri-nucléotides, qui sont amarrés individuellement sur la protéine avant d'être assemblés pour reconstituer la chaîne native d'ARN lié. Cette étape d'assemblage est

difficile du fait de la combinatoire qu'engendre le grand nombre de positions possibles pour chaque fragment. Pour diminuer la combinatoire, deux solutions ont été explorées au cours de la thèse : l'une sur la manière d'assembler les fragments, l'autre sur la réduction de la cardinalité des bibliothèques de fragments.

2 Motivation des travaux effectués

La méthode de modélisation des complexes ARN simple brin-protéine ssRN'ATTRACT ne permet d'obtenir des modèles complets de l'ARN lié que lorsque des interactions de certains nucléotides de l'ARN avec des résidus particuliers de la protéine peuvent être considérées comme connues. C'est le cas en particulier pour certaines familles conservées de protéines liant les ARN. Cette hypothèse se traduit sous la forme de contraintes : la position des bases des nucléotides ainsi ancrés sur la protéine est connue avec une certaine précision (contraintes de boîte) grâce aux structures expérimentales de complexes homologues. Cependant ces interactions ne sont pas strictement conservées au sein d'une même famille de protéines, et sont inexistantes ou non caractérisées expérimentalement dans d'autres familles. Il apparaît donc utile de pouvoir travailler avec une information plus facilement disponible, plus fiable et/ou recouvrant d'autres cas. Le relâchement de la contrainte sur les points d'ancrage s'accompagne invariablement d'un accroissement de la complexité du problème (explosion de la combinatoire), qui appelle des innovations à la fois dans la spécification de l'approche et dans sa mise en œuvre algorithmique.

Une première innovation a consisté en la caractérisation d'une nouvelle contrainte de fermeture de boucle dans l'algorithme d'assemblage de fragments, lors de son application à une région simple brin insérée dans une structure d'ARN en épingle à cheveux. L'amélioration des bibliothèques de fragments structuraux permet également de limiter l'accroissement de la complexité. Les ϵ-réseaux sont les meilleures structures pour obtenir des bibliothèques de fragments structuraux de cardinalité la plus faible possible à partir d'un nombre fini de fragments. La caractérisation de ces bibliothèques de fragments en ϵ-réseaux nous a amené à explorer l'inférence de ces ϵ-réseaux dans des espaces métriques sur un nombre fini de points. La deuxième innovation est l'application des méthodes de classification ascendante hiérarchique pour cette inférence. Au regard de ces applications, nous avons défini une nouvelle méthode dédiée à l'inférence d'ϵ-réseaux. De plus, la création partiellement automatisée de bases de données structurales d'ADN/ARN et de telles bibliothèques de fragments a fait l'objet du développement d'un nouvel outil.

3 Plan du document

Le chapitre 1 est un chapitre d'état de l'art introduisant les principales méthodes d'assemblage de complexes protéines/ARN. Il s'appuie sur une présentation synthétique des concepts biologiques dont la connaissance est nécessaire à la compréhension du phénomène et des méthodes.

Avec le chapitre 2 débute l'exposé de la contribution originale de la thèse. Il porte sur le développement de l'outil ProtNAff pour la création des bases de données nécessaires à l'application et la validation de notre méthode de modélisation. Ce nouvel outil permet des sélections complexes, dans les bases de données existantes, de jeux de structures d'ARN ayant des caractéristiques choisies, et la création de bibliothèques de fragments à partir de ce jeu. Il a été conçu de façon à dépasser la seule satisfaction de nos besoins, afin d'être adopté par la communauté pour le traitement de problèmes aussi variés que possible autour des interactions protéine-ARN/ADN. Nous l'avons ainsi appliqué aux épingles à cheveux, dont le *docking* fait l'objet d'un autre chapitre de cette thèse.

Le chapitre 3 présente l'application des méthodes de classification ascendante hiérarchique de la littérature pour l'inférence d'ϵ-réseaux. Puis nous décrivons notre méthode dédiée à cette inférence, avant de définir les hypothèses sur l'espace nécessaire à sa mise en œuvre. Enfin, sa spécification aux bibliothèques de fragments structuraux est détaillée.

Le chapitre 4 présente deux parties. La première partie est consacrée au docking de la boucle des épingles à cheveux. Elle introduit une contrainte traduisant la connaissance approximative de la distance entre les nucléotides fermant cette boucle. Le résultat majeur est la démonstration par l'expérience de la possibilité de remplacer la contrainte antérieure relative à l'interaction de certains nucléotides avec des résidus connus de la protéine, sous réserve d'un échantillonnage (au sens du docking) suffisant des positions de fragments. L'information qui s'y substitue, relative à la distance séparant les nucléotides aux extrémités de la boucle, s'avère pleine de promesses pour réaliser le docking d'autres types d'éléments structuraux d'ARN pour lesquels le contact fait intervenir une partie simple brin. La deuxième partie du chapitre 4 vient compléter cette contribution en vue du docking des parties doubles brins de ces éléments structuraux. Elle présente la création et l'analyse de bibliothèques de fragments doubles brins, ainsi que des perspectives pour l'assemblage complet des structures en épingle.

Chapitre 1

Contexte et état de l'art

Les interactions macromoléculaires telles que les interactions ADN-protéine, ARN-protéine ou protéine-protéine sont essentielles au bon fonctionnement de la cellule. Ces interactions interviennent notamment au cours des mécanismes de transcription, de traduction et de régulation (pour ne citer que les plus connus). Savoir comment ces interactions fonctionnent permet d'en limiter les dérèglements. Ceux-ci peuvent amener à des maladies auto-immunes, des cancers, etc.

C'est pour cela que de nombreuses études portent sur ces interactions, parmi lesquelles des études expérimentales, mais aussi de la modélisation pour les cas insolubles expérimentalement et pour éviter ces études expérimentales qui peuvent être coûteuses en temps et en argent.

Ce chapitre présente rapidement les concepts nécessaires à la compréhension de la suite du document, ainsi que les réflexions qui ont amené le sujet de cette thèse. La première partie est une présentation des rôles biologiques des macromolécules et de leur structuration. Puis l'obtention de ces structures de manière expérimentale est présentée avant d'expliquer les méthodes computationnelles pour la prédiction des structures 3D des macromolécules. Enfin, la dernière partie se concentre sur une méthode particulière très importante dans la suite du document : le *docking* par fragments.

1.1 Les macromolécules d'intérêt

Cette section vise à présenter rapidement les macromolécules étudiées dans le cadre de cette thèse : les Acides RiboNucléiques (ARN) et les protéines. Nous nous concentrons tout d'abord sur leurs rôles biologiques, et nous verrons ensuite comment elles se structurent.

1.1.1 Les rôles des macromolécules biologiques et de leurs interactions

Les Acides RiboNucléiques

Les ARN ont tout d'abord été considéré comme une simple copie de l'information génétique contenue dans l'ADN, en accord avec le premier énoncé du dogme central ([90]). Cependant, depuis cette époque, la connaissance des ARN s'est grandement étoffée, et la compréhension des nombreux mécanismes dans lesquels l'ARN est impliqué prouve le rôle crucial que ces molécules jouent.

L'ARN est notamment une molécule permettant la formation des protéines [58]. En effet, plusieurs types d'ARN participent à la transcription et à la traduction. Parmi ceux-ci, les ARN messagers (ARNm) portent l'information génétique codant pour la protéine. Les ARN de transfert (ARNt) déplacent les acides aminés, les briques de base des protéines, et reconnaissent le code génétique (sous forme de triplets de nucléotides). Les ribosomes utilisent les ARN de transferts pour lier les acides aminés et former la future protéine. Ces ribosomes sont des assemblages d'ARN ribosomiques (ARNr) et de protéines. Il existe d'autres classes aux rôles variés, comme les ARNlnc, les miARN, etc.).

Les ARN sont aussi impliqués dans de nombreux processus de signalisation cellulaire, et dans la régulation de certains autres mécanismes cellulaires [56].

Les complexes ARN-protéine

Les protéines sont des macromolécules qui comprennent une ou plusieurs chaînes d'acides aminés. Elles interviennent dans la majorité des fonctions biologiques au sein des organismes : les réactions catalytiques, la réplication de l'ADN, le transport des molécules, les réponses aux stimuli, etc.

Les protéines qui lient les ARN (*RNA-Binding Proteins* : RBPs) ont des rôles divers dans les mécanismes post-transcriptionnels de l'expression des gènes, tels que la régulation du découpage des ARNm, la localisation et le transport de l'ARNm, la stabilisation et la traduction de l'ARNm [81, 25]. Les RBPs participent à la régulation de nombreux processus cellulaires, et sont donc impliquées dans des maladies telles que les cancers [69]. En cas de mutations de la protéine, ses interactions avec l'ARN peuvent être modifiées. Par exemple, la protéine SRSF2 [31] est impliquée dans la maturation de l'ARNm. Certaines mutations de cette protéine . En particulier, la transformation d'une proline en Histidine, Leucine ou Arginine entraîne une modification de la séquence d'ARN reconnus et liée dans l'ARNm, et la surexpression de gènes impliqués dans des leucémies. L'étude des structures et des complexes ARN-protéines a donc un grand intérêt

en médecine, pour comprendre certaines maladies.

1.1.2 Les interactions physico-chimiques

En biologie, il existe de nombreux types d'interactions non covalentes entre macromolécules. Celles qui seront mentionnées dans le reste du document sont présentées ici.

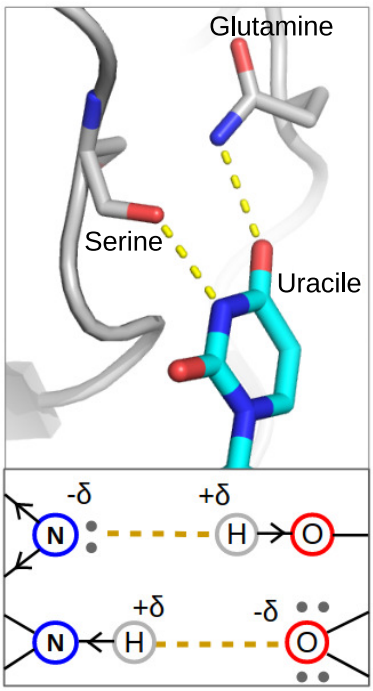
Les liaisons hydrogène sont les interactions les plus répandues dans les macromolécules biologiques. Elles font intervenir un atome électronégatif, c.-à-d., qui porte une pseudo-charge négative, tel que l'Oxygène ou l'Azote (liés à un ou deux carbone(s)), et un atome d'Hydrogène électropositif. Elles déterminent les structures secondaires des protéines et des ARN (voir section 1.1.3 et section 1.1.4). Les molécules d'eau forment aussi de très nombreuses liaisons hydrogène. L'énergie de ces liaisons varie entre 1 et 15 $\mathrm{kJ \cdot mol^{-1}}$.

Les interactions de van der Waals résultent d'une force attractive entre les nuages électroniques de deux atomes adjacents. Ces interactions sont relativement faibles (environ 4 $\mathrm{kJ \cdot mol^{-1}}$) mais restent énergétiquement importantes du fait de leur très grand nombre.

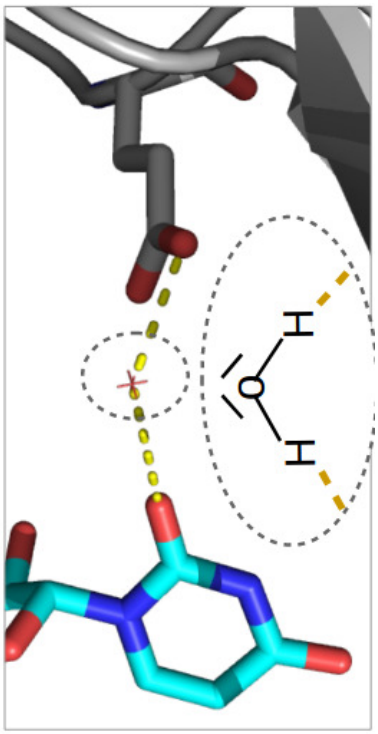
Les interactions ioniques impliquent deux atomes chargés. Si les charges sont opposées, les deux atomes s'attirent, au contraire si les deux charges sont de même polarité, les deux atomes se repoussent. L'énergie de ces liaisons varie entre 170 et 1500 $\mathrm{kJ \cdot mol^{-1}}$. Elles sont très fortes dans le cas où elles impliquent des transferts d'électrons.

Le π - π *stacking*, qui sera nommé empilement par la suite, est une interaction entre deux cycles aromatiques. Cette interaction nécessite une position particulière des deux cycles, soit parallèles, soit perpendiculaires. Ces interactions sont relativement fortes, et celles entre les bases azotées sont très importantes pour la structuration des AN. Dans le cadre des appariements protéine-AN, les acides aminés aromatiques (tyrosine, phénylalanine et tryptophane) constituent des points d'ancrage pour les interactions avec les AN simple brin [36].

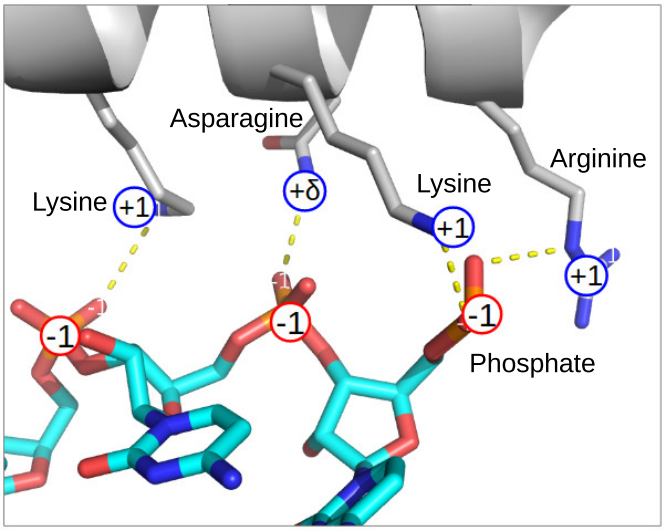
Le π - charges *stacking* sont les interactions faisant intervenir les cycles aromatiques comme des accepteurs ou donneurs d'électrons [52]. Ceux-ci interagissent avec un autre atome porteur de charge, complète (pour des cations ou des anions), ou partielle (pour les hydrogènes liés à des atomes électronégatifs). Dans le cadre de l'ARN l'interaction entre un phosphate et une base azotée permet de stabiliser la structure 3D.



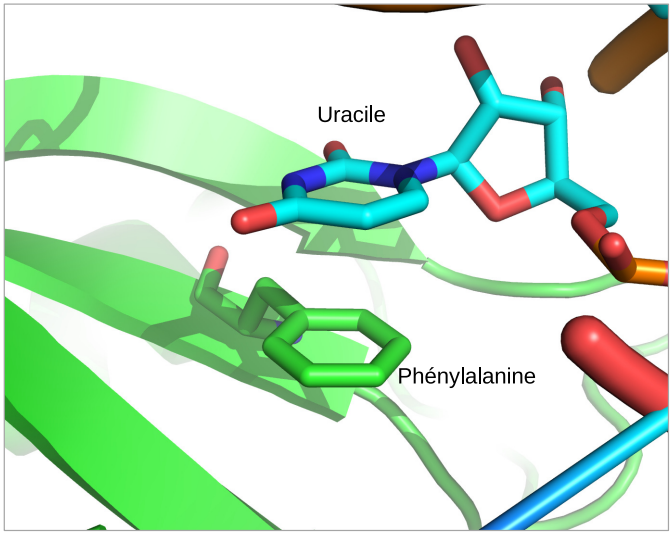
(a) Les liaisons hydrogène.



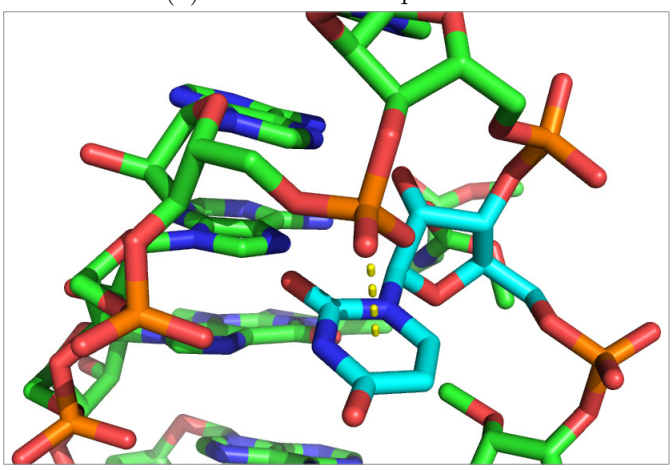
(b) Exemple d'un pont formé par une molécule d'eau.



(c) Les liaisons ioniques.



(d) Interaction d'empilement.



(e) Interaction d'empilement base - phosphate.

FIGURE 1.1 – Représentations des interactions physico-chimiques entre un ARN (en bleu) et une protéine (en gris et vert).

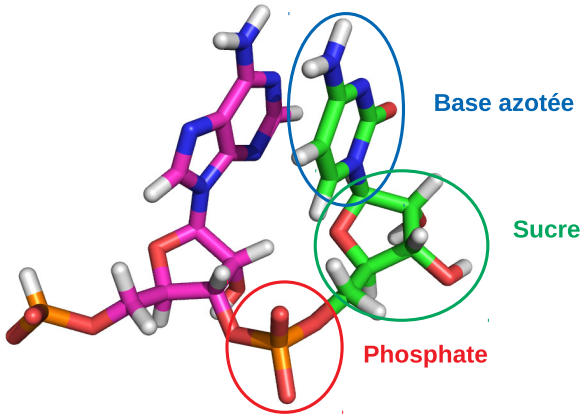


FIGURE 1.2 – Représentation 3D de deux nucléotides (Adénine en rose et cytosine en vert).

1.1.3 La structuration des ARN

Les acides nucléiques (AN) sont des macromolécules qui existent sous forme de polymères, des polynucléotides. Chaque nucléotide est lui-même constitué de trois parties : une base azotée, un monosaccharide à cinq atomes de carbone (nommé sucre dans la suite), ainsi qu'un ou plusieurs groupements phosphates. Dans les ARN, le sucre est oxydé, et il n'y a qu'un seul groupement phosphate par nucléotide. La figure 1.2 montre la représentation 3D de deux nucléotides liés. On y voit notamment les phosphates, les sucres et les bases azotées.

Le polymère est constitué d'une série de nucléotides, liés par des liaisons phosphodiester entre le phosphate et le sucre. Cette chaîne de phosphates-sucres est nommée le *squelette*. Les bases azotées ne font pas partie du squelette de l'ARN. Cette succession de nucléotides constitue la structure primaire de l'ARN : la séquence. Comme seules les bases azotées changent d'un nucléotide à l'autre, on représente la séquence par les lettres représentant les bases azotées. Dans le cas de l'ARN, c'est donc une séquence composée de 4 lettres possibles : A, G, C et U.

Il existe deux familles de bases azotées : les purines (l'adénine A et la guanine G) constituées de deux cycles aromatiques et les pyrimidines (la cytosine C, l'uracile U pour l'ARN et la thymine T pour l'ADN) constituées d'un seul cycle aromatique (voir figure 1.3).

Ces différentes bases azotées peuvent s'apparier entre elles. L'appariement canonique de Watson et Crick implique que A et C interagissent respectivement avec U et G. Il existe d'autres types d'appariement entre bases azotées non abordés dans cette thèse [65]. Une représentation des appariements canoniques est donnée en figure 1.3. L'interaction entre A et U comporte deux liaisons hydrogène, alors que celle entre G et C comporte trois liaisons hydrogène et est donc plus stable. Ces appariements permettent de former des doubles hélices. Ces hélices sont formées par des brins antiparallèles. Dans le cadre de l'ADN, ce sont très souvent deux molécules différentes,

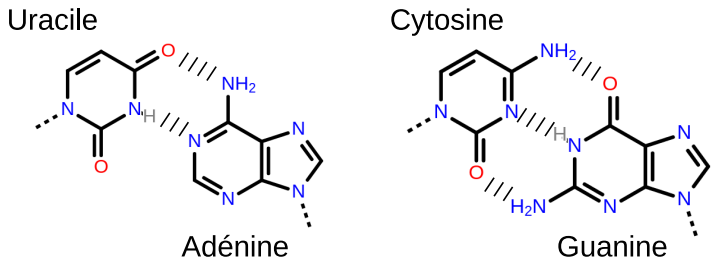


FIGURE 1.3 – Présentation des appariements des bases selon le modèle de Watson et Crick.

parfaitement complémentaires. Au contraire pour l'ARN ce peut être une même molécule qui se replie sur elle-même et forme ces hélices. Cette structure hélicoïdale est une structure très stable du fait des nombreuses liaisons hydrogène ainsi que des interactions d'empilement des bases azotées. Ces appariements entre bases azotées définissent la structure secondaire de l'ARN.

Certaines structures secondaires portent des noms particuliers : les boucles en épingles, les hélices, les bulges, et les boucles internes. On différencie les parties simple brin (ou non appariées) des parties double brin (ou appariées).

Le troisième niveau de structure est la structure tertiaire, qui correspond à la structure 3D de l'ARN. En effet, de nombreuses autres interactions que les appariements peuvent être présentes au sein d'une molécule d'ARN : des liaisons hydrogène entre les bases azotées, les sucres ou les phosphates ou des interactions d'empilements hors hélices, notamment entre les phosphates et les bases azotées. Ces interactions permettent de donner une conformation 3D particulière à l'ARN et cette conformation peut avoir un rôle important pour l'activité biologique de l'ARN. C'est le cas par exemple pour les ARNt qui ont une conformation 3D permettant de faire ressortir 3 nucléotides simple brin particuliers pour la traduction de l'ARNm en protéine.

La figure 1.4 présente ces trois niveaux de structuration des ARN.

Il existe un quatrième niveau de structuration qui n'intervient pas pour chaque ARN : la structure quaternaire. Cette structuration fait intervenir plusieurs chaînes d'ARN qui interagissent entre elles, par des liaisons bases à bases ou par des interactions non séquence spécifique. Le caractère symétrique de cette structuration est présentée dans [40]. Un exemple est présenté en figure 1.5, c'est la structure PDB 1XP7 qui montre deux boucles d'ARN qui forment des interactions base à base (cette structure est nommée *kissing loops*). Cette interaction est essentielle pour la réplication rétrovirale du SIDA notamment.

1.1.4 La structuration des protéines

Quatre types de structuration sont également définis pour les protéines.

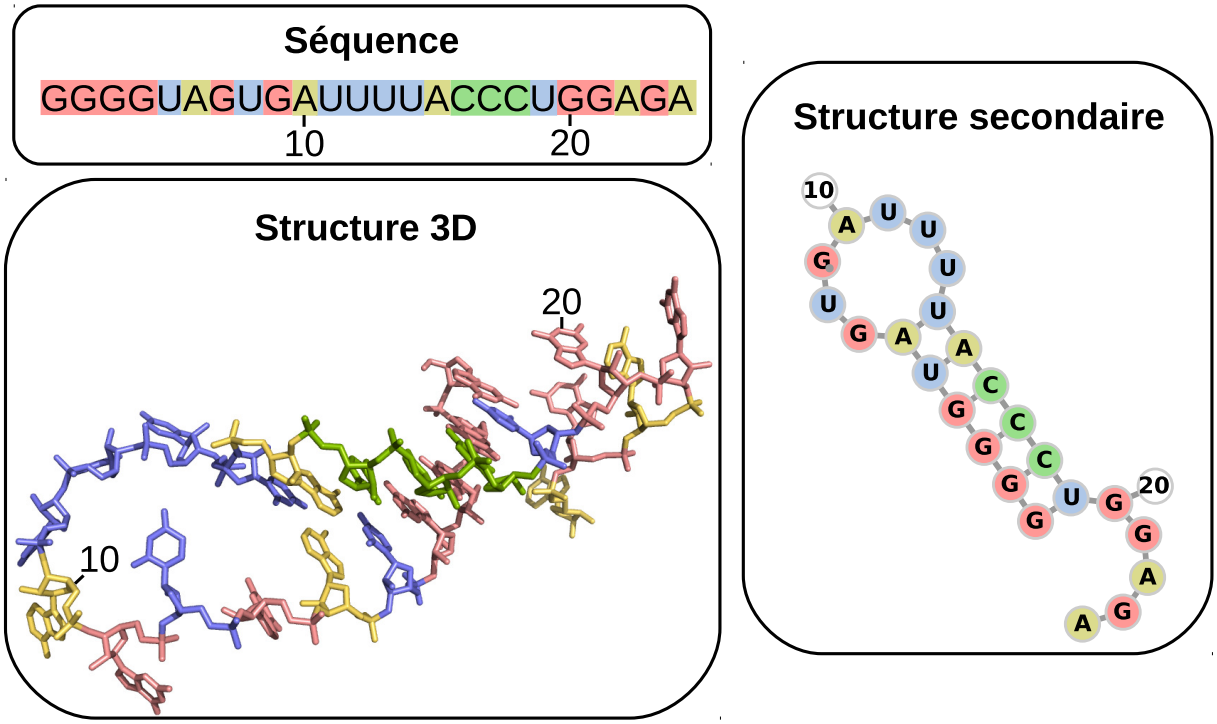
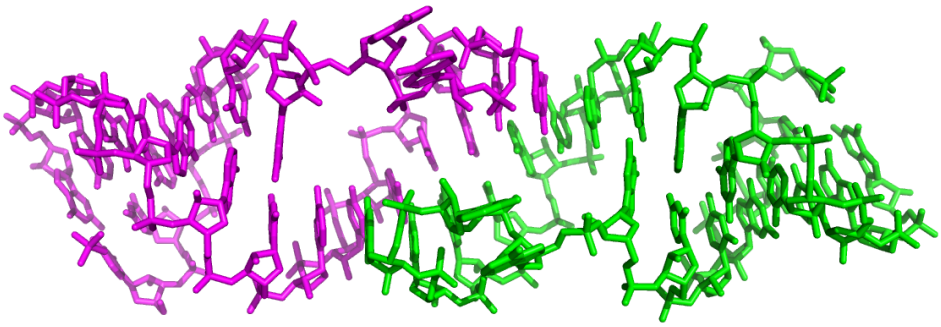


FIGURE 1.4 – La structuration des ARN.

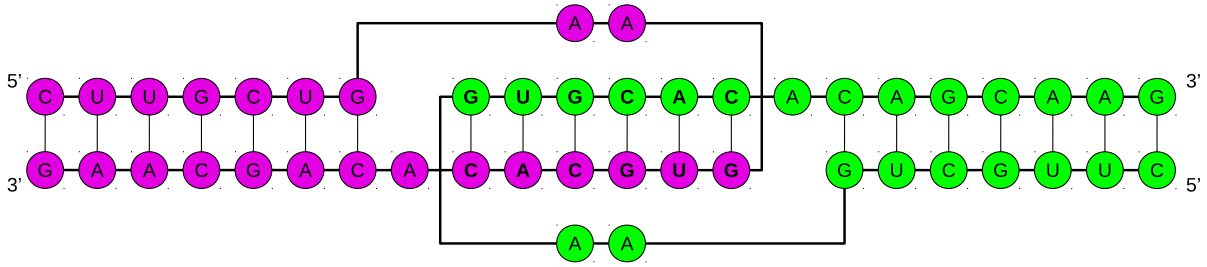
La structure primaire correspond à la séquence d'acides aminés. Les 21 acides aminés possèdent tous une base commune (un carbone α sur lequel sont fixés un groupe amine, un groupe carboxyle, un hydrogène) et l'élément qui les différencie : la chaîne latérale. Les propriétés chimiques de cette chaîne latérale déterminent les caractéristiques de l'acide aminé, et son rôle dans la protéine. Ces acides aminés sont fixés les uns aux autres par une liaison peptidique, dans laquelle la chaîne latérale n'intervient pas. Ainsi, le squelette des protéines fait intervenir le carbone α, les groupements amine et carboxyle.

La structure secondaire des protéines est formée par certains segments enroulés ou pliés de façon répétitive. Ces motifs proviennent des liaisons hydrogènes au sein du squelette de la protéine. Les chaînes latérales n'interviennent toujours pas dans cette structure. Il existe deux grands types de forme particulière induite par ces liaisons hydrogène : les hélices α et les feuillets β. Les hélices α sont des enroulements maintenus en place par des liaisons hydrogène tous les 4 acides aminés. Cette structure implique des acides aminés consécutifs. Les feuillets β sont formés par des brins qui se déploient côte à côte dans un même plan. Ces brins peuvent être parallèles (dans le même sens) ou antiparallèles. Différents brins ne sont pas forcément formés par des acides aminés consécutifs ni par des acides aminés provenant de la même chaîne polypeptidique. Enfin, entre ces différentes structurations, les acides aminés restants forment des boucles flexibles.

La structure tertiaire des protéines, correspondant à l'arrangement 3D des structures secon-



(a) Structure 3D de 1XP7 qui présente deux dimères appareillés.



(b) Structure secondaire de 1XP7 qui présente deux dimères appareillés.

FIGURE 1.5 – Les structures secondaires et quaternaires de 1XP7.

daires, fait intervenir les interactions des chaînes latérales entre elles ou avec le squelette. Ces interactions peuvent être de différentes natures : les liaisons hydrogène, les liaisons ioniques, les interactions de van der Waals, les ponts disulfures et les interactions d'empilements. De manière générale lors du repliement 3D, les acides aminés hydrophobes vont se retrouver à l'intérieur de la protéine, alors que les acides aminés hydrophiles vont se retrouver à la surface. Cela est dû au fait que le milieu cellulaire est un milieu hydrophile.

La structure quaternaire est l'association de plusieurs polypeptides en une protéine. De nombreuses protéines se composent de deux ou trois chaînes polypeptidiques assemblées de façon à former une macromolécule fonctionnelle. Les liaisons entre ces sous-unités peuvent se faire par leurs chaînes latérales, mais aussi par leurs squelettes.

Cette structuration de la protéine lui donne sa forme 3D très importante dans le rôle de son activité biologique. La figure 1.6 présente les quatre niveaux de structurations des protéines ainsi que la liaison peptidique.

1.1.5 Les interactions physico-chimiques entre protéines et ARN

La banque de données des protéines (*Protein Data Bank* : PDB) contient actuellement (le 06 octobre 2022) 10 942 structures résolues experimentallement d'assemblages d'ADN/ARN avec des protéines, incluant une grande variété de systèmes biologiques. Les ribosomes comptent pour approximativement 20% de tous ces assemblages, ce qui laisse environ 8 534 chaînes nucléotidiques non ribosomiques en contact avec une protéine dans la PDB. Ce chiffre est bien inférieur au nombre total de chaînes de protéines (environ 191 996) dans la PDB. Pourtant, environ 16% de toutes les protéines du génome humain (d'après UniProt) contiennent au moins un domaine de liaison à l'ARN. Cela signifie que les AN sont sous-représentés dans la PDB par rapport aux protéines.

Les RBPs peuvent être rangées en plusieurs classes qui interagissent de manières différentes avec l'ARN. Il est cependant possible de décrire des fonctionnements généraux. Les interactions entre protéine et ARN peuvent être spécifiques de la séquence de l'ARN ou non spécifiques. Cette différence permet à une protéine d'interagir avec un ARN très particulier ou au contraire avec un grand nombre d'ARN différents. Les interactions spécifiques impliquent les acides aminés et des bases azotées des nucléotides, principalement via des liaisons hydrogène. Les interactions non spécifiques impliquent la reconnaissance de forme par les repliements globaux de la protéine et de l'ARN. Les phosphates de l'ARN étant chargés négativement, les protéines possédant des acides aminés chargés positivement (ou des patchs de surface chargés positivement) ont tendance a interagir avec le squelette de l'ARN et former des liaisons non spécifiques. Les acides aminés

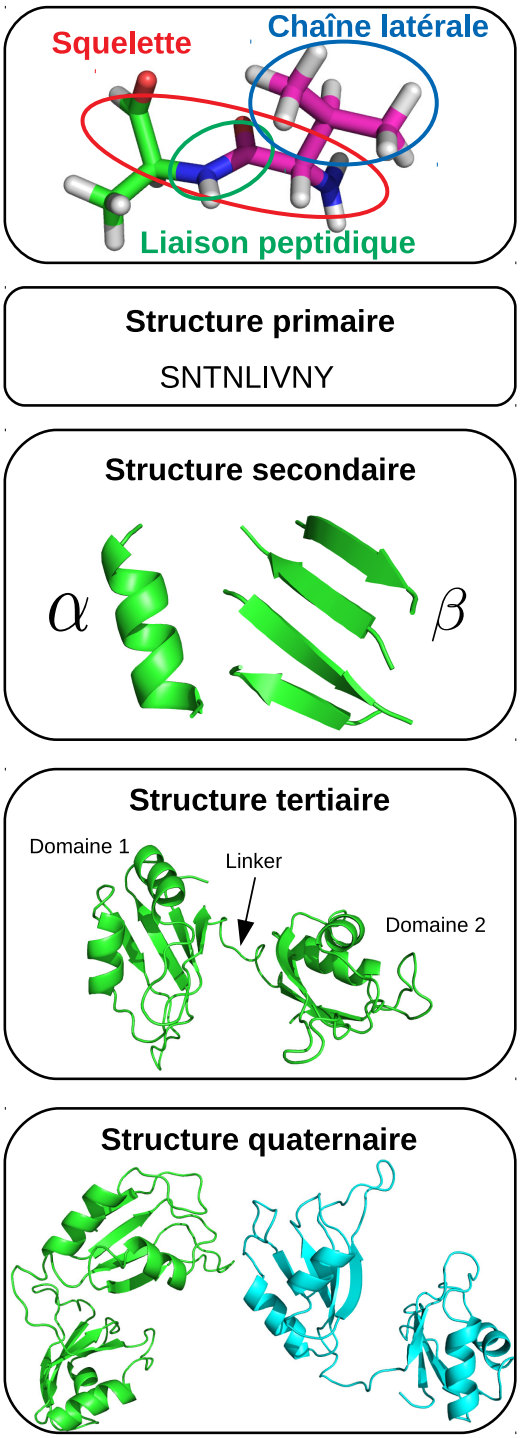


FIGURE 1.6 – De la structure primaire à la structure quaternaire des protéines.

chargés positivement sont l'arginine, la lysine et l'histidine (voir la figure A.1 en annexe). Enfin, les empilements entre acides aminés aromatiques (tyrosine, phénylalanine et tryptophane, voire les différents acides aminés en annexe A.1) et bases azotées sont spécifique du point de vue de la proteine et non spécifique du point de vue de l'ARN.

Les domaines qui lient les ARN sont souvent en contact avec seulement quelques nucléotides, mais il peut exister plusieurs domaines liant les ARN au sein d'une même protéine.

Les différentes catégories de protéines liant les ARN et leurs spécificités structurales sont présentées ci-dessous.

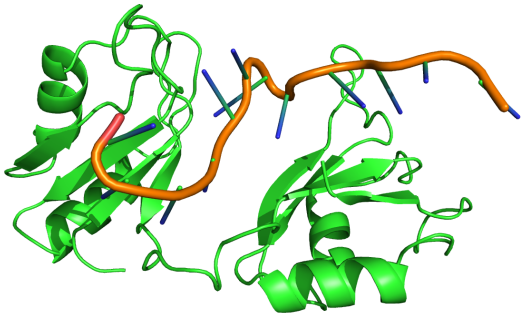
Les *RNA Recognition Motifs* (RRM) : Les RRM sont les domaines d'interactions avec les ARN les plus courants, présents dans 2% de l'ensemble des protéines du génome humain [55]. Un RRM est généralement constitué d'environ 90 acides aminés et contient deux motifs conservés RNP1 et RNP2 constitués respectivement de 8 et 6 acides aminés, dont certains chargés positivement ou aromatiques (permettant donc un empilement). Ces domaines ont une reconnaissance plus ou moins stricte d'une séquence.

***K homology domain* :** Le domaine d'homologie K est un domaine d'environ 70 acides aminés organisés en 2 hélices α ainsi qu'un feuillet β. Cette structure particulière permet d'interagir avec 4 nucléotides de manière spécifique à la séquence.

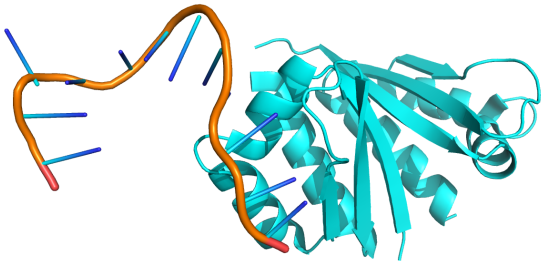
***Double-stranded RNA binding domain* :** Ces domaines sont constitués de 65 à 70 acides aminés organisés en 2 hélices α et en 3 feuillets β. Du fait de la forme particulière des hélices d'ARN, ces domaines interagissent majoritairement avec les squelettes de l'ARN, ce qui rend ces interactions très peu séquence spécifique.

***Pumilio homology domain* :** Ce domaine consiste généralement en une succession de plusieurs motifs de 36 acides aminés, chacun de ces motifs interagissant avec un nucléotide. Le domaine complet forme une structure courbée permettant l'interaction avec 8 à 11 nucléotides [88]. Du fait de la très bonne compréhension de l'interaction entre ces domaines et les ARN, il est possible de créer artificiellement ces domaines pour lier des séquences spécifiques de l'ARN.

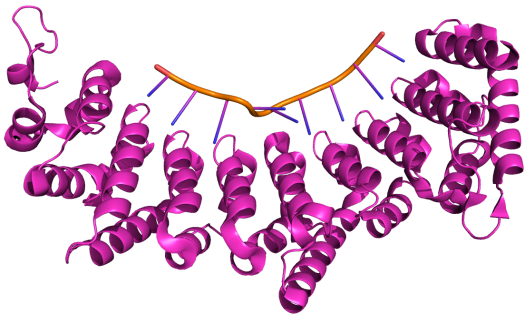
Les doigts de Zinc Cette classe regroupe une grande diversité de domaines qui ont pour point commun de faire intervenir le zinc. Le génome humain code pour environ 60 protéines avec des doigts de zinc. Ces domaines peuvent se lier à l'ARN, mais aussi à l'ADN ou à des protéines. Les interactions entre l'ARN et la protéine sont des empilements et des liaisons hydrogènes avec le



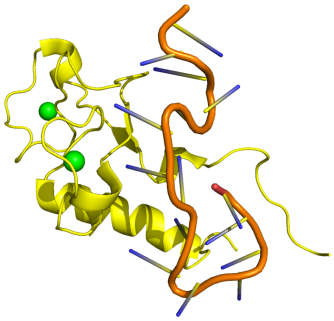
(a) Exemple d'un RRM : 1B7F (sex-lethal et poly-U)



(b) Exemple d'un KH : 2PY9 (human poly-C binding)



(c) Exemple d'une pumilio : 3K62



(d) Exemple d'un doigt de Zinc (le zinc en vert) : 5U9B (human cardiac troponin T)

FIGURE 1.7 – Exemples de RBPs, avec ici leurs interactions avec des ARN.

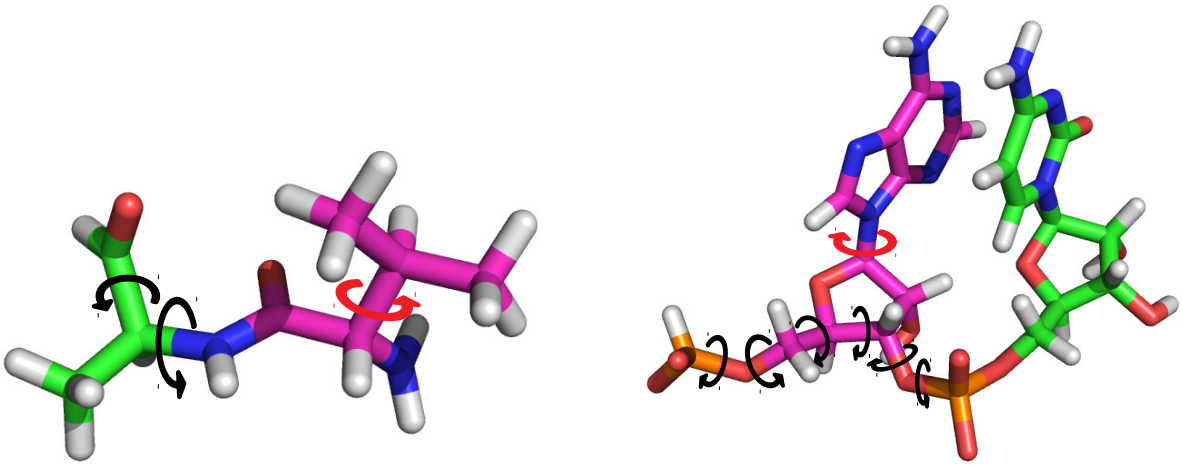
squelette de la protéine, ce qui rend ces interactions peu spécifiques à la séquence de l'ARN.

1.1.6 La flexibilité des macromolécules

Les structures 3D des protéines et des ARN sont flexibles et adoptent de nombreuses conformations dans la cellule. Dans le cadre des interactions, ces flexibilités jouent un rôle, qui peut être pris en compte.

La flexibilité de l'ARN et des protéines

Les liaisons covalentes des macromolécules pouvent etre considerées comme de longueur fixes aux échelles de modélisation qui nous intéressent, ainsi que les angles entre 3 atomes (voir la section 3.5.4). Les degrés de liberté sont alors les angles dièdres, ou angles de torsion formés par 4 atomes consécutifs. Les degrés de liberté du squelette déterminent la flexibilité globale des macromolécules. Le squelette des protéines comporte deux degrés de liberté par acide aminé, contre six par nucléotide pour l'ARN, ce qui fait de ce dernier une molécule beaucoup plus flexible. Les degrés de liberté des chaines laterales déterminent quant a eux une flexibilité locale.



(a) Degrés de liberté associés à un acide aminé. (b) Degrés de liberté associés à un nucléotide.

FIGURE 1.8 – Les flèches noires symbolisent les rotations possibles au sein du squelette pour une unité (nucléotide ou acide aminé). Les flèches rouges symbolisent les rotations d'une base azotée ou d'une chaîne latérale d'un acide aminé.

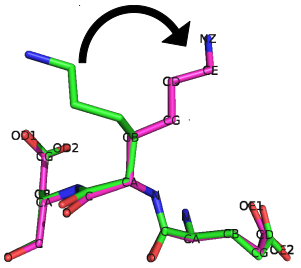
Ce sont les rotations des bases azotées autour de leur liaison avec le ribose dans les ARN, et les rotations des chaînes latérales autour de leur liaison avec le squelette ou les rotations au sein des les chaînes latérales longues pour les protéines. Ces différentes rotations sont présentées dans la figure 1.8.

Plusieurs amplitudes de flexibilité sont définies, de la plus faible à la plus forte (voir figure 1.9) :

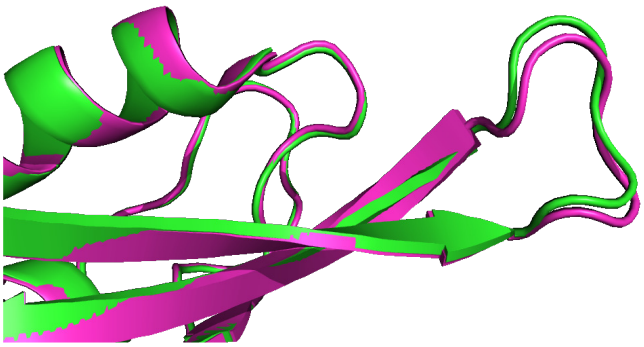
— les chaînes latérales ou les bases azotées (amplitude de 1-5 Å) ;
— les structures secondaires (amplitude de 1-5 Å) ;
— les boucles (amplitude de 1-10 Å) ;
— les domaines (amplitude $>$ 5 Å).

La flexibilité des complexes ARN-protéine Dans un assemblage entre un ARN et une protéine, ces deux macromolécules ont une structure particulière, adaptée à leur interaction, nommée la forme liée. Lorsque ces molécules n'interagissent pas, elles sont sous forme libre. Ces formes liées et libres peuvent avoir des structures très proches ou très éloignées selon les cas (voir figure 1.10). Cela sera discuté plus en détail dans la suite.

Il existe deux modèles pour expliquer les changements de conformation 3D des partenaires entre forme libre et forme liée. Le premier est appelé la sélection conformationnelle. Les molécules étant flexibles, elles adoptent plusieurs conformations à l'état libre, et certaines peuvent interagir par complémentarité de forme et électrostatique alors que d'autres ne peuvent pas. Ainsi, quand

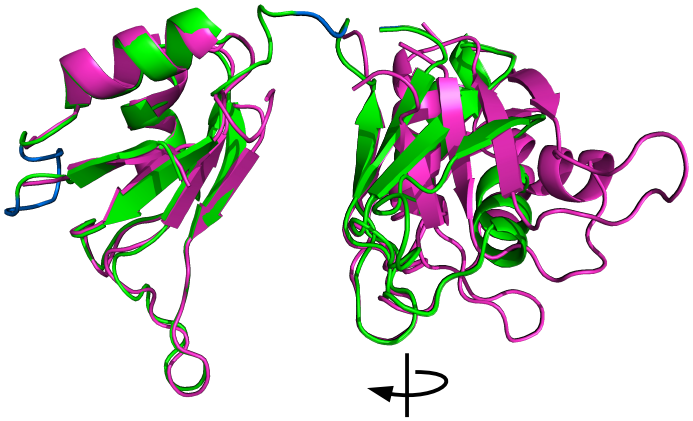


(a) Mouvement d'une chaîne latérale de protéine.



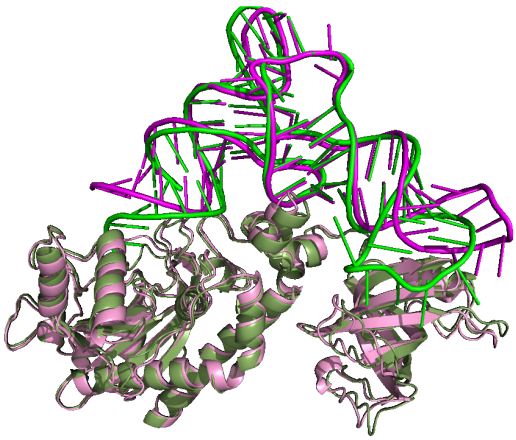
(b) Mouvement d'une boucle au sein d'une protéine.

(c) Mouvement d'une structure secondaire d'une protéine.

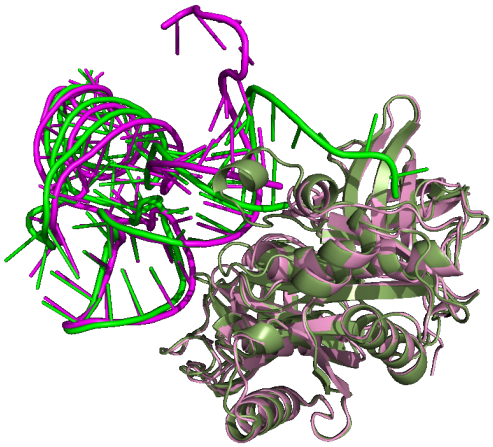


(d) Mouvement d'un domaine complet d'une protéine.

FIGURE 1.9 – Exemples de la flexibilité au sein des protéines.

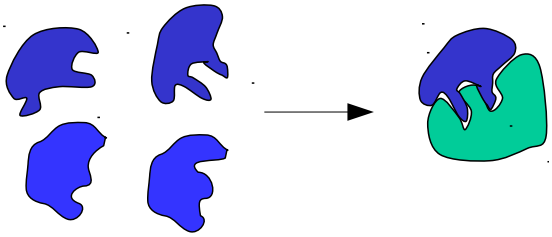


(a) Exemple de la structure 1ASY (structure d'une aminoacyl transferase en complexe avec un ARNt), dans ce cas il y a peu de différences entre la forme liée et la forme libre.

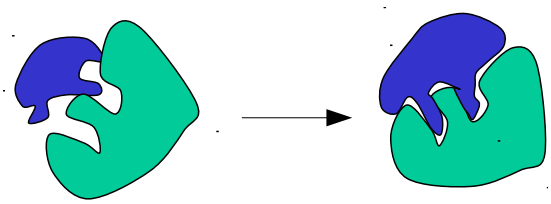


(b) Exemple de la structure 1B23 (structure d'un facteur d'élongation en complexe avec un ARNt), dans ce cas la forme liée de l'ARN a sa partie simple brin orientée différemment de la forme libre.

FIGURE 1.10 – Exemples de structures liées (vertes) et libres (roses) d'ARN.



(a) Représentation de la sélection conformationnelle, plusieurs conformations de ligands existent dans la cellule et celui ayant la bonne forme se lie avec le récepteur.



(b) Représentation de l'induced fit, le ligand se déforme au contact du récepteur en se liant.

FIGURE 1.11 – Les deux représentations de la flexibilité des macromolécules.

les deux partenaires ont la bonne conformation, il y a appariement. Le deuxième modèle est nommé *induced-fit*. C'est l'interaction entre les partenaires qui les fait adopter leur conformation correspondant à la structure du complexe.

En réalité, c'est plutôt un mélange de ces deux modèles qui intervient : lorsque les conformations sont proches de celles complémentaires, les deux molécules s'apparient, puis du fait de l'appariement des changements de conformations renforcent les liaisons entre les deux molécules.

Une hypothèse importante en biologie structurale est que l'activité biologique d'un complexe a lieu à son minimum d'énergie globale. Il est aussi considéré que les structures obtenues expérimentalement sont à ce minimum d'énergie globale. C'est donc cette structure que l'on cherche lors de la modélisation d'un complexe. Le fait qu'un complexe soit à son minimum d'énergie glo-

bale ne signifie cependant pas que chacun des partenaires est à son minimum d'énergie globale. En effet, l'énergie d'interaction de l'appariement peut compenser l'augmentation de l'énergie conformationnelle d'un ou des partenaires par rapport à sa forme libre. C'est pourquoi il est difficile de prédire une structure liée avant de prédire la structure globale du complexe.

1.2 Les Méthodes expérimentales en biologie structurale

Des méthodes expérimentales permettent d'observer les structures des macromolécules. Ces structures expérimentales sont regroupées dans la PDB, la plus grosse base de données de structures existante. Cette base de données contient à ce jour (le 06 octobre 2022) 192095 structures. Parmi ces structures, 14495 contiennent des AN. Les principales méthodes sont présentées ici.

1.2.1 La cristallographie

La cristallographie aux rayons X est une méthode d'analyse fondée sur la diffraction des rayons X par des cristaux. Le principe est d'envoyer sur un cristal de macromolécules un faisceau de rayons X qui sont renvoyés dans des directions déterminées par la position des atomes dans le réseau cristallin et par la longueur d'onde des rayons X. La mesure des angles et de l'intensité des rayons diffractés permet d'obtenir un maillage tridimensionnel. À partir de ce maillage, il est possible d'avoir la position moyenne des atomes du cristal. La qualité du maillage donne la résolution.

Cette méthode permet d'obtenir des structures cristallines de très bonne résolution (1-3 Å), cependant, il faut que le complexe ARN-protéine d'intérêt soit cristallisable. Cela peut s'avérer très compliqué dans le cadre de protéines membranaires par exemple, ou pour des molécules très flexibles (pour lesquelles la cristallisation ne fonctionne pas).

De plus, cette méthode peut créer des artefacts de cristallisation, notamment le cas d'un contact entre protéine et ARN qui n'existe pas forcément dans la cellule. Les structures obtenues peuvent être annotées par les biologistes pour expliquer ces possibles contacts. Le complexe présenté après une curation est appelé assemblage biologique.

Enfin, les structures obtenues peuvent être différentes de celles existant dans le vivant. En effet, l'arrangement cristallin n'est pas forcément celui d'énergie globale minimale *in vivo*.

1.2.2 La résonance magnétique nucléaire

La Résonance Magnétique Nucléaire (RMN) est une autre méthode d'analyse des structures couramment utilisée. Elle est basée sur la propriété de certains noyaux atomiques possédant un

spin nucléaire. Ces noyaux peuvent absorber l'énergie d'un rayonnement électromagnétique et le restituer selon une fréquence très précise qui dépend de l'environnement atomique. La mesure de ces fréquences permet la reconstruction de la structure des molécules.

Cette méthode est très efficace dans le cadre des petites molécules (moins de 15 kDa). Cependant, dans le cadre de larges protéines ou d'ARN, les fréquences deviennent trop mélangées pour pouvoir les interpréter, et la méthode n'est donc pas utilisable.

À l'inverse de la cristallographie, cette méthode ne nécessite pas de cristal, ce qui permet d'obtenir les structures d'objets peu structurés ou très flexibles. En effet, la RMN fonctionne en milieu soluble, il est donc possible d'obtenir de nombreuses structures d'un même objet (molécule ou complexe) et donc d'en comprendre la flexibilité.

1.2.3 Cryo-ME

La Cryo-Microscopie Électronique (Cryo-ME) consiste à congeler les échantillons dans de l'éthane liquide, obtenant ainsi une glace non cristalline. Une projection d'électrons permet ensuite d'obtenir des projections 2D des objets congelés, et ces images peuvent être assemblées pour obtenir la structure 3D des objets.

Cette méthode reste récente et en pleine expansion. Il est maintenant possible d'avoir des structures 3D avec une meilleure résolution que la cristallographie. Cette méthode utilise des échantillons plus simples à obtenir (puisqu'en solution) et permet d'observer les structures en solution plutôt que sous la forme de cristal, ce qui donne un grand nombre de conformations comme résultat. Elle permet aussi d'obtenir des structures de grosses molécules, de leurs assemblages et de leurs changements de conformations (par exemple les transitions du fonctionnement biologique d'un complexe).

Cependant, même si la préparation des échantillons est plus simple, la postproduction des résultats est bien plus complexe et les machines coûtent encore très cher.

1.3 Les méthodes computationnelles

Les différentes méthodes expérimentales ont presque toutes leurs avantages ainsi que leurs inconvénients. Cependant, deux des inconvénients sont communs à toutes ces méthodes : le temps qu'elles demandent et leur coût. En effet, c'est plusieurs semaines voir mois/années de travail pour obtenir expérimentalement une structure d'un complexe ARN-protéine. Ces méthodes impliquent des machines complexes à utiliser et donc du personnel qualifié, et peuvent demander de gros investissements.

C'est pourquoi les méthodes computationnelles qui sont présentées ci-dessous existent pour prédire les structures des complexes macromoléculaires. Ceci étant, les méthodes expérimentales et informatiques restent très complémentaires, la partie expérimentale venant vérifier les résultats obtenus de manière informatique, et les méthodes expérimentales fournissant des informations permettant de définir des contraintes pour réduire l'espace de recherche par les méthodes informatiques.

Dans cette partie, nous verrons les méthodes computationnelles pour la prédiction des structures d'ARN et de protéine, et surtout des complexes ARN-protéines.

1.3.1 Le clustering et l'échantillonnage de structures 3D

Les différentes méthodes computationnelles peuvent nécessiter d'échantillonner des structures d'intérêts, afin d'obtenir toutes les conformations possibles ou observées d'une (fraction de) molécule donnée. Pour réaliser cet échantillonnage, il existe plusieurs méthodes. La première que nous présentons consiste à échantillonner les conformations déjà observées expérimentalement, par exemple dans la PDB. L'échantillonnage des conformations peut aussi être réalisé en calculant les conformations possibles *de novo*, par exemple en utilisant des combinaisons de valeurs d'angles autorisées, pour obtenir des conformations qui ne sont pas observées mais qui restent biologiquement plausibles.

Il peut être utile de sélectionner un certain nombre de représentants de ces conformations, afin de couvrir l'espace conformationnel de façon homogène et/ou de limiter le nombre de conformèree. Le clustering est classiquement utilisé pour cette étape. Parmi les nombreuses méthodes de clustering existantes, les suivantes sont utilisées en biologie structurale :

— La méthode k-means permet de sélectionner le nombre de représentants que l'on veut, et réalise un clustering basé sur la minimisation des variabilités inter-clusters [33].
— La méthode star-shape présentée dans Gromacs [66] permet de sélectionner un seuil de telle manière que l'ensemble des clusters aient un rayon d'au plus la valeur seuil.
— Les Self Organizing Maps (SOM) sont quant à elle souvent utilisées pour réaliser des clusters des différentes étapes de dynamiques moléculaires [54]. Les clusters obtenus sont basés sur la densité des conformations.

Ces méthodes de clustering sont basées sur des calculs de distances entre conformations. Généralement c'est la RMSD qui est utilisée, mais il est possible de la remplacer par des calculs de distances entre coordonnées internes par exemple.

L'échantillonnage des conformations peut aussi être réalisé en calculant les conformations possibles, en utilisant des valeurs d'angles par exemple, cela permet d'obtenir des conformations

qui ne sont pas observées mais qui resteraient biologiquement viables.

1.3.2 Les représentations discrètes

L'espace de recherche qui est exploré lors de la modélisation de la structure d'une macromolécule a pour dimension $3 \times N$, N étant le nombre d'atomes. En réalité, cette dimension est réduite par les contraintes géométriques qui régissent cette structure, à commencer par les liaisons covalentes (au minimum $N - 1$) de longueur et angles fixes (aux échelles de modélisation qui nous intéressent). Les principaux degrés de liberté (DL) sont les angles de torsion formés par 4 atomes consécutifs. Dans les protéines et ARN, les angles de torsion du squelette influent sur la structure globale, et représentent 2 et 6 DL par acide-aminé/nucléotide. Les angles de torsion des chaînes latérales influent sur la structure locale, et varient en fonction du type de résidu : de 0 à 5 dans les protéines, et 5 (4 couplés dans la sucre et une rotation de la base) dans les ARN. L'espace de recherche peut être encore réduit en considérant seulement les intervalles de valeurs autorisés ou favorables pour certains angles. En effet, il a été montré que les acide-aminés [21] comme les nucléotides sont rotamérique, c'est-à-dire que leurs conformations peuvent être approximées par un ensemble fini de conformations types, appelées rotamers. Ce type de représentation discrète est largement utilisé pour la modélisation des protéines et ARN dans divers contextes, par des approches combinatoires : prédiction de structure *ab initio* (ARN [24, 92], protéine [34]), affinement de structure 3D (ARN [2], protéine [91]), étude du processus de repliement (ARN [22]), design (peptides [73], protéines [35]). Des représentations en fragments de plusieurs résidus, de diverses tailles, sont également largement utilisées, ce que nous détaillons ci-après. Ces représentations permettent de réduire encore la dimensionnalité du problème de prédiction (ou design) de structure, mais pose le problème de la construction d'une bibliothèque de fragments adéquate au problème abordé.

1.3.3 Les prédictions de structures

L'évaluation des méthodes de prédiction des structures 3D se fait sur des cas tests en comparant les résultats de la méthode et la structure expérimentale connue (considérée comme celle de plus faible énergie et correspondant à la fonction biologique). Le plus souvent, la mesure utilisée pour évaluer cette différence est la *Root Mean Squarred Deviation* (RMSD), avec n le nombre d'atomes et $x_{j,i}$ les coordonnées cartésiennes du $i^{ème}$ atome de la $j^{ème}$ structure :

$$\text{RMSD} = \sqrt{\frac{\sum_{i=1}^{n}(x_{1,i} - x_{2,i})^2}{n}} \tag{1.1}$$

La RMSD est une mesure de déviation entre deux structures 3D qui s'exprime en Angström (Å $= 10^{-10}$ m). Cette mesure peut être calculée après superposition des deux structures. Cela signifie que des rotations/translations globales sont appliquées à une des deux molécules pour minimiser la RMSD entre les deux molécules : dans ce cas, c'est le RMSD conformationnel (cRMSD).

Prédiction des structures de l'ARN

Le principe de la prédiction de la structure secondaire est de chercher les jeux d'appariements entre bases azotées qui minimisent l'énergie interne de l'ARN. Les outils les plus connus sont dans le paquet ViennaRNA [50], mais il existe de nombreuses méthodes recensées dans [94]. Cependant, même si cette solution est fiable dans de nombreux cas, la prédiction de la structure secondaire reste un défi pour les longs ARN et du fait de structures très particulières qui sont difficiles à prédire, tels que les *pseudo-knots* ou les *G-quadruplexes*.

Dans la suite de la thèse, ces défis ne jouent pas un grand rôle. En effet, les structures qui seront présentées sont les épingles à cheveux, or ces structures sont globalement très bien prédites.

La prédiction de la structure secondaire est souvent la première étape pour la prédiction 3D des ARN. Dans le cadre de la prédiction de la structure 3D, les régions simple brin posent une difficulté ; car elles sont trop flexibles pour avoir une structure 3D fixe au-delà de quelques nucléotides.

Une expérience nommée RNA-Puzzles [11] est organisé tous les 2 ans (4 itérations jusque aujourd'hui) pour permettre aux dernières méthodes d'être appliquées sur des séquences dont la structure a été résolue par des biologistes, mais pas encore publiée. L'idée est de réaliser un classement des méthodes de prédiction de la structure 3D des ARN. Parmi les méthodes qui se sont démarquées, FARFAR [13] a donné de très bons résultats en s'appuyant sur la reconstruction à partir de structures homologues. Cette reconstruction superpose des fragments sur les structures homologues selon une fonction de score qui récompense les appariements base à base et les empilements. Puis un affinement est appliqué sur la structure finale, avec cette fois la prise en compte des interactions physico-chimiques décrite en section 1.1.2. Une nouvelle version accessible en ligne nommée FARFAR2 [89] propose des benchmarks pour l'utilisation de cette méthode. Une limite de ces méthodes reste la capacité à discriminer la bonne solution du grand nombre de modèles obtenus (entre 1000 et 100 000 modèles). Actuellement ce sont jusque 10 clusters qui sont formés et les centres représentent les solutions proposées.

Une méthode parue récemment et basée sur le *Deep Learning* nommée ARES [82] propose une première solution pour l'évaluation des modèles obtenus par FARFAR2. Cette méthode propose

de très bons résultats sur les défis proposé par RNA-Puzzle (81 % du benchmark pour lequel une solution dans le top 10 est à moins de 5 Å de la structure native). Cette avancée qui repose sur l'intelligence artificielle s'est inspirée des résultats obtenus pour la prédiction des structures de protéines présentée dans la partie suivante.

Ces méthodes basées sur le principe de l'assemblage de fragments ont ensuite été adaptées pour la modélisation 3D de l'ARN. Pour des prédictions ab initio, FARNA [13, 14] correspond à la méthode ROSETTA appliquée à l'ARN. FARNA assemble une structure 3D d'ARN à partir de courts fragments linéaires issus de la grande sous-unité ribosomale de *Haloarcula marismortui* (code PDB : 1ffk). Il utilise une fonction d'énergie basée sur la connaissance, qui prend en compte les préférences des conformations du squelette et des chaînes latérales, et des interactions d'appariement et d'empilement de bases, dérivées de structures d'ARN déterminées expérimentalement. FARFAR est une extension de FARNA avec un affinement tout-atome. MC-Fold/MC-Sym [67] est basé sur un principe apparenté à FARNA : il assemble des structures d'ARN à partir d'une bibliothèque de 'motifs cycliques de nucléotides', c'est-à-dire des fragments dans lesquels tous les nucléotides sont connectés circulairement par des interactions covalentes, d'appariement ou d'empilement. RNACOMPOSER [71] se base sur une prédiction de structure secondaire et le dictionnaire FRABASE [72]. Pour des predictions template-based, ModeRNA [77] est inspiré de SWISS-MODEL coté protéines : il interprète un alignement de séquences comme un ensemble d'instructions utilisées pour créer un modèle en copiant la partie conservée d'une structure modèle et en introduisant dans les parties variables des fragments issus de structures expérimentales.

Prédiction des structures de protéines

Le principe de la prédiction de la structure des protéines est de chercher le repliement 3D le plus probable d'une séquence donnée. Ce challenge est complexe et comporte de nombreuses difficultés, notamment dues à la flexibilité des protéines, et donc au très grand nombre de conformations à prendre en compte. Les premières approches utilisées étaient basées sur les homologies entre protéines de séquences proches [48]. Les approches se focalisent en général sur la prédiction du squelette, qui définit la structure globale, et laisse de côté le placement des chaînes latérales.

Un concours nommé *Critical Assessment of Structure Prediction* (CASP) propose la même chose que RNA-Puzzles mais pour les protéines. Il est organisé tous les 2 ans lui aussi, avec 14 itérations à aujourd'hui.

Au cours de ma thèse a eu lieu CASP14, qui a fait beaucoup de bruit dans la communauté du fait des très bons résultats d'Alphafold2, un outil développé par DeepMind [42]. Ces résultats sont montrés en figure 1.12, et on peut observer qu'Alphafold2 semble avoir des années d'avance

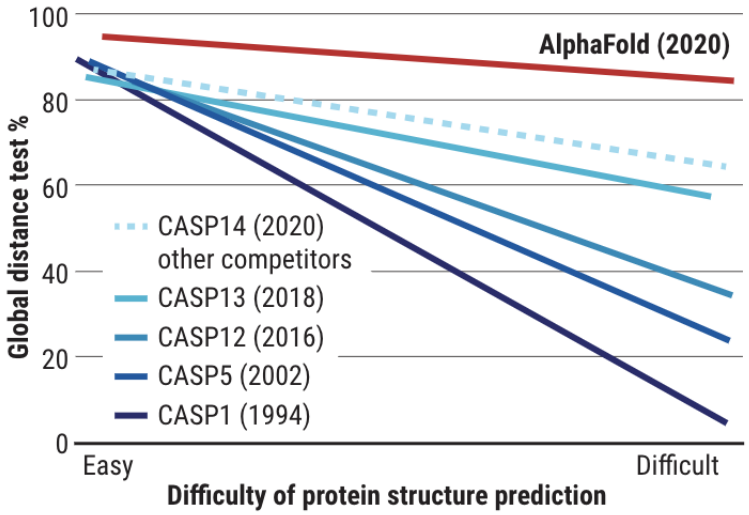


FIGURE 1.12 – Vue des résultats sur les différentes années de CASP en comparaison avec les résultats de Alphafold2.

sur les résultats de CASP14 (par rapport à la tendance qui montre une amélioration des résultats à chaque nouvelle instance de CASP).

La révolution provoquée par Alphafold (en 2018 à CASP13) est basée sur l'utilisation du *Deep Learning* et plus précisément sur l'utilisation de réseaux de neurones convolutifs (*Convolutional Neural Network* : CNN). Le deep learning est efficace pour résoudre des problèmes structurés et dont la structure est hiérarchique. C'est ce qu'a utilisé DeepMind pour la prédiction de structure 3D des protéines. Les informations de départ sont des alignements de séquences multiples. À partir de ces alignements sont extraites les co-évolutions entre acides aminés, qui donnent des informations sur la proximité de ces acides aminés dans la structure 3D. À partir de ces co-évolutions, une carte des distances inter-résidus est approximée et la structure est prédite à partir de cette carte. Alaphafold2 a ajouté le mécanisme d'attention à son réseau de neurones, ce qui a grandement amélioré les résultats.

Le fonctionnement d'Alphafold2 repose donc sur la connaissance de séquences de protéines proches de la séquence initiale, et laisse donc une faille sur la prédiction de structure d'une séquence complètement nouvelle. Cependant avec les séquençages automatiques, de plus en plus de séquences sont disponibles, et ce problème ira diminuant. Les protéines désordonnées sont une autre faille de la méthode : pour ces protéines c'est plutôt un ensemble de conformations possibles qui est recherché (là où Alphafold prédit une structure moyenne). Enfin, cette méthode de repliement protéique est aussi utilisée dans le cadre du docking (présenté juste après), et obtient les meilleurs résultats parmi toutes les méthodes existantes [8]. Même si ces résultats ne

s'appliquent pas (encore) au docking ARN-protéine, et donc n'ont pas influencé directement ma thèse, ils vont influencer le domaine de la biologie structurale sur le long terme.

Il a été annoncé que DeepMind travaille actuellement sur le repliement 3D des AN. L'entraînement de leur méthode sur les ARN posera des difficultés différentes de celles résolues pour les protéines. D'une part il y a beaucoup moins de structures d'ARN ou d'ADN disponibles actuellement que de structures de protéines (comme discuté section 1.1.5). D'autre part la structuration de l'ARN est particulière au vu de celle des protéines, les G-quadruplexes ainsi que les pseudo-knots sont difficiles à prédire pour les structures secondaires. Ces structures de G-quadruplexe sont peu nombreuses dans la PDB, ce qui limite l'application du deep learning. Ces difficultés seraient d'autant plus importantes pour une prédiction directe par deep learning de structures des complexes ARN-protéine.

Cependant, cette méthode de prédiction très précise permet d'avoir accès à des structures de protéines que nous n'avions pas jusqu'alors. Ces structures de protéines pourraient être utilisées, en l'absence d'une structure non liée, dans nos méthodes de docking ARN-protéine. Et à l'avenir les structures des ARN prédites par deep learning pourraient également être utilisées pour le docking des complexes ARN-protéine. Cependant cela ne résout pas le docking des parties simple brin de l'ARN qui peuvent changer radicalement de conformation entre les formes liées et non liées.

Les méthodes ab initio consistent à prédire la structure native d'une protéine à partir de sa seule séquence d'acides aminés. Le coût de calcul pour échantillonner et évaluer un paysage énergétique très complexe afin d'identifier le minimum global est trop élevé, même pour les petites protéines. De nombreuses méthodes de prédiction de structure de protéine performantes telles que ITASSER [93] et ROSETTA [76] utilisent des bibliothèques de fragments pour améliorer la précision et réduire le nombre de degrés de liberté pendant la recherche conformationnelle. Elles exploitent le principe selon lequel les séquences locales d'acides aminés favorisent certaines structures locales, par leurs propriétés telles que la charge, la présence de résidus aromatiques et l'hydrophobicité [30]. La structure de segments de séquence cible peuvent ainsi être modélisée à l'aide de fragments d'autres protéines non homologues. Une fois qu'une insertion ou un déplacement de fragment a été effectué par l'algorithme d'échantillonnage, il est soit accepté soit rejeté, en fonction de son impact sur l'énergie ou le score du système. Les bibliothèques de fragments sont également utilisées pour la prédiction de conformations de boucle dans des protéines de structures connues par ailleurs (structure experimental incomplete, ou modele par homologie [75, 44]. Ces régions montrent la plus grande diversité conformationnelle, et la prédiction de leurs conformations est particulièrement utile car elles sont principalement à la surface des protéines

et sont donc fréquentes dans les sites de liaison.

1.3.4 Le docking

Les méthodes de docking ont pour but de prédire la structure d'un complexe de macromolécules, en général à partir de la structure des molécules constituant le complexe. Dans le cadre de la thèse, je me suis intéressé au docking entre deux molécules, et plus particulièrement entre une protéine (récepteur) et un ARN (ligand).

Le principe

Un docking est effectué quand l'hypothèse est faite qu'il y a une interaction entre deux molécules. Le docking ne permet pas directement de savoir s'il y a interaction entre les deux molécules. Quoiqu'il se passe, la méthode trouvera une structure de complexe.

Un postulat est que la structure minimisant l'énergie globale donne la structure active biologiquement. L'approche courante est de trouver la structure d'un complexe de deux molécules minimisant l'énergie d'interaction : on néglige souvent l'énergie du changement conformationnel des partenaires. En pratique, on a seulement accès à une approximation de l'énergie d'interaction (champ de force ou fonction de score).

Le docking se compose en général de deux étapes. La première est l'étape d'*échantillonnage* de positions du ligand sur le récepteur. La deuxième est l'étape d'*évaluation* des différentes solutions. L'objectif est d'avoir un échantillonnage suffisamment exhaustif pour explorer des minimums locaux d'énergie qui comprennent le minimum global, puis d'avoir une évaluation suffisamment performante pour différencier les minimums locaux du minimum global.

La figure 1.13 représente schématiquement en une dimension le paysage énergétique correspondant à la pose d'un ligand sur un récepteur.

Échantillonnage

Il existe plusieurs méthodes d'échantillonnage. Les plus connues d'entre elles sont présentées ici.

La *Fast Fourier Transform* (FFT) est historiquement l'un des premiers algorithmes capables de réaliser les docking les plus rapides. Une solution d'accélération par FFT consiste à discrétiser la forme 3D de chaque protéine sur une grille (en une partie "intérieure" et une "surface", voir figure 1.14), puis évaluer simultanément toutes les translations possibles en termes de complémentarité de forme. Cette évaluation doit être répétée pour chaque rotation (3 degrés de

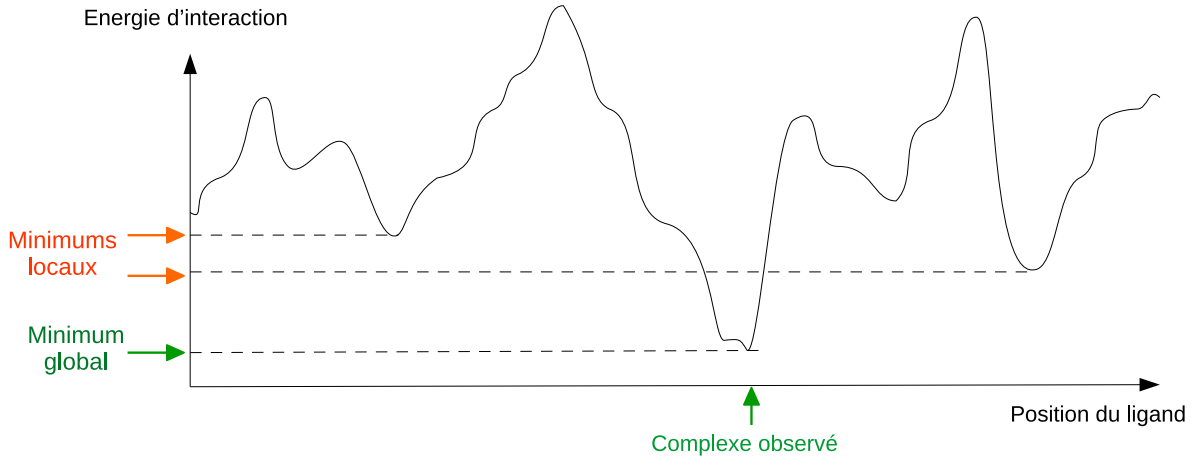


FIGURE 1.13 – Représentation d'un paysage énergétique correspondant à l'interaction entre un ARN et une protéine. On observe des minimums locaux ainsi que le minimum global (qui correspond à la structure expérimentale observée).

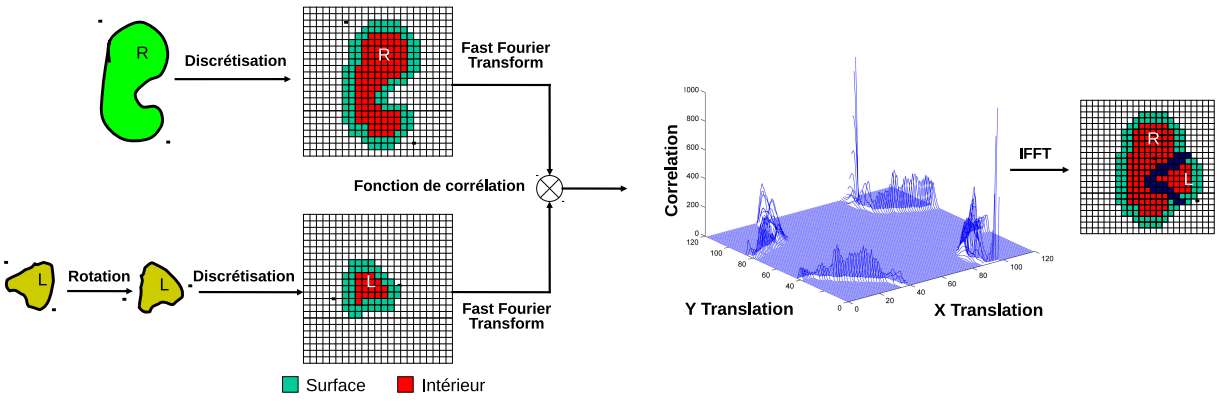


FIGURE 1.14 – Schéma sur le fonctionnement de la FFT.

liberté) dans un espace discrétisé. Ainsi, cette méthode est très rapide, mais avec une perte de précision importante. Une représentation particulière des protéines en harmoniques sphériques est proposée dans l'outil Hex créé par Dave Ritchie en 2000. L'accélération des méthodes utilisant des représentations de protéines à haute précision dans la dernière décennie a réduit l'intérêt de l'accélération par FFT par rapport à la perte de précision.

Les méthodes haute résolution explorent le paysage énergétique de manière continue. Pour se déplacer sur ce paysage, deux solutions : la descente de gradient ou la méthode Monte-Carlo. L'idée de la descente de gradient est de répéter de manière itérative le calcul du gradient à une position donnée puis de descendre d'un pas selon ce gradient. Pour cela, il faut que l'approximation de l'énergie d'interaction soit dérivable. Dans la méthode Monte-Carlo, un déplacement aléatoire est appliqué. Si l'énergie d'interaction diminue alors le changement est gardé, si elle

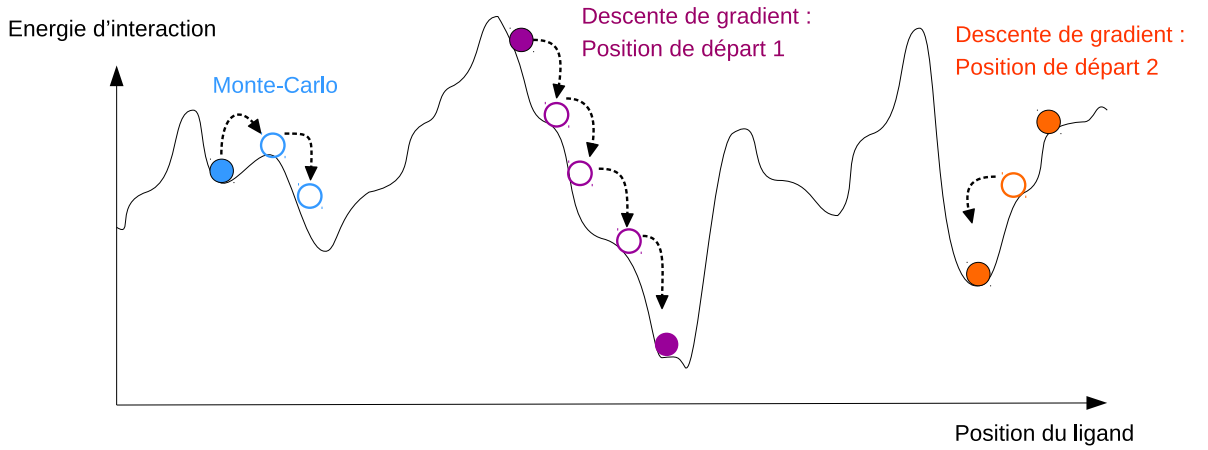


FIGURE 1.15 – Exemple de deux positions de départ pour une descente en gradient et d'une position de départ en Monte-Carlo sur une représentation de l'énergie d'interaction en fonction de la position du ligand sur le récepteur.

augmente, une décision aléatoire est appliquée, permettant de parfois garder une augmentation.

L'avantage de la méthode Monte-Carlo est qu'elle peut sortir des minimums locaux, alors que la descente de gradient ne peut pas. Cependant, l'utilisation de nombreuses positions initiales permet à la descente de gradient de contourner partiellement ce problème. La figure 1.15 présente ces deux solutions. L'inconvénient de la méthode Monte-Carlo est qu'il n'est pas garanti de converger vers une solution (la méthode du recuit simulé peut résoudre ce problème). Rosetta [28] est l'outil le plus connu pour le docking utilisant la méthode Monte-Carlo. Pour la descente de gradient, les deux outils les plus connus sont HADDOCK [83] qui permet une utilisation simple des données expérimentales (voir section 1.3.4) et ATTRACT (qui sera celui utilisé dans le reste de la thèse).

Évaluation

Nous avons vu comment les méthodes d'échantillonnage permettent d'obtenir plusieurs poses. Une fois ces poses obtenues, il faut les discriminer. Pour cela, une fonction d'évaluation est nécessaire. Cette étape d'évaluation, très délicate, consiste à approximer au mieux l'énergie d'interaction. Cependant, cela reste une approximation et comme le montre la figure 1.16, il est possible (et même courant) d'évaluer un minimum local comme le minimum global. De plus, il est possible que la position expérimentale soit différente de la position de plus basse énergie *in vivo*.

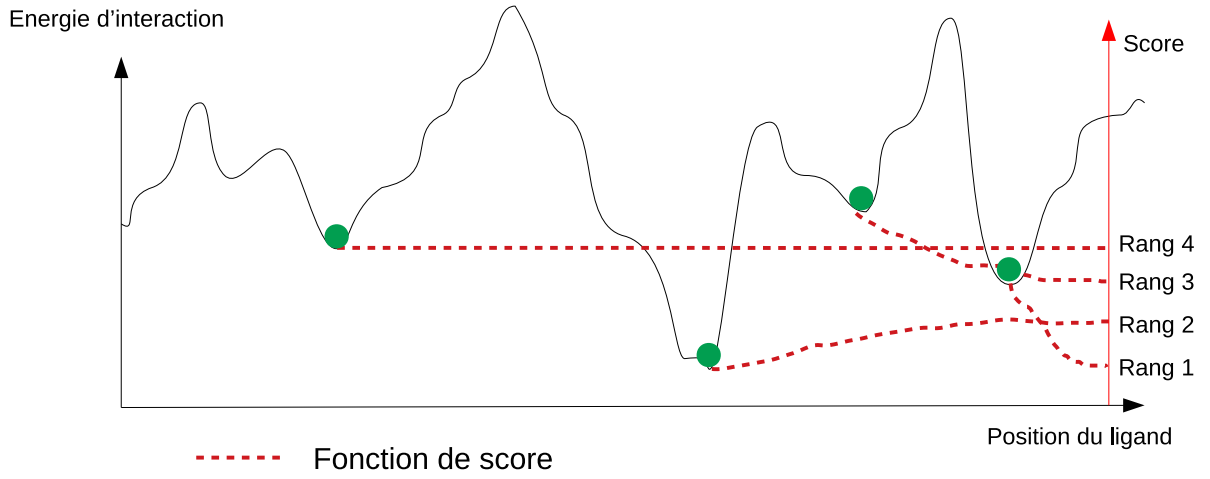


FIGURE 1.16 – Exemple d'une fonction de score qui classe 4 poses en utilisant une approximation de l'énergie d'interactio,n réelle. La meilleure pose n'étant pas celle d'énergie minimum réelle.

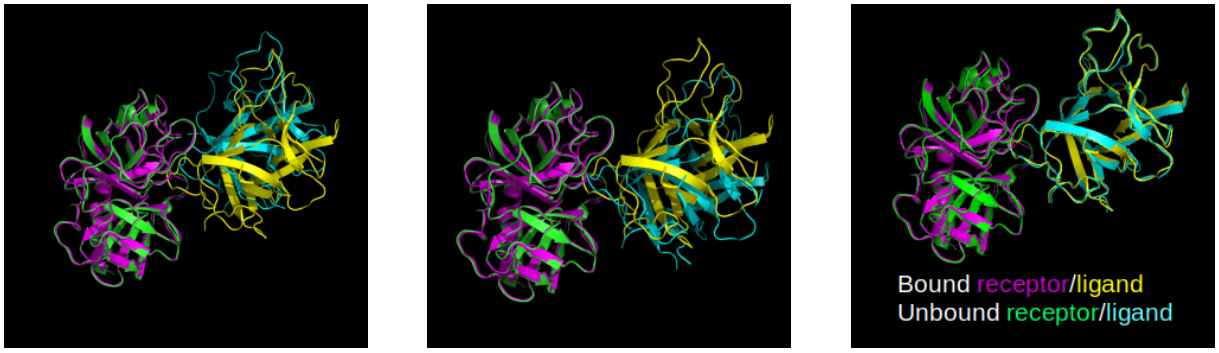
Il existe d'autres facteurs que l'énergie d'interaction pouvant être pris en compte pour réaliser l'échantillonnage et l'évaluation, notamment la complémentarité de forme (qui prend aussi en compte la surface non accessible au solvant). Ils ne sont pas utilisés dans cette thèse.

Validations des résultats de docking

Pour évaluer les résultats de docking, la comparaison est faite avec les structures expérimentales. Il existe un concours annuel permettant aux différentes équipes de recherche de comparer leurs méthodes de docking. Ce concours s'appelle *Critical Assessment of PRediction of Interactions* (CAPRI [47]). Le principe est que des biologistes résolvent expérimentalement une structure, mais avant qu'ils ne publient cette structure des participants prédisent cette structure. Ils connaissent la séquence des molécules impliquées, mais ne connaissent pas la structure du complexe. CAPRI utilise plusieurs notations pour comparer des résultats à la structure expérimentale :

- la RMSD du ligand (RMSD) est la mesure du RMSD entre le ligand amarré et celui de la structure expérimentale après avoir superposé le récepteur ;
- la RMSD de l'interface (iRMSD) est la mesure du RMSD entre les atomes étant à l'interface entre les deux molécules (c.-à-d. à moins de 10 Å de l'autre molécule) après superposition sur l'interface ;
- la fraction de contact natif (Fnat) est le pourcentage de contacts natifs retrouvés, un contact entre résidus étant considéré si deux atomes lourds sont à moins de 5 Å.

Ces mesures permettent de définir quatre catégories de qualités des solutions :



(a) iRMSD : 3.5 Å
*

(b) iRMSD : 1.7 Å
**

(c) iRMSD : 0.5 Å
***

FIGURE 1.17 – Les notations CAPRI, permettant de différencier la qualité d'une solution.

— élevées, si $f_{nat} \geqslant 0.5$ ET (LRMSD $\leqslant 1.0$ OU iRMSD $\leqslant 1.0$) ;
— moyennes, si ($0.3 \leqslant f_{nat} < 0.5$ ET (LRMSD $\leqslant 5.0$ OU iRMSD $\leqslant 2.0$)) OU ($f_{nat} \geqslant 0.5$ ET LRMSD > 1.0 ET iRMSD > 1.0) ;
— acceptables, si ($0.1 \leqslant f_{nat} < 0.3$ ET (LRMSD $\leqslant 10.0$ OU iRMSD $\leqslant 4.0$)) OU ($f_{nat} \geqslant 0.3$ ET LRMSD > 5.0 ET iRMSD > 2.0) ;
— incorrectes, si $f_{nat} < 0.1$ OU (LRMSD > 10.0 ET iRMSD $\leqslant 4.0$).

De plus, le concours définit des difficultés liées aux cibles en fonction des homologies présentes ou non dans la PDB. Un cas difficile ne possède aucune structure homologue connue, alors qu'un cas simple possède plusieurs homologues déjà connus.

Prise en compte de la flexibilité dans le docking

Historiquement, le récepteur et le ligand étaient considérés comme rigides lors du docking (pour des raisons de coût de calculs). Cette limite a pu être levée petit à petit en prenant en compte la flexibilité des macromolécules dans le cadre du docking.

On distingue plusieurs structures, celles liées (c.-à-d., appariées avec un partenaire), et celles non liées (c.-à-d., seules en solution). De plus, il y a aussi les structures connues (c.-à-d., observées expérimentalement) et les structures approximées (c.-à-d., prédites de manière computationnelle).

Dans le cas où les structures liées des partenaires sont proches de structures liées connues (par homologie), le docking rigide permet d'obtenir de bons résultats. De la même manière, si les changements conformationnels entre les structures liées et non liées sont faibles (ou les erreurs d'approximation, dans le cas de structures approximées), les résultats du docking rigide peuvent être bons. Cependant, s'il n'existe pas d'homologue, et que la structure non liée est très différente de la structure liée, alors le docking rigide ne va pas fonctionner.

La prise en compte de la flexibilité peut limiter ces problèmes et permet d'obtenir de meilleurs résultats que le docking rigide dans certains de ces cas plus complexes.

La flexibilité peut être prise en compte pour chacun des partenaires de différentes manières. Les plus connues sont développées dans cette partie.

Le docking multi-conformations consiste à réaliser des docking rigides de différentes conformations pour la macromolécule flexible. Pour réaliser cette méthode, il faut cependant avoir un ensemble de conformations contenant au moins une conformation proche de celle liée. Cet ensemble peut être nommé bibliothèque conformationnelle. Ces bibliothèques peuvent être compliquées à créer. Cette méthode permet de modéliser l'effet de la sélection conformationnelle (voir section 1.1.6).

Les modes normaux sont une approximation harmonique des mouvements globaux d'une molécule comme une combinaison de modes de vibration indépendants les uns des autres (fréquences différentes). Ainsi, la flexibilité d'une macromolécule ou de l'assemblage est modélisée comme la somme de mouvements élémentaires. Cette méthode permet de modéliser l'induced-fit (voir section 1.1.6) en déformant la molécule au cours du docking (à moindre coût computationnel et de manière continue).

Le docking par fragments consiste à découper le ligand en fragment et à réaliser un docking multi-conformations de chaque fragment. Ensuite, les poses des fragments de positions compatibles sont réassemblées en une structure complète. Cette méthode permet de modéliser la flexibilité locale à l'échelle des fragments (en réalisant un docking multi-conformations), et de modéliser la flexibilité globale en assemblant des fragments de positions (et de conformations) différentes. Elle modélise l'effet de l'induced-fit en assemblant les fragments à même le récepteur. Ce type de docking est inutile pour les ligands peu flexibles, et n'est applicable que pour de petites structures linéaires.

Le gros-grain est une manière de simplifier la représentation des macromolécules en regroupant plusieurs atomes en un pseudo-atome. Cette "simplification" permet de lisser les incertitudes sur la position des atomes, ce qui donne une certaine amplitude à la position des atomes, et modélise la flexibilité locale. Ce "lissage" permet de limiter le nombre de minimums locaux, ce qui augmente la probabilité de trouver le minimum global et diminue le nombre de faux positifs (voir la figure 1.18). Un autre avantage est la complexité en termes de temps de calcul. En effet,

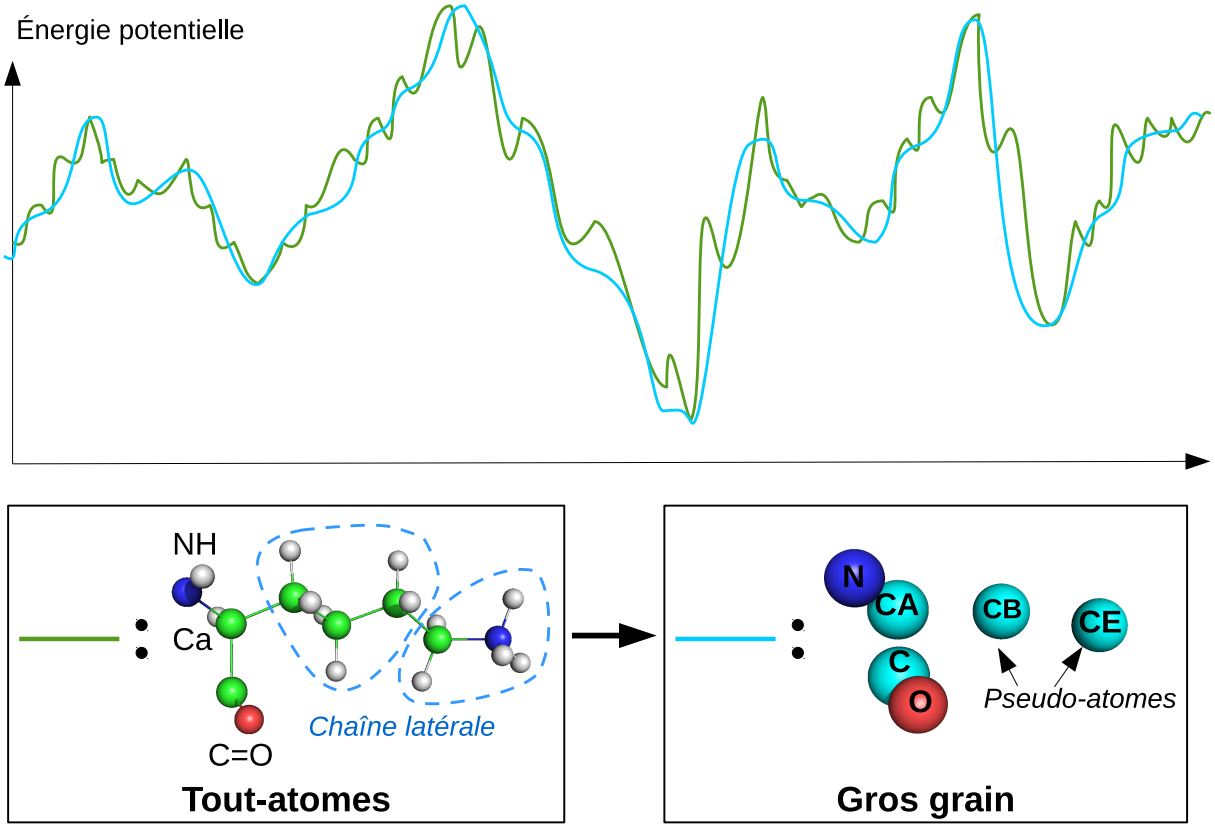


FIGURE 1.18 – Représentation schématique réduite à une dimension du paysage énergétique d'interaction entre deux molécules en représentation tout-atome et dans le cadre gros-grain. Un exemple représentatif de gros-grain ATTRACT pour un acide aminé est donné sous la courbe.

comme il y a moins d'atomes, il y a moins d'interactions à calculer, dans le cas où l'on calcule l'énergie pour chaque paire d'atomes.

Le plus gros problème de l'approche gros-grain est qu'il faut calibrer un nouveau champ de force avec les pseudo-atomes. Calibrer un tel champ de force peut être un vrai casse-tête pour prendre en compte les différentes énergies impliquées entre deux pseudo-atomes contenant plusieurs atomes. De plus, il est impossible de calibrer ces champs de forces de manière expérimentale (en tout cas de manière indirecte en observant chacune des interactions de manière seule), à l'inverse des champs de force tout-atome existants dans la littérature (Amber [68], Charmm[19]). Il existe cependant des représentations gros grain utilisées par la communauté et qui possèdent des champs de force utilisable dans le cadre général (Martini [70]), ou dans des cadres spécifiques au docking ARN-protéine (Attract [80]). Il faut cependant faire attention aux méthodes ciblées par ces champs de force. En effet, dans le cadre d'une minimisation en descente de gradient, il faut pouvoir dériver le champ de force (ce qui n'est pas nécessaire dans l'application du Monte-Carlo).

L'affinement des résultats de docking peut être réalisé par une courte dynamique moléculaire. Cette dynamique moléculaire va en général améliorer les résultats localement (réarrangement de l'interface) mais ne va pas réaliser de changements d'orientation globale du fait des limites de temps de calcul.

Docking guidé par les données

Une des solutions pour améliorer le docking est d'avoir accès à des informations biologiques pour ajouter des contraintes sur le docking. Les données obtenues expérimentalement peuvent être utilisées, ou alors ces données peuvent être prédites. Parmi les informations intéressantes, on peut noter les suivantes :

— la forme globale du complexe
— les contacts entre les partenaires
— l'interface de contact

Des informations sur la forme globale peuvent être obtenues par la méthode *Small X-Ray Scattering* (SAXS), mais aussi par Cryo-ME. Ces données peuvent être utilisées pour ajuster ou replier la ou les macromolécule(s) dans un volume, ou pour vérifier l'adaptation de la forme globale obtenue avec les données. Les méthodes ATTRACT EM et SAXS-ATTRACT permettent d'utiliser ces informations [17, 79].

Les contacts entre partenaires peuvent être étudiés grâce aux transferts d'énergie entre molécules fluorescentes (notamment la GFP : *Green Fluorescent Protein*), appelés FRET (*Fluorescence resonance energy transfer*). Les *cross-link* (induction de liaisons covalentes) entre partenaires aident aussi à l'identification des contacts pouvant exister. Ces informations de contacts permettent de contraindre le docking en forçant les contacts à exister dans les modèles.

Les interfaces ou sites de liaison peuvent être situés à l'aide de mutagenèses *in vitro*. Pour cela, on regarde si modifier un acide aminé empêche le contact d'avoir lieu avec le ligand. Il est aussi possible de trouver quels résidus sont à l'interface d'une interaction à l'aide de sondes chimiques, les résidus à l'interface étant protégés d'une modification chimique. Il existe aussi des méthodes computationnelles permettant de prédire un site de liaison de l'ARN sur une protéine. Ces méthodes sont souvent basées sur la complémentarité chimique et cherchent donc les zones chargées positivement [86].

Un outil de docking particulier nommé HADDOCK permet de prendre en compte une large majorité de ces informations biologiques pour contraindre le docking et ainsi obtenir de meilleurs

résultats.

1.4 Le docking par fragments

1.4.1 Le principe du docking par fragment

Le principe du docking par fragment est de découper le ligand en petits fragments, puis de réaliser le docking de chacun de ces fragments avant de réassembler les meilleures poses. Cette méthode présente plusieurs avantages. Le premier est que l'exploration de la conformation globale se fait par assemblage des fragments liés à la protéine. Ceci implique de ne garder que les conformations liées. Le deuxième avantage est l'exploration indépendante des conformations locales : certaines conformations ne vont pas bien se lier à la protéine et ne seront pas retenues. Ceci réduit l'échantillonnage des conformations globales du ligand puisqu'on le modélise directement sur la surface de liaison.

Cependant, cette méthode implique plusieurs challenges informatiques. Parmi ceux-ci, le premier est la discrétisation de l'espace conformationnel des fragments. En effet, il faut un équilibre entre la cardinalité de l'ensemble (pas trop grande sinon il y a trop de docking) et la représentativité de l'ensemble (c.-à-d., s'assurer qu'il y a au moins une conformation proche de la conformation réelle). Le deuxième challenge survient lors de l'assemblage des fragments : s'il y a beaucoup de fragments et beaucoup de poses par fragment, la combinatoire peut exploser et empêcher la résolution du problème. Ces deux challenges seront évoqués à plusieurs reprises au cours de la thèse.

L'enjeu capital pour réaliser l'assemblage des fragments en une chaîne complète est qu'il est nécessaire d'avoir au moins une pose correcte (proche de la pose native) pour chacun des fragments. Cette difficulté inhérente à la méthode implique que pour la réussite du docking il est nécessaire que tous les fragments interagissent suffisamment avec la protéine (nombre de contacts) pour être correctement échantillonnés lors du docking. De plus l'hypothèse de départ qui nous dit que l'ensemble protéine-ligand est au minimum d'énergie globale n'est plus valable pour les ensembles protéine-fragments. En effet, certains fragments *hot-spot* peuvent induire l'interaction globale alors que d'autres fragments ne se positionnent sur la protéine que de façon à accommoder la position de ces hot-spots. Cela signifie que le champ de force doit essayer de prendre en compte ces positions non optimales comme étant des positions plutôt bonnes (représentation en figure 1.19).

Un deuxième aspect de cette difficulté est lié à l'évaluation des poses (voir annexe B). En effet, même si l'échantillonnage donne des solutions quasi-natives, elles peuvent être très mal

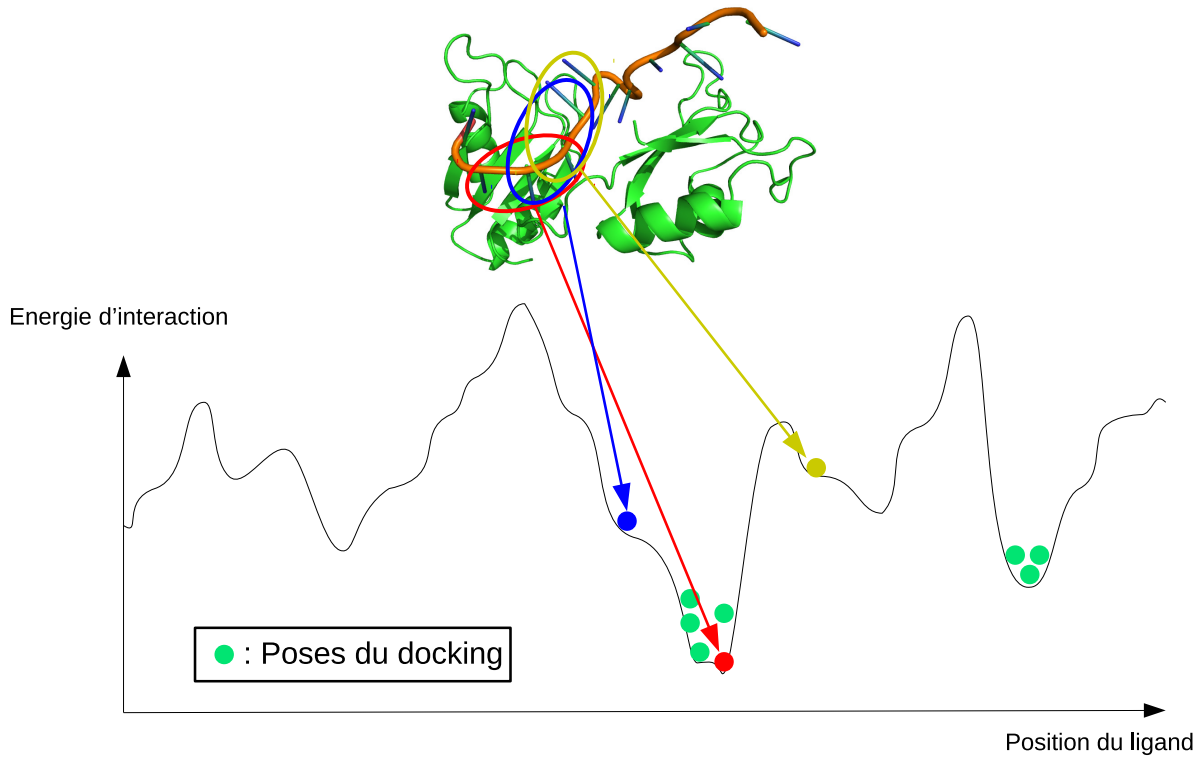


FIGURE 1.19 – Représentation du problème d'évaluation des fragments. Ici la séquence d'ARN évaluée (5 nucléotides) est séquence une poly-U. La courbe représente l'énergie d'interaction d'un tri-nucléotide UUU avec la protéine. Les poses réelles sont montrées, ainsi que les poses du docking. On peut comprendre que les poses réelles auront du mal à être échantillonnées.

évaluées. Dans ce cas, la reconstruction de la chaîne globale peut être très compliquée.

Les caractéristiques déterminant les cas particuliers où le docking par fragment est appliqué sont définies dans les sections suivantes.

1.4.2 Les applications possibles hors ARN

L'origine du docking par fragment

La première application du docking par fragment a été le *Fragment-Based Drug Design* (FBDD). Le fondement du FFBD a été décrit par Jencks [38] et soutenu par Nakamura et Abeles [62], qui ont montré que les molécules thérapeutiques peuvent être considérées comme la combinaison de plusieurs fragments. La FBDD présente deux avantages principaux par rapport aux méthodes de criblage. Le premier est que les bibliothèques de fragments peuvent couvrir une plus grande partie de l'espace chimique que les bibliothèques de criblage. Même les très grandes bibliothèques de criblage, avec plus d'un million de composés, ne peuvent explorer qu'une infime partie des 10^{60} composés estimés [7] ayant jusqu'à 30 atomes lourds. Étant donné que le nombre de composés possibles augmente de façon exponentielle avec la taille de la molécule, une bibliothèque de seulement quelques milliers de fragments avec moins de 17 atomes lourds est capable de couvrir une plus grande fraction de l'espace chimique. Le deuxième avantage réside dans le concept d'efficacité du ligand, c'est-à-dire la contribution moyenne de chaque atome de la molécule à l'affinité de liaison.

La FBDD présente aussi l'avantage de prendre en compte la flexibilité du ligand de par la construction de la molécule. En effet, les différents fragments représentent la flexibilité locale, alors que l'assemblage de ces différents fragments permet de prendre en compte la flexibilité globale. De plus, la prise en compte de la flexibilité de cette manière permet de modéliser à la fois la sélection conformationnelle et l'induced-fit.

L'utilisation dans le docking protéine-protéine, et dans le repliement protéique

L'utilisation des fragments dans le docking protéine-protéine se limite au docking de petits peptides sur des protéines, la taille des fragments de peptides étant variable selon les différentes méthodes existantes. Ici aussi, les fragments permettent de limiter le nombre de degrés de liberté par docking associés au squelette de la chaîne d'acides aminés, réduisant ainsi la complexité de la prise en compte de la flexibilité.

Historiquement, la méthode Rosetta utilisait les peptides comme des fragments pour modéliser le repliement protéique [76]. Suite à cette utilisation, ils ont aussi utilisé les fragments

pour faire du docking protéine-protéine [12]. D'autres équipes de recherche s'intéressent aussi à la création de bibliothèques de fragments protéiques [53].

1.4.3 Son utilisation dans le docking ARN/prot

Dans le cadre du docking ARN-protéine, les méthodes à base de fragments sont plus nombreuses que dans le cas protéine-protéine, et plus spécifiquement pour le cas de l'ARN simple brin, car c'est une structure linéaire facile à découper en fragments. De plus les méthodes classiques de docking n'arrivent pas à modéliser l'énorme flexibilité de cette molécule. Plusieurs méthodes par fragments sont présentées ici.

RNA-lim

RNA-lim [29] est la première méthode à avoir utilisé le docking par fragment pour l'ARN simple brin. Cette méthode a besoin de connaître le site de liaison pour ensuite réaliser le docking de nucléotides. Ces nucléotides sont représentés par un pseudo-atome caractérisé par un centre et un rayon, ce qui implique que l'orientation des nucléotides n'est pas connue. Cependant, cela permet quand même d'obtenir la position de la chaîne d'ARN simple brin à 15 Å près.

FBDRNA

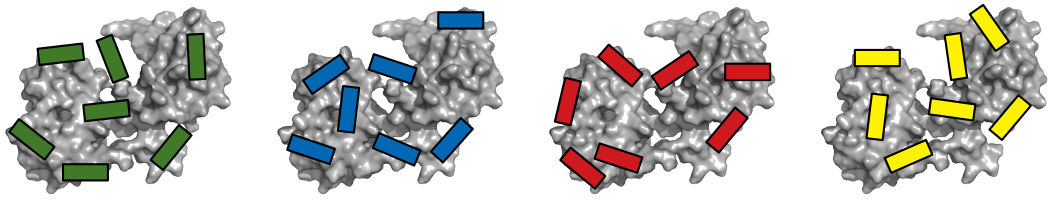
FBDRNA [26] est une méthode de docking par fragment, avec comme fragments des mono-nucléotides. Le docking de ces mono-nucléotides est réalisé, sur une région prédéfinie de la protéine, connue/prédite comme le site de liaison, avec une conformation en tout-atome pour chaque nucléotide. Si la séquence de l'ARN est connue, seuls les nucléotides d'intérêt sont amarrés. Si la séquence est inconnue (prédiction d'une séquence d'interaction), alors les quatre nucléotides sont amarrés.

Les meilleures poses sont ensuite sélectionnées par une fonction de score approximant l'énergie enthalpique d'interaction. La matrice de connexion entre les poses est calculée selon plusieurs critères de distance entre atomes. Puis les chaînes sont construites en utilisant la matrice de connexion. L'utilisateur peut spécifier la taille de la chaîne voulue, ou alors spécifier la séquence recherchée. Enfin, un processus de minimisation de l'énergie d'interaction permet d'optimiser les résultats obtenus.

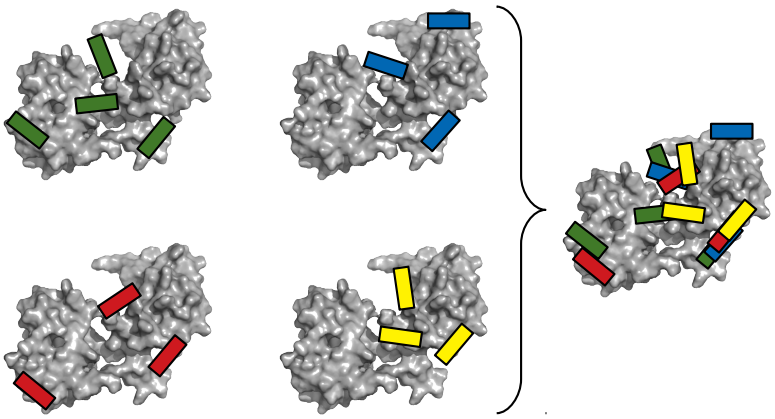
1.4.4 ATTRACT

Dans cette partie, nous allons développer le docking d'ARN simple brin basé sur les fragments réalisé par ATTRACT. Dans la suite du document, c'est ce docking qui sera utilisé. Pour cela,

(a) Les données en entrée.



(b) L'échantillonnage des fragments.



(c) L'évaluation des poses.

FIGURE 1.20 – Les différentes étapes du docking par fragment réalisé par ATTRACT.

plusieurs étapes sont réalisées, ces étapes sont schématisées dans la figure 1.20.

La création des bibliothèques de fragments

Notre méthode utilise comme fragments des tri-nucléotides en représentation gros-grain de ATTRACT. L'avantage des tri-nucléotides par rapport à des di-nucléotides ou des mono-nucléotides est qu'ils ont plus d'interactions avec la protéine donc ils se lient avec des positions plus spécifiques. De plus, le nombre de conformations possibles pour un tri-nucléotide est d'un ordre de grandeur beaucoup plus faible que pour un tetra-nucléotide (il y aurait trop de conformations différentes à gérer dans ce cas).

La première étape de la méthode de docking par fragments est la construction des bibliothèques de tri-nucléotides. Ces bibliothèques ont pour objectif de représenter toutes les conforma-

tions possibles d'une séquence donnée de tri-nucléotide. Pour cela, des représentants (prototypes) sont choisis tels que toutes les conformations observées dans la PDB se trouvent à 1 Å cRMSD ou moins d'un représentant. Le chapitre 3 est dédié au choix des représentants.

Le docking des fragments

Une fois ces bibliothèques obtenues (il n'est pas nécessaire de les faire pour chaque docking), la séquence que l'on veut amarrer sur la protéine est découpée en tri-nucléotides se chevauchant de 2 nucléotides.

Si la séquence contient n nucléotides, on obtient ainsi $n-2$ fragments. Pour chacune des séquences obtenues, la bibliothèque de fragments correspondante est amarrée sur la protéine. Pour cela, un docking rigide est réalisé avec un grand nombre de répétitions pour chaque conformation dans la bibliothèque. Il existe deux manières de démarrer l'échantillonnage dans ATTRACT, la méthode Randsearch qui place aléatoirement les fragments (en termes de position et rotation) sur une sphère centrée sur le centre de masse de la protéine, et la méthode Systsearch qui place les fragments à des positions régulières autour de la protéine (à une certaine distance de la surface), chaque conformation étant posée sur chaque position selon 228 orientations.

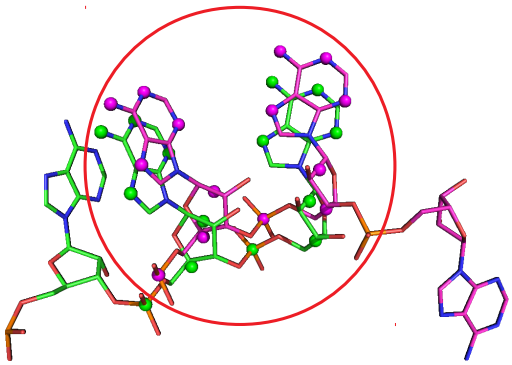
Une fois les positions obtenues, elles sont minimisées selon le score ATTRACT en descente de gradient. Les poses obtenues sont classées par score, les poses redondantes (selon un seuil de 0.2 Å RMSD) sont éliminées. Les poses de meilleur score sont conservées pour l'assemblage.

L'assemblage des fragments

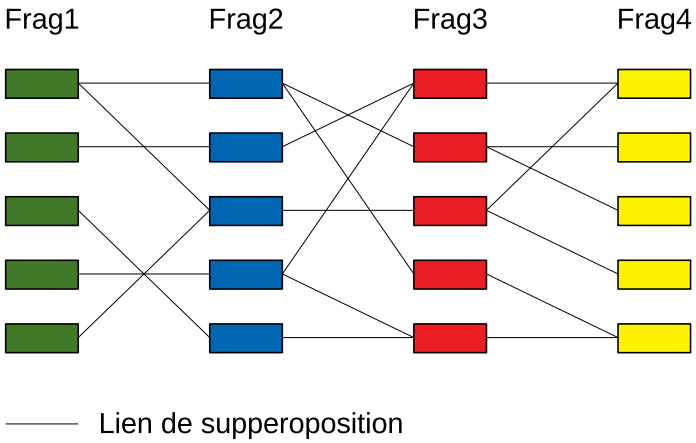
L'étape suivante consiste à créer un graphe de connexion entre les poses conservées. Pour cela, la RMSD entre les deux nucléotides chevauchants de deux poses de deux fragments successifs est calculée. Si cette RMSD est inférieure à un seuil défini alors un lien est ajouté dans le graphe. Une fois tous les liens ajoutés dans le graphe, il devient possible de le parcourir pour obtenir les chaînes complètes (voir figure 1.21).

Enfin, la dernière étape consiste à discriminer les chaînes obtenues. En effet, le nombre de chaînes obtenues peut être très grand, par exemple, sur une séquence de 7 nucléotides (donc 5 fragments possibles), si l'on décide de conserver les 100000 premières poses, le graphe peut contenir jusqu'à $(10^6)^5 = 10^{30}$ chaînes, ce qui rend impossible l'énumération des chaînes.

En l'absence de connaissances supplémentaires sur le système, le nombre de chaînes est donc trop important. On peut cependant obtenir un ensemble des poses les plus connectées. Pour cela, un comptage du nombre de chaînes auxquelles chaque pose participe est effectué en aller-retour par programmation dynamique. Les poses les plus connectées sont gardées comme solution et



(a) Chevauchement de deux nucléotides de deux tri-nucléotides (un en vert et un en rose).



(b) Graphe des poses

FIGURE 1.21 – Le graphe des poses, les liens sont ajoutés en fonction du seuil de distance des chevauchements.

définissent le site de liaison. Dans l'hypothèse où le nombre de poses à considérer est suffisamment faible et que l'on peut énumérer les chaînes, une discrimination des chaînes est possible à l'aide d'un seuil de rang moyen (approche *branch and bound*). Pour cela, un parcours du graphe est réalisé avec un calcul du rang moyen de la chaîne à chaque fois qu'une pose est ajoutée. Si le seuil est dépassé, on s'arrête et on supprime la chaîne. Cette étape peut quand même prendre beaucoup de temps (quelques heures pour énumérer 10 millions de chaînes contenant six fragments).

Les spécificités d'ATTRACT

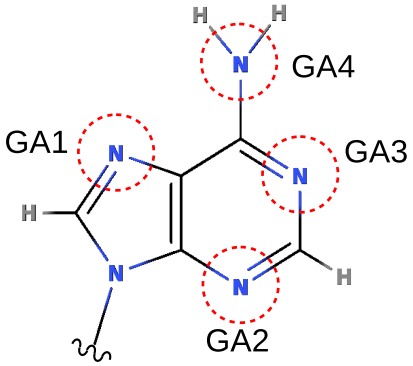
ATTRACT peut effectuer l'échantillonnage par une descente de gradient et possède sa propre représentation gros-grain (voir figure 1.22). Cette représentation permet de garder l'information sur l'orientation des bases azotées dans l'espace. Le champ de force associé à cette représentation a été développé en 2011, et il est actuellement amélioré par une doctorante de l'équipe. La paramétrisation d'un tel champ de force est un enjeu majeur, puisque c'est en suivant ce champ de force que la minimisation et l'évaluation sont faites. Au cours de ma thèse, j'ai pu voir qu'il existe certaines erreurs dans le champ de force défini en 2011 (voir annexes B).

Le docking ancré est une méthode qui permet d'utiliser la connaissance de points d'ancrage (c.-à-d., des contacts connus entre un acide aminé et un nucléotide) pour contraindre très fortement le docking par fragments. Elle permet d'obtenir des modèles de chaînes complètes avec une bonne précision, mais nécessite des connaissances biologiques spécifiques au système. Nous ne développons pas cette méthode, car elle n'est pas utilisée dans la suite du document.

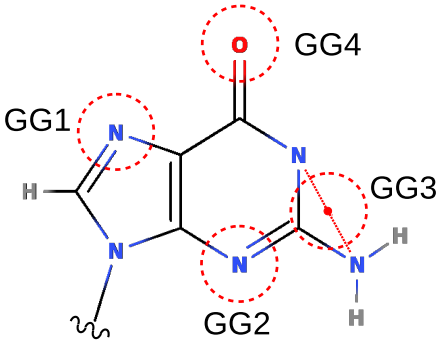
Comparaison avec FBDRNA : une différence avec l'approche FBDRNA est l'absence de la connaissance du site de liaison. Comme nous ne connaissons pas le site de liaison, il faut échantillonner sur toute la protéine. Cela implique de très nombreux minimums locaux pour de petits fragments tels que les nucléotides. C'est pour limiter ces minimums locaux que nous utilisons des tri-nucléotide et la représentation gros-grain, plutôt que des nucléotides en tout-atome.

1.4.5 Bibliothèques de fragments d'AN

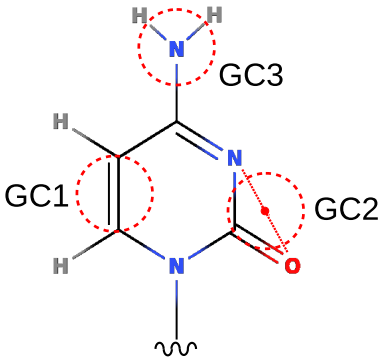
Selon l'utilisation des bibliothèques de fragments, il peut être intéressant d'avoir des bibliothèques différentes. Ces contextes peuvent se baser sur la taille des fragments (nombre de nucléotides/acides aminés) ou sur les environnements des fragments (par exemple en contact ou



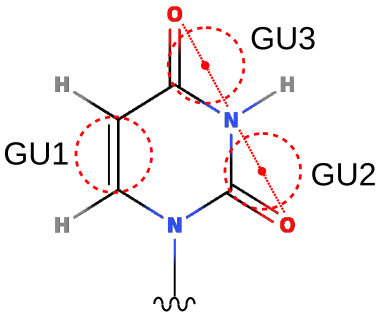
(a) Adénine



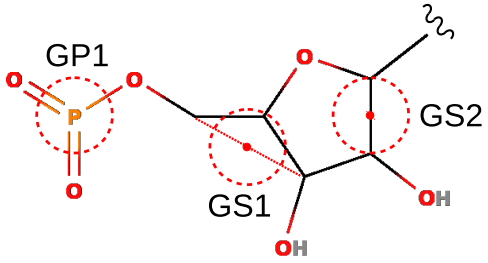
(b) Guanine



(c) Cytosine



(d) Uracile



(e) Sucre/Phosphate

FIGURE 1.22 – Les cercles rouges représentent les pseudo-atomes du gros-grain ATTRACT. Ils sont placés sur les atomes ou sur les centres géométriques de plusieurs atomes.

non avec une protéine). Il existe de très nombreuses bibliothèques de fragments qui répondent à des contextes plus ou moins complexes.

Pour les AN, celles décrites dans ([95, 78]) sont dédiées à l'ADN. Elles permettent de chercher des structures conformationnelles particulières dans la PDB ainsi que de les visualiser. Pour l'ARN, les bases de données sont souvent centrées sur un contexte particulier. C'est le cas de **RNA CoSSMos 2.0** ([74]) qui se concentre sur les motifs en forme de boucle. Une exception est **RNANet** ([3]) qui analyses toutes les structures cristallographiques de la PDB. Elle recense un grand nombre de descripteurs géométriques (tous les angles de torsion des structures). De plus, cet outil fournit un alignement des séquences provenant des familles Rfam. Ce grand nombre de descripteurs est rassemblé dans une base de données SQL. Cependant, les contacts avec les protéines sont très peu développés.

Dans la suite du document, nous présenterons notre outil qui permet de créer des bibliothèques de fragments les plus modulables possibles pour répondre à nos besoins en termes de docking.

Les bibliothèques de fragments peuvent aussi avoir pour objectif d'être utilisées dans le cadre d'affinement de structures existantes. Dans ce cas, les bibliothèques doivent être aussi précises et exactes que possible, et ce n'est plus l'exhaustivité qui est recherchée. C'est le cas de la bibliothèque créée par l'équipe de Černý Jiří qui définit un alphabet des conformations exactes d'AN [84].

Chapitre 2

ProtNAff et bibliothèques de fragments

Ce chapitre recouvre les possibilités proposées par ProtNAff (*protein-bound Nucleic Acids filters and fragments libraries*) : la création d'une base de données, la sélection d'un sous-ensemble de structures et la construction de bibliothèques de fragments, et les différentes étapes de chaque processus sont détaillées dans la figure 2.1. Les résultats de quatre analyses quantitatives effectuées avec protNAff sont présentés à titre d'exemple. Chaque exemple utilise une combinaison complexe de filtres. Nous nous concentrons sur l'ARN parce que les structures d'ARN sont plus diverses dans le PDB que les structures d'ADN, mais les mêmes analyses ont été appliquées à l'ADN (elles sont présentées de manière plus succincte).

Les quatre exemples sont les suivants.

— la spécificité de la séquence des conformations tri-nucléotidiques à une résolution de 1 Å est mesurée pour chaque motif ;

— la diversité conformationnelle des tri-nucléotides est mesurée à différents niveaux de résolution ;

— la fraction des conformations locales d'ARN existantes qui sont spécifiquement induites par la liaison aux protéines et qui n'existent pas dans les régions d'ARN non liées est mesurée pour différentes séquences d'ARN et différents états structurels 2D ;

— la propension à la liaison protéique des boucles en épingle à cheveux de l'ARN est comparée pour différentes longueurs de boucle purine/pyrimidine.

Cet outil a amené à une publication dans un journal [59] (hal-03765772).

2.1 Contexte

Comme nous l'avons expliqué en section 1.3.3, la grande flexibilité des AN doit être prise en compte pour leur modélisation et peut rendre difficile cette modélisation [41]. En particulier, la

flexibilité élargit l'espace conformationnel à modéliser. Alors que pour les protéines, il existe des bibliothèques de fragments qui décrivent l'espace conformationnel possible [5], de telles bibliothèques pour les ANS sont trop restrictives pour nos utilisations. Les AN adoptent en général des conformations beaucoup plus spécifiques au contexte. Pour l'ADN, il y a souvent un ajustement induit lors de la liaison à une protéine [57]. Pour l'ARN purement simple brin, la conformation non liée est soit désordonnée, soit dans une autre structure secondaire dans l'état non lié, et ne devient définie que lors de la liaison.

La modélisation de ces conformations nécessite la création de bases de données d'assemblage protéine-AN et de bibliothèques de fragments spécifiques à un contexte biologique ou méthodologique donné (par exemple AN simple brin, contact avec une protéine, méthode expérimentale de résolution, etc.). Un exemple de contexte très spécifique est le sous-ensemble de structures d'ARN provenant uniquement de la cristallographie, avec au moins 2 Å de résolution, avec une boucle en épingle à cheveux d'ARN d'au moins cinq nucléotides, dont au moins 70% sont en contact avec une protéine. Ce contexte a été utilisé pour créer un *benchmark* pour une méthode spécifique de docking d'épingle à cheveux présenté dans la section 4.1. En principe, chaque nouveau contexte et chaque nouvelle utilisation nécessitent une nouvelle base de données.

En outre, de nouvelles structures 3D expérimentales des AN sont disponibles en permanence. Par conséquent, la création de bases de données d'AN nouvelles et mises à jour continuera d'être nécessaire pour faire avancer le domaine. Cependant, la construction de telles bases de données représente une grande quantité de travail, souvent manuel, car il n'existe pas d'outil générique pour construire des bases de données d'AN de manière automatisée. Nous avons développé un tel outil.

2.2 Présentation de l'outil ProtNAff

2.2.1 Principe et avantages

Le pipeline ProtNAff présenté dans ce chapitre apporte une solution pour la création de filtres pour sélectionner des ensembles de structures 3D issues de la PDB, et pour la création de bibliothèques de fragments, il est capable de :

— corriger et analyser les structures de protéines-AN issues de la PDB (en utilisant partiellement les outils de la suite ATTRACT [16]) ;
— extraire un large ensemble de leurs caractéristiques structurelles (en utilisant partiellement les outils DSSR de la suite 3DNA [51]) ;
— organiser ces données dans un fichier JSON.

La sortie peut être traitée par n'importe quel langage de programmation, pour appliquer des filtres personnalisés. Les filtres permettent d'identifier les structures sur la base d'un ensemble complexe de critères, d'extraire leurs régions d'intérêt et de créer des bibliothèques contextuelles de fragments présentant différents niveaux de redondance. Ils permettent également d'effectuer une analyse automatique de l'effet d'un contexte sur les caractéristiques locales d'AN, telles que la diversité des conformations des fragments ou leur type d'interactions avec les protéines. Dans ce chapitre, comme preuve de concept, un ensemble de filtres écrits en Python est proposé, correspondant à différents contextes, qui peuvent être facilement adaptés aux besoins de l'utilisateur.

ProtNAff se spécialise actuellement dans la création et l'analyse de bibliothèques de tri-nucléotides d'ARN. Un point fort particulier par rapport à l'utilisation de bibliothèques existantes est la possibilité de moduler la résolution de l'espace conformationnel. Nous avons réalisé l'analyse statistique des caractéristiques des conformations locales dans différents contextes et leur interprétation biologique. De plus, le seuil approprié pour regrouper (c.-à-d., le rayon des *clusters*) les fragments d'AN est étudié pour détecter les caractéristiques spécifiques au contexte. Par exemple, pour détecter si certaines conformations locales d'AN sont spécifiquement induites par la liaison à une famille de protéines donnée, le seuil de regroupement choisi doit être suffisamment élevé pour créer des clusters spécifiques à la famille, mais suffisamment bas pour que ces clusters contiennent des fragments provenant d'un nombre statistiquement suffisant de structures liées à différents membres de la famille.

En outre, la granularité de l'espace conformationnel a des implications pour l'utilisation des bibliothèques de fragments d'AN dans la modélisation 3D. Dans ce cas, le seuil de regroupement des fragments doit équilibrer le niveau de précision souhaité et la taille maximale de la bibliothèque (c.-à-d., le nombre de conformations), selon le processus de modélisation et de son application. Dans le cadre de l'étape d'échantillonnage pour le docking basé sur les fragments avec des tri-nucléotides, plusieurs options sont possibles. La première consiste à utiliser un petit nombre de fragments très différents, correspondant à un seuil de regroupement élevé, et à ajouter une flexibilité explicite (par exemple, la dynamique moléculaire) pour couvrir le paysage conformationnel autour de chaque fragment. La seconde est de faire un échantillonnage rigide avec des conformations beaucoup plus proches les unes des autres (donc une plus grande bibliothèque), correspondant à un seuil de regroupement faible, pour obtenir un échantillonnage suffisamment exhaustif des paysages conformationnels.

Cette polyvalence permet à protNAff de couvrir une très grande partie des multiples contextes possibles mentionnés ci-dessus. De plus, la polyvalence de l'outil, en évitant les critères stricts

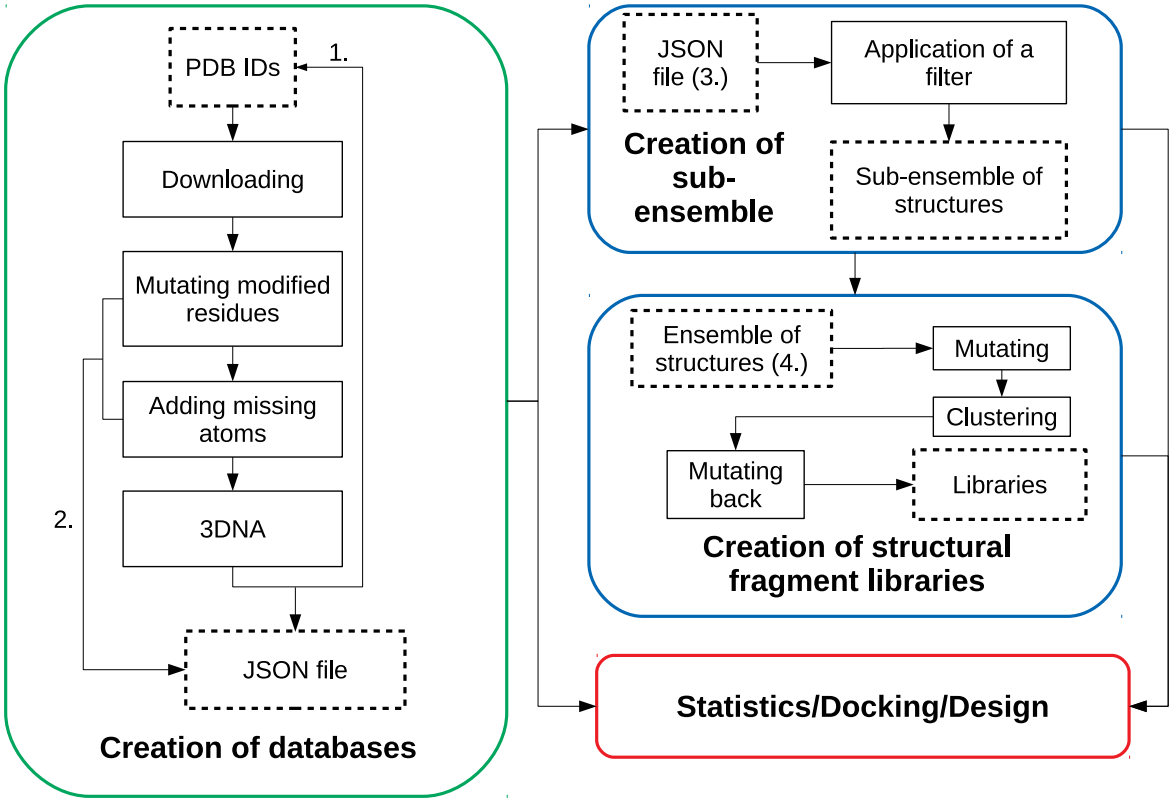


FIGURE 2.1 – Ce schéma global présente le pipeline et ses étapes les plus importantes. Chaque sortie - base de données, sous-ensembles, fragments - peut être utile indépendamment : recherche d'exemples particuliers de complexes, analyses statistiques sur les motifs, modélisation à base de fragments, etc. Les boîtes en pointillés sont les fichiers d'entrée et de sortie. (1.) Toutes les étapes sont effectuées pour chaque identifiant pdb. (2.) Les données concernant les modifications apportées aux fichiers pdb sont stockées dans la base de données. (3.) Le fichier JSON est celui obtenu comme base de données. (4.) L'utilisateur va créer son ensemble de structures pour faire des bibliothèques de fragments structuraux.

prédéfinis, le place dans une meilleure position pour exploiter les nouveaux types de données structurelles ou les nouvelles caractéristiques d'intérêt qui pourraient devenir disponibles dans un avenir proche. L'un des points forts de l'outil est que la base de données résultante est stockée localement, de sorte que l'utilisateur a le contrôle sur la base de données et peut l'utiliser comme il l'entend. Les différents cas d'utilisation peuvent être mis en œuvre de manière collaborative au fil du temps, le code étant accessible sur GitHub.

La base de données et la bibliothèque (clusters à 1 Å) créées par protNAff pour ces analyses sont disponibles en téléchargement à l'URL `https://zenodo.org/record/6483823#.YmbVhFxByV4`, ce qui permet de les utiliser et de les analyser sans exécuter le pipeline complet.

2.2.2 Création de la base de données

Méthode de création

Pour initier le processus de création, l'utilisateur fournit une liste d'identifiants pdb. Cette liste peut être obtenue à la suite d'une requête sur le site de la PDB, ce qui permet un premier filtrage par l'utilisateur. Les fichiers `.pdb` sont téléchargés puis édités. La méthode expérimentale pour obtenir la structure, et la résolution si disponible sont extraites (en analysant la description dans le fichier pdb). L'eau, les ions et autres petits ligands sont supprimés.

Les atomes manquants dans les nucléotides ou les acides aminés sont détectés à l'aide de l'outil ATTRACT `aareduce.py`. Les nucléotides qui contiennent moins de quatre atomes sont écartés, les autres sont conservés pour ajouter les atomes manquants. Pour un acide aminé, l'outil pdb2pqr [20] est utilisé. Pour un nucléotide, une bibliothèque de mono-nucléotides est alignée sur les atomes existants en utilisant l'outil `fit.py` de ATTRACT, et les atomes du nucléotide qui produit le meilleur alignement (cRMSD résiduel le plus faible) sans interpénétration sont ajoutés (voir Annexes). Pour réaliser cela, nous avons créé une bibliothèque structurelle de mono-nucléotides par base ARN/ADN, en extrayant tous les nucléotides de la pdb, en supprimant ceux qui sont non canoniques ou avec des atomes manquants, en les regroupant à 0,3 Å cRMSD, en éliminant les clusters avec moins de 10 membres, et en gardant un prototype dans chaque cluster.

Les informations sur les nucléotides d'une structure ayant des atomes manquants et dans quelle partie (phosphate, sucre, base) sont stockées dans la base de données. Cela permet à l'utilisateur de choisir comment traiter ces nucléotides. Par exemple, les rejeter ou les conserver dans une bibliothèque de fragments en fonction de contraintes d'exhaustivité ou d'exactitude.

Les nucléotides modifiés avec des atomes supplémentaires sont également détectés à l'aide de l'outil ATTRACT `aareduce.py`. Si possible, ils sont transformés en leur nucléotide canonique le plus proche (nombre le plus élevé d'atomes communs) en enlevant quelques atomes choisis selon une table `[nom de résidu - atomes à retirer]` que nous avons pré-établie manuellement par observation de cas sous PyMol (ces modifications sont répertoriées dans `protNAff/data/[r/d]nalib/mutate.list`). Ils sont éliminés s'ils sont trop éloignés de tout nucléotide canonique.

La base de données finale stocke une liste des nucléotides qui ont été canonisés et des nucléotides qui n'ont pas été canonisés. Cela permet à l'utilisateur de conserver soit tous les fragments (pour se rapprocher de l'exhaustivité), soit seulement les fragments authentiques (pour se rapprocher de l'exactitude), en fonction des applications de la bibliothèque. Le script protNAff `filter_no_modified.py` est fourni pour écarter les fragments modifiés.

Une renumérotation des nucléotides à partir de 1 est appliquée. S'il existe des positions

alternatives pour certains résidus/atomes, plusieurs fichiers `.pdb` sont créés, un pour chaque ensemble de positions alternatives.

Pour chaque partie (phosphate, sucre, base) de chaque nucléotide, la distance de chaque atome lourd par rapport à un atome lourd de protéine ou d'un cofacteur est calculée, et enregistrée si elle est inférieure à un seuil choisi (par défaut 5 Å). L'outil DSSR du package 3DNA est appliqué sur le fichier `.pdb` modifié. Cela fournit des informations exhaustives sur la structure 2D et 3D de l'AN, dont une grande partie est intégrée dans la base de données. Ces informations sont ensuite analysées et converties du niveau de la structure au niveau des tri-nucléotides sur les bibliothèques de fragments. Nous voulons l'information au niveau des nucléotides pour pouvoir appliquer les filtres plus facilement, et aussi pour les études statistiques au niveau des nucléotides. Par exemple, à partir d'un énoncé "les nucléotides 5 à 15 forment une tige-boucle" de la sortie du DSSR, on extrait un état "le nucléotide 5 est en position 1 dans une tige-boucle de 11 nucléotides" à stocker dans notre base de données. La base de données est placée dans un fichier JSON, pour une analyse facile avec le langage de programmation désiré.

Contenu de la base de données

La première sortie du pipeline protNAff est un fichier JSON contenant des informations par structure (par exemple, la résolution cristallographique), par chaîne d'acide nucléique (par exemple, les positions des cassures dans le squelette) et par nucléotide (par exemple les liaisons hydrogène avec la protéine). La deuxième sortie de protNAff est une bibliothèque de fragments (tri-nucléotides consécutifs dans l'implémentation actuelle), constituée des coordonnées des fragments (dans des matrices `.npy` [32] ou dans le format pdb) et des métadonnées/informations structurelles par fragment, stockées dans un autre fichier JSON (identifiant pdb de la structure originale, indice de cluster, atomes manquants dans la structure initiale, etc.).

La combinaison des deux fichiers JSON contenant les informations, l'un par structure et l'autre par fragment, permet de calculer toutes sortes de statistiques sur la bibliothèque. Quelques exemples sont présentés dans la section 2.3. Une liste complète des données par structure et par fragment est présentée ci-dessous.

Par structure PDB :

— la méthode expérimentale ;
— la résolution (si cristallographie) ;
— les noms des chaînes d'AN et protéiques ;
— le nombre de modèles ;
— les noms des cofacteurs.

Par chaîne d'AN' :

— la position des cassures dans le squelette ;
— la séquence.

Par nucléotide :

— toutes les liaisons hydrogène faites avec la protéine, avec pour chaque liaison (i) l'acide aminé impliqué ainsi que l'atome, (ii) la partie du nucléotide (sucre, phosphate ou base azotée), (iii) la distance de la liaison H, et (iv) le fait que 3DNA considère le donneur ou l'accepteur "questionnable" ;
— toutes les liaisons hydrogène faites avec un autre acide nucléique, avec (i) la position de l'autre nucléotide dans la séquence (n-2, n-1, n+1, n+2 ou autre), (ii) la sous partie du nucléotide, (iii) la distance de la liaison H, et (iv) le fait que 3DNA considère le donneur ou l'accepteur "questionnable" ;
— le nombre total de liaisons hydrogène avec la protéine pour chaque sous partie, avec un poids de 0.5 si "questionnable" ;
— les types d'appariements de bases dans lesquels il est impliqué ;
— le nom initial du résidu dans le fichier PDB (s'il a été canonisé) ;
— la distance minimale de chacune de ses sous partie à la protéine ou à un cofacteur (cette distance est notée si elle est inférieure au seuil défini par l'utilisateur) ;
— les parties dans lesquelles il manquait des atomes dans le fichier PDB initial ;
— la structure secondaire (simple brin terminal, boucle d'une épingle à cheveux, jonction, double brin) ;
— la présence d'une interaction d'empilement avec les nucléotides en position n-2, n-1, n+1, n+2 ou autre.

Par fragment :

— le nom de la structure PDB de laquelle il est extrait ;
— l'index de son modèle (pour les structures PDB multi-modèles) ;
— le nom de la chaîne PDB ;
— les indices des résidus dans le fichier PDB ;
— la séquence initiale ;
— si des atomes étaient manquants alors dans quel(s) nucléotide(s) et dans quelle(s) partie(s) ;
— si le fragment est un prototype de cluster pour les différents seuils de *clustering* ;
— l'index du cluster auquel il appartient pour les différents seuils.

Filtres personnalisés

L'utilisateur peut appliquer des filtres sur la base de données pour la création d'ensembles de structures correspondant à un contexte donné. Les utilisations de ces ensembles peuvent être multiples et dépendent donc de l'utilisateur. Par exemple, ils peuvent être utilisés pour créer des benchmarks de complexes pour tester une méthode de docking. Cette étape est la moins automatisée, afin de laisser le plus de liberté possible à l'utilisateur. Elle consiste à écrire un script qui parcourt le fichier JSON et collecte les données d'intérêt. Un notebook d'explication est fourni (`filters/explanation_filters.ipynb`) sur la façon d'utiliser les fichiers JSON pour créer des filtres. L'exemple donné à la fin du notebook sélectionne les nucléotides appariés (appariement de Watson et Crick) qui sont en contact avec une protéine.

2.2.3 Création des bibliothèques de fragments

ProtNAff permet la création de bibliothèques de fragments structuraux. Ces bibliothèques peuvent être utilisées comme blocs de construction pour la prédiction de structure (modélisation d'ARN, docking, design) ou pour calculer des statistiques sur des assemblages particuliers. Pour créer ces bibliothèques, les structures d'intérêt sont sélectionnées et leurs fragments sont identifiés via le fichier JSON. Ces fragments sont extraits des structures PDB nettoyées, regroupés par séquence et un clustering est appliqué pour obtenir un ensemble de prototypes.

Dans le cas de notre méthode de docking basé sur les fragments (voir partie 1.4.4), les fragments d'intérêt sont les tri-nucléotides consécutifs. Il est toutefois possible, moyennant quelques modifications, de créer des bibliothèques d'autres structures, par exemple des di-nucléotides ou des hélices double brin.

Clustering

Une fois que tous les fragments des structures choisies ont été récupérés et extraits au format pdb, un clustering est effectué. Le but de ce clustering est d'obtenir des représentants de tous les fragments existants (observés ou non). L'utilisateur a la possibilité de choisir le rayon des clusters, ce qui permet une représentation plus ou moins fine de l'espace conformationnel.

Historiquement, l'algorithme de clustering ressemble au suivant : l'initialisation est réalisée en choisissant aléatoirement un fragment comme premier prototype de cluster. Ensuite, pour chaque fragment, la distance (RMSD) à chacun des prototypes de l'ensemble actuel est mesurée. Dès que l'une de ces distances est inférieure au rayon choisi, alors le fragment est affecté au cluster. Sinon, il est ajouté à l'ensemble des prototypes. Une fois que cette première passe sur l'ensemble

des fragments est terminée (c'est-à-dire une fois que tous les prototypes ont été obtenus), une deuxième passe est effectuée : les fragments qui ne sont pas des prototypes sont réaffectés au cluster du prototype le plus proche.

Ce clustering est effectué de manière itérative, avec des rayons différents : un premier clustering rapide avec un rayon RMSD de 0,2 Å est appliqué pour éliminer les redondances (en particulier, les fragments de deux assemblages biologiques de la même structure cristalline), puis les fragments non redondants sont regroupés avec un rayon RMSD de 1 Å puis de 3 Å.

Une deuxième méthode de clustering est implémentée dans protNAff. Cette méthode est présentée en détail au chapitre suivant. Il s'agit d'une classification agglomérative hiérarchique caractérisée par sa fonction de *linkage* qui prend en entrée deux clusters et renvoie le rayon de la plus petite boule les englobant. Les prototypes résultants ne sont pas des fragments d'entrée, mais les centres des boules. Si l'utilisateur a besoin d'une bibliothèque composée d'un sous-ensemble des fragments initiaux, la première méthode de clustering est celle à privilégier, sinon la seconde produit moins de clusters. Une comparaison entre ces deux méthodes de clustering appliquées à des fragments de tri-nucléotides peut être trouvée dans le notebook `clustering_comparison.ipynb` et ces résultats sont discutés dans le chapitre suivant.

Mutations systématiques

Pour augmenter artificiellement le nombre de conformations pour chaque séquence, toutes les combinaisons de mutations sont calculées pour transformer les purines (R) entre elles et les pyrimidines (Y) entre elles (voir section 1.1.3). Ceci repose sur l'hypothèse que de telles mutations, qui ajoutent ou enlèvent un atome lourd, ont un impact négligeable sur la structure globale du tri-nucléotide. Nous avons vérifié cette hypothèse par la quatrième analyse donnée en exemple dans ce chapitre.

Pour optimiser le calcul des clusters, toutes les bases sont d'abord mutées en A ou C, le clustering (voir section 2.2.3) est appliqué, puis les fragments regroupés sont mutés en retour pour toutes les combinaisons de mutations U/T vers C et A vers G. Pour les fragments tri-nucléotidiques, chaque tri-nucléotide est muté en 8 nouvelles séquences, donc la taille des ensembles est multipliée par environ 8 (un peu moins à cause des conformations redondantes entre les différentes séquences).

2.3 Exemples d'application

La liste des identifiants PDB donnée en entrée pour obtenir les résultats exposés dans cette section est l'ensemble des structures pdb qui contiennent des chaînes de protéines et des chaînes d'ARN, mais pas d'ADN, où soit la résolution est inférieure à 3 Å soit la méthode est la RMN. Les ribosomes sont retirés de cette base de données en raison de leur taille. Toutes les statistiques et figures peuvent être trouvées dans le notebook `figures_protnaff.ipynb` sur la page web GitHub de protNAff `https://github.com/isaureCdB/protNAff`. La bibliothèque de fragments peut être téléchargée à l'adresse suivante `https://zenodo.org/record/6483823#.YmbVhFxByV4`.

2.3.1 Spécificité de la séquence des conformations de l'ARN

Nous avons exploré dans quelle mesure les conformations tri-nucléotidiques sont spécifiques à la séquence pour un motif purine/pyrimidine (R/Y) donné. ProtNAff applique toutes les mutations possibles de purine à purine et de pyrimidine à pyrimidine, et regroupe les fragments séparément pour chaque motif (voir la section 2.2.3).

Pour répondre à notre question, nous avons analysé la population des clusters de 1 Å de rayon (RMSD) en termes de séquence originale des fragments. Si un cluster ne contient que des fragments avec une même séquence, alors nous pouvons considérer que cette conformation est spécifique à la séquence.

Les détails de cette analyse peuvent être trouvés dans le notebook appelé `sequence-specific_conformations.ipynb`. Pour un cluster de n fragments, la probabilité d'avoir un cluster pur (ne contenant qu'une seule séquence) diminue avec n. Dans nos clusters à 1 Å RMSD, pour tous les motifs R/Y-R/Y-R/Y ensemble, nous avons trouvé moins de 5% de clusters purs pour $n \geqslant 3$ (32/715), moins de 3% (11/461) pour $n \geqslant 4$ et moins de 2% (6/346) pour $n \geqslant 5$. Ainsi, la proportion de clusters de séquences uniques reste très faible, ce qui signifie que la grande majorité des conformations (définies à 1 Å) ne sont pas spécifiques d'une séquence pour un motif purine/pyrimidine donné.

Ces résultats soulignent la pertinence d'effectuer des mutations lors de la création de bibliothèques de tri-nucléotides avec une exhaustivité conformationnelle maximale, au détriment de l'autorisation de quelques conformations irréelles (correspondant aux conformations mutées alors que spécifiques à la séquence d'origine). Pour les applications où il est plus important d'être exact que d'être exhaustif, les fragments mutés doivent être éliminés par filtrage. Notez que ces résultats ne s'appliquent pas à d'autres types de fragments où un appariement de bases peut se produire, et pour lesquels le modèle de mutation doit être adapté. Un tel exemple est disponible

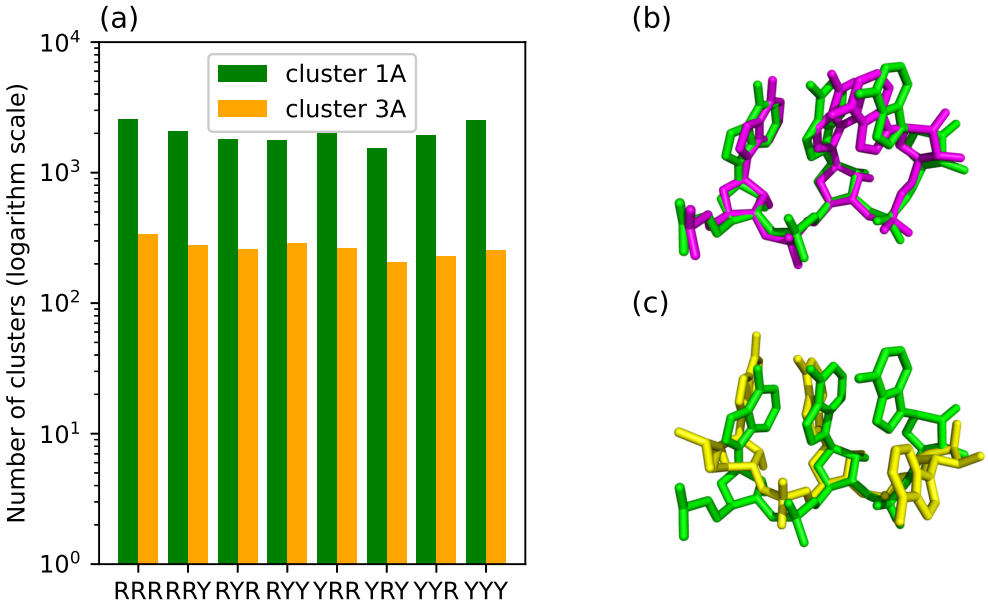


FIGURE 2.2 – Clustering avec différents seuils de RMSD : (a) nombre de clusters en fonction de la largeur des clusters (1 ou 3 Å), en échelle logarithmique ; exemples aléatoires d'alignement de deux tri-nucléotides différents à 0,9 Å RMSD (b) ou 2,5 Å RMSD (c), pour illustrer la diversité conformationnelle à l'intérieur des clusters en fonction de leur largeur.

dans le dépôt Github de protNAff (voir `create_helices_library.sh`).

2.3.2 Diversité conformationnelle des fragments d'ARN à différentes échelles

La plupart des utilisations de la base de données et des bibliothèques nécessitent un traitement approprié et spécifique au contexte des redondances. Au niveau des fragments, différents seuils de regroupement peuvent moduler la granularité de la représentation de l'espace conformationnel. Par conséquent, l'implémentation de protNAff permet à l'utilisateur d'ajuster le rayon des clusters en fonction de son objectif. Cela permet de trouver un équilibre entre la taille maximale de la bibliothèque et la précision souhaitée de la discrétisation de l'espace conformationnel pour ce type de fragment.

À titre d'exemple, le nombre de clusters de tri-nucléotides obtenus avec un 1 Å ou 3 Å comme seuil de cRMSD pour les différentes séquences pyrimidine/purine est présenté dans la figure 2.2. La diversité des clusters à 1 Å et 3 Å est présentée dans la figure 2.3. Les chiffres indiquent le nombre de clusters d'une certaine taille parmi les différents motifs Y/R. La distribution pour les motifs est similaire pour les deux seuils de clustering (la différence est qu'il y a moins de clusters, mais qu'ils ont plus d'individus à 3 Å qu'à 1 Å, comme attendu).

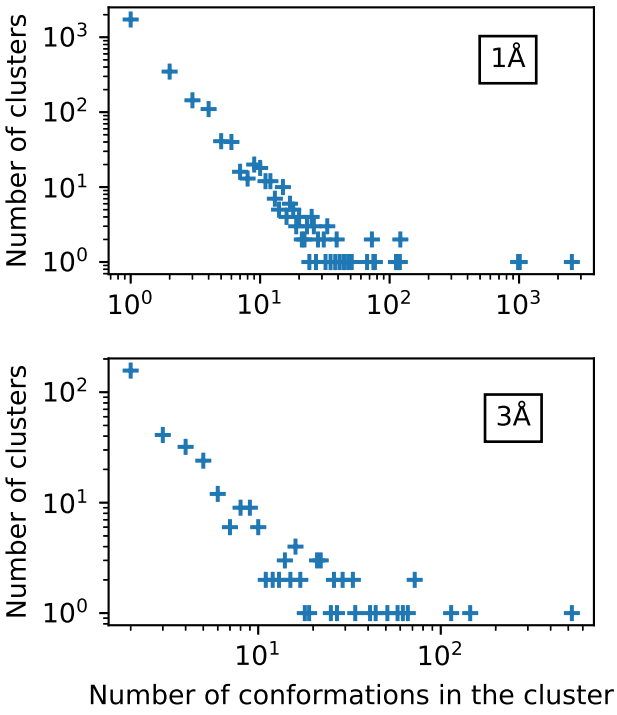


FIGURE 2.3 – Nombre de clusters en fonction du nombre d'individus par cluster après clustering à 1 Å (en haut) et 3 Å (en bas) pour l'ARN. Les deux axes sont en échelle logarithmique

2.3.3 Conformations locales de l'ARN induites par la liaison des protéines

On pose l'hypothèse (nommée H1) selon laquelle l'ARN en contact avec les protéines peut adopter des conformations locales spécifiques par rapport à l'ARN non lié, c'est-à-dire des conformations induites par la liaison à la protéine (avec une énergie d'interaction du ligand plus élevée pour minimiser celle du complexe). Cette hypothèse peut être testée en regroupant les conformations des fragments sur la base de leurs différences structurelles. Si certains clusters ne contiennent que des fragments en contact avec une protéine, la conformation correspondante est probablement spécifiquement induite par la protéine. La figure 2.4 explicite la distinction entre les différents types de clusters.

De plus, on pourrait s'attendre à ce que ces conformations induites par les protéines soient plus nombreuses dans les ARN simple brin, qui sont moins contraints par les liaisons intra-ARN (nommons H2 cette hypothèse). Ceci peut être testé en distinguant les clusters constitués exclusivement de fragments d'ARN simple brin.

Pour vérifier ces deux hypothèses, tous les tri-nucléotides d'ARN qui se chevauchent ont été extraits de l'ensemble complet d'ARN liés à des protéines dans la PDB, et une classification avec un seuil de 1,0 Å RMSD a été effectuée. Nous définissons plusieurs types de fragments et plusieurs

types de clusters. Les fragments sont entièrement simple brin (les trois nucléotides sont simple brin), ou alors ils sont entièrement double brin ou enfin mixtes. De même manière les fragments sont entièrement non liés si les trois nucléotides ne sont pas en contact avec une protéine, et sinon ils sont liés. Pour les clusters, les clusters seulement simple brin contiennent uniquement des fragments entièrement simple brin. Les clusters seulement double brin contiennent uniquement des fragments entièrement double brin. Les clusters contenant tout type de fragments sont des clusters mixtes. Pour les contacts au niveau des clusters, nous distinguons trois états : seulement liés (tous les fragments sont en contact), seulement non liés (aucun des fragments n'est en contact), et mixte. Ces distinctions sur les contacts et les appariements nous donnent 9 types de clusters possibles, schématisés en figure 2.5. Nous notons : informations sur appariements/informations sur contacts, ainsi un cluster mixte/mixte contient des fragments simple brin et double brin, qui sont pour certains liés et pour d'autres non liés.

Les clusters contenant moins de trois conformations ont été considérés comme non représentatifs et donc retirés. Puisque 39% de tous les fragments d'ARN ne se trouvent pas à l'interface de la protéine, nous avons supposé que cet ensemble couvre toutes les conformations possibles de l'ARN non lié. Les résultats sont présentés dans la table 2.1.

Le faible nombre de clusters entièrement double brin était attendu ($< 5\%$ des clusters), car les conformations rencontrées dans les hélices peuvent également être adoptées par un fragment contenant au moins un nucléotide simple brin (fragment considéré comme mixte ici). De plus, les conformations double brin sont moins diversifiées que le simple brin car elles sont beaucoup plus contraintes. Les clusters mixtes/mixtes sont les clusters contenant les conformations les plus fréquentes, ce qui explique qu'ils regroupent la majorité des fragments. Environ la moitié des clusters sont constitués uniquement de tri-nucléotides en contact avec une protéine, ce qui suggère que ces conformations sont spécifiquement (à une résolution de 1 Å) induites par le contact avec la protéine.

Comme attendu selon l'hypothèse H2, le pourcentage de conformations spécifiquement induites par la liaison protéique est plus élevé pour les clusters constitués exclusivement de fragments simple brin, pour tous les motifs purine/pyrimidine. Un test Z a été appliqué entre la population de fragments simple brin et le reste. Pour les huit motifs possibles, la *p-value* était inférieure à 10^{-5} et le test était positif.

Il est intéressant de noter que les clusters entièrement simple brin ou entièrement double brin sont soit entièrement liés, soit entièrement non non liés avec une protéine. Cela signifie que ni les conformations spécifiques du simple brin ni les conformations spécifiques du double brin n'existent dans les deux états : liés et non liés. En d'autres termes, le contact d'un tel fragment

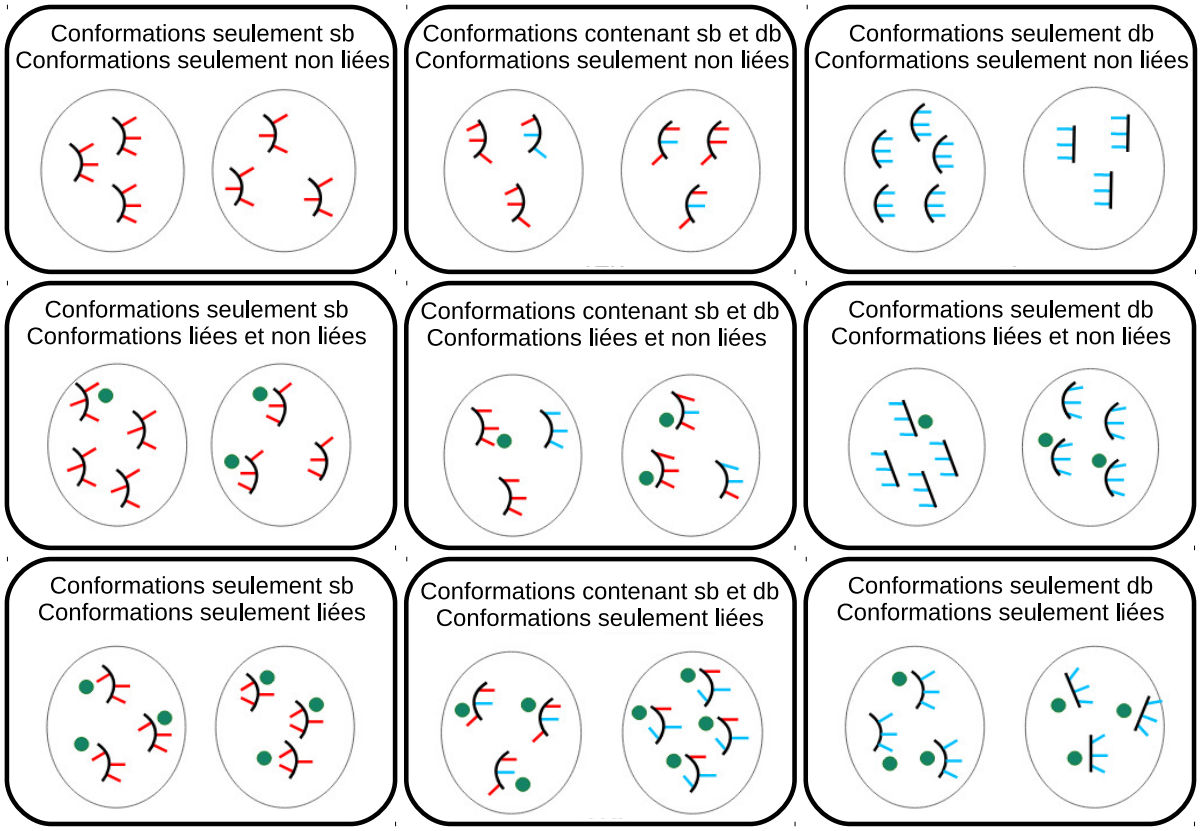


FIGURE 2.4 – Définitions pour les clusters simple brin (sb) / double brin (db) / mixte et lié / non lié / mixte. Simple brin signifie que les trois nucléotides ne sont pas appariés, double brin signifie que les trois nucléotides sont appariés. Lié signifie qu'au moins un nucléotide est en contact (distance inférieure à 5 Å) avec la protéine. Non lié signifie qu'il n'y a pas de contact avec la protéine.

avec une protéine induit toujours une conformation différente de celles qui peuvent être observées dans l'ARN non lié.

Toutes les observations faites précédemment peuvent être faites pour les clusters à 3 Å. Une différence importante est qu'il n'y a presque pas de clusters entièrement double brin (seulement 2 clusters pour les 8 séquences). De plus, nous aurions dû utiliser ces informations pour retirer les clusters contenant uniquement des conformations non liées de nos bibliothèques utilisées pour le docking (puisque ces conformations n'existent pas en contact des protéines).

2.3.4 Taille des interfaces dans les boucles en épingle à cheveux d'ARN liées à des protéines

Nous avons utilisé un filtre pour déterminer si la taille de l'interface protéine-ARN dans les boucles en épingle à cheveux est en corrélation avec la taille de ces boucles. Le filtre sélectionne

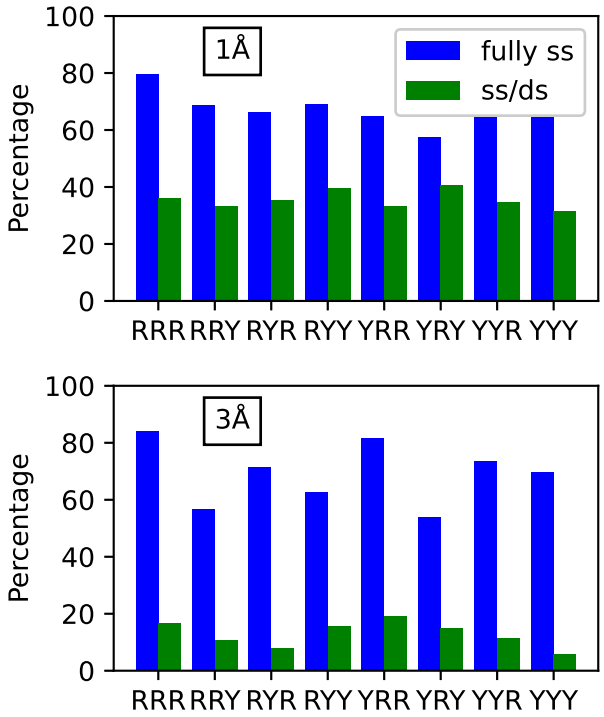


FIGURE 2.5 – Pourcentage de conformations entièrement simple brin induites par le contact avec la protéine et les conformations qui ne sont pas simple brin. En haut, les conformations sélectionnées à 1 Å, en bas, les conformations sélectionnées à 3 Å.

RRR 1 Å	sb	mixte	db	total	RRR 3 Å	sb	mixte	db	total
lié	184(1255)	146(823)	9(47)	339(2125)	lié	32(197)	26(189)	0(0)	58(386)
mixte	0(0)	194(12286)	0(0)	194(12286)	mixte	0(0)	119(16860)	0(0)	119(16860)
non lié	47(315)	72(427)	8(62)	127(804)	non lié	6(24)	10(81)	0(0)	16(105)
total	231(1570)	412(13536)	17(109)	660(15215)	total	38(221)	155(17130)	0(0)	193(17351)
RRY 1 Å	sb	mixte	db	total	RRY 3 Å	sb	mixte	db	total
lié	101(712)	107(562)	8(45)	216(1319)	lié	13(108)	13(72)	0(0)	26(180)
mixte	0(0)	149(10432)	0(0)	149(10432)	mixte	0(0)	99(13933)	0(0)	99(13933)
non lié	46(242)	73(466)	9(43)	128(751)	non lié	10(65)	10(39)	0(0)	20(104)
total	147(954)	329(11460)	17(88)	493(12502)	total	23(173)	122(14044)	0(0)	145(14217)
RYR 1 Å	sb	mixte	db	total	RYR 3 Å	sb	mixte	db	total
lié	108(782)	91(418)	6(26)	205(1226)	lié	20(139)	9(75)	0(0)	29(214)
mixte	0(0)	115(8264)	0(0)	115(8264)	mixte	0(0)	96(11550)	0(0)	96(11550)
non lié	55(400)	50(415)	13(82)	118(897)	non lié	8(61)	9(60)	0(0)	17(121)
total	163(1182)	256(9097)	19(108)	438(10387)	total	28(200)	114(11685)	0(0)	142(11885)
RYY 1 Å	sb	mixte	db	total	RYY 3 Å	sb	mixte	db	total
lié	78(444)	83(426)	7(37)	168(907)	lié	15(77)	18(164)	0(0)	33(241)
mixte	0(0)	94(9370)	0(0)	94(9370)	mixte	0(0)	89(12021)	0(0)	89(12021)
non lié	35(206)	40(295)	4(19)	79(520)	non lié	9(54)	7(33)	0(0)	16(87)
total	113(650)	217(10091)	11(56)	341(10797)	total	24(131)	114(12218)	0(0)	138(12349)
YRR 1 Å	sb	mixte	db	total	YRR 3 Å	sb	mixte	db	total
lié	106(558)	122(738)	9(39)	237(1335)	lié	22(113)	20(153)	1(3)	43(269)
mixte	0(0)	168(8706)	0(0)	168(8706)	mixte	0(0)	86(12575)	0(0)	86(12575)
non lié	58(402)	86(601)	9(67)	153(1070)	non lié	5(38)	3(18)	0(0)	8(56)
total	164(960)	376(10045)	18(106)	558(11111)	total	27(151)	109(12746)	1(3)	137(12900)
YRY 1 Å	sb	mixte	db	total	YRY 3 Å	sb	mixte	db	total
lié	83(631)	83(477)	5(34)	171(1142)	lié	7(65)	13(77)	1(4)	21(146)
mixte	0(0)	89(7072)	0(0)	89(7072)	mixte	0(0)	77(10041)	0(0)	77(10041)
non lié	62(436)	35(222)	5(48)	102(706)	non lié	6(30)	2(9)	0(0)	8(39)
total	145(1067)	207(7771)	10(82)	362(8920)	total	13(95)	92(10127)	1(4)	106(10226)
YYR 1 Å	sb	mixte	db	total	YYR 3 Å	sb	mixte	db	total
lié	86(484)	84(437)	7(27)	177(948)	lié	14(58)	13(70)	0(0)	27(128)
mixte	0(0)	112(7879)	0(0)	112(7879)	mixte	0(0)	92(10962)	0(0)	92(10962)
non lié	39(282)	57(321)	3(21)	99(624)	non lié	5(35)	9(69)	0(0)	14(104)
total	125(766)	253(8637)	10(48)	388(9451)	total	19(93)	114(11101)	0(0)	133(11194)
YYY 1 Å	sb	mixte	db	total	YYY 3 Å	sb	mixte	db	total
lié	143(912)	81(391)	4(15)	228(1318)	lié	16(84)	7(28)	0(0)	23(112)
mixte	0(0)	133(10183)	0(0)	133(10183)	mixte	0(0)	101(14277)	0(0)	101(14277)
non lié	61(376)	52(321)	0(0)	113(697)	non lié	7(43)	10(43)	0(0)	17(86)
total	204(1288)	266(10895)	4(15)	474(12198)	total	23(127)	118(14348)	0(0)	141(14475)

TABLE 2.1 – Cette table contient le nombre de clusters à 1 Å et à 3 Å (nombre de fragments) pour les différentes possibilités présentées dans la figure 2.4, dans le cadre de l'ARN.

toutes les structures en épingle à cheveux où la boucle simple brin est en contact (< 5 Å) avec la protéine (voir `filters/query_hairpin.py` sur GitHub). Ici, la possibilité est donnée à l'utilisateur de choisir la taille minimale/maximale de la boucle simple brin et de l'hélice double brin. Dans notre cas, nous avons choisi une longueur minimale de 3 nucléotides pour la boucle, c'est la longueur minimale possible. Pour les hélices, nous avons choisi une longueur minimale de 3 paires de bases, c'est la taille de nos bibliothèques d'hélices (voir section 4.4). De plus, des hélices plus courtes induisent des structures très particulières. Avec ces valeurs, nous obtenons 1160 épingles à cheveux qui appartiennent à 437 structures pdb. Ce filtre nous permet d'étudier certaines caractéristiques de liaison aux protéines des épingles à cheveux d'ARN, comme le

nombre de nucléotides en contact avec la protéine (simple brin et double brin) pour des épingles à cheveux de différentes tailles. Nous avons étudié les valeurs aberrantes correspondant à une boucle trop longue ou à une partie en contact trop petite. La structure 7KA0 (structure d'une Phe-ARNt synthétase en complexe avec un ARNt) a une boucle de 11 nucléotides de long avec seulement 2 nucléotides en contact, ce qui peut être expliqué par une interaction double brin avec une autre chaîne d'ARN. Un graphique comparant la taille des boucles et le nombre de nucléotides en contact avec la protéine est présenté dans la figure 2.6. Nous avons constaté que 75% des épingles à cheveux ont une boucle de 7 nucléotides ou moins et qu'un tiers des épingles à cheveux a tous ses nucléotides à moins de 5 Å d'une protéine.

Les deux boucles les plus longues possèdent 16 nucléotides, dont pour l'une 11 sont liés à la protéine. Pour 6ZDQ (structure d'une télomérase tronquée liée à un ARN modèle), seulement 6 nucléotides sont en contact avec la protéine, ce qui est peu considérant la taille de la boucle (après avoir regardé la structure sur PyMol, nous avons constaté que seul le squelette de la chaîne d'ARN interagit avec la protéine). Pour les boucles d'une longueur allant jusqu'à 10 nucléotides, la plupart des nombres possibles de nucléotides liés aux protéines sont observés. Au-dessus de 10 nucléotides, presque toutes les boucles ont la moitié ou plus de leurs nucléotides liés. Au maximum, neuf nucléotides ne sont pas en contact avec la protéine, ce qui signifie que pour les petites boucles, une forte proportion peut être éloignée de la protéine. Dans le cas de longues boucles, la proportion de la boucle qui n'est pas en contact devient petite. Cela peut venir du fait que si une longue boucle simple brin n'est pas en contact avec une protéine, alors elle est trop flexible pour avoir une bonne résolution.

2.3.5 Analyses sur l'ADN

Nous avons reproduit les mêmes chiffres et statistiques sur l'ADN, tous les résultats se trouvent ci-après. Pour l'ADN, la requête sur la PDB est : tous les complexes ADN-protéine (sans ARN) avec une résolution d'au plus de 3 Å. Cette requête donne 1906 structures. Ensuite, le pipeline est appliqué à ces structures et les figures peuvent être reproduites à l'aide d'un notebook sur le GitHub (`figures_dna_protnaff.ipynb`).

La diversité conformationnelle des fragments d'ADN à différentes échelles

Pour l'ADN, il y a moins de clusters que pour l'ARN, ce qui signifie moins de diversité conformationnelle, mais le facteur entre les clusters à 1 Å et ceux à 3 Å est plus faible (voir figure 2.7). Les distributions des clusters en termes de taille sont présentées dans la figure 2.8 : on peut voir qu'il y a moins de clusters que pour l'ARN et qu'ils sont moins peuplés. Nous

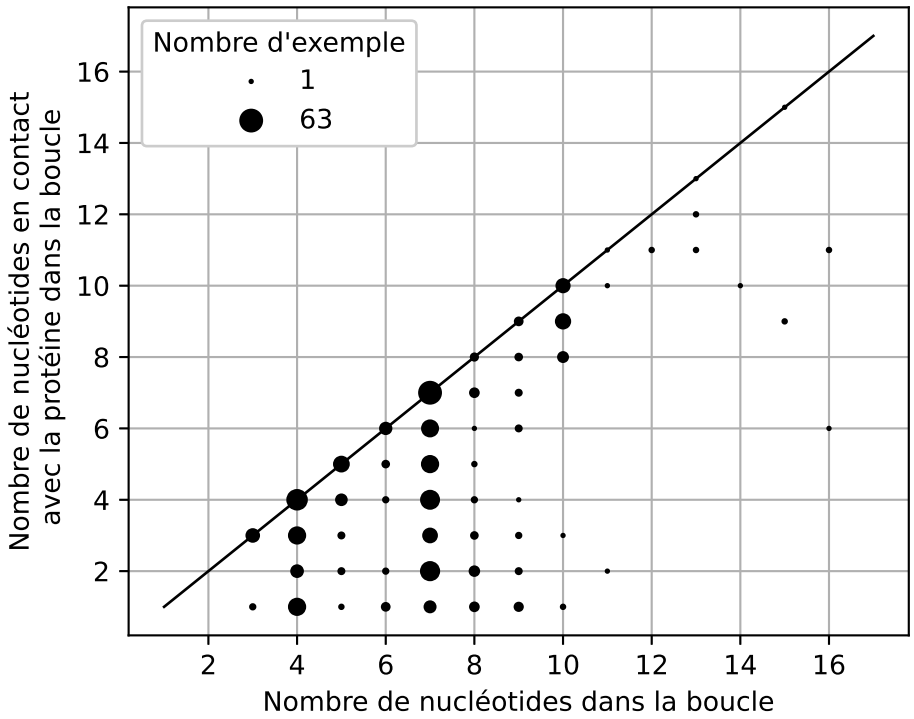


FIGURE 2.6 – Nombre de nucléotides de la boucle en épingle à cheveux d'ARN en contact avec la protéine en fonction du nombre de nucléotides de la boucle. La taille des points indique le nombre d'épingles à cheveux dans cette configuration, de 1 à 63.

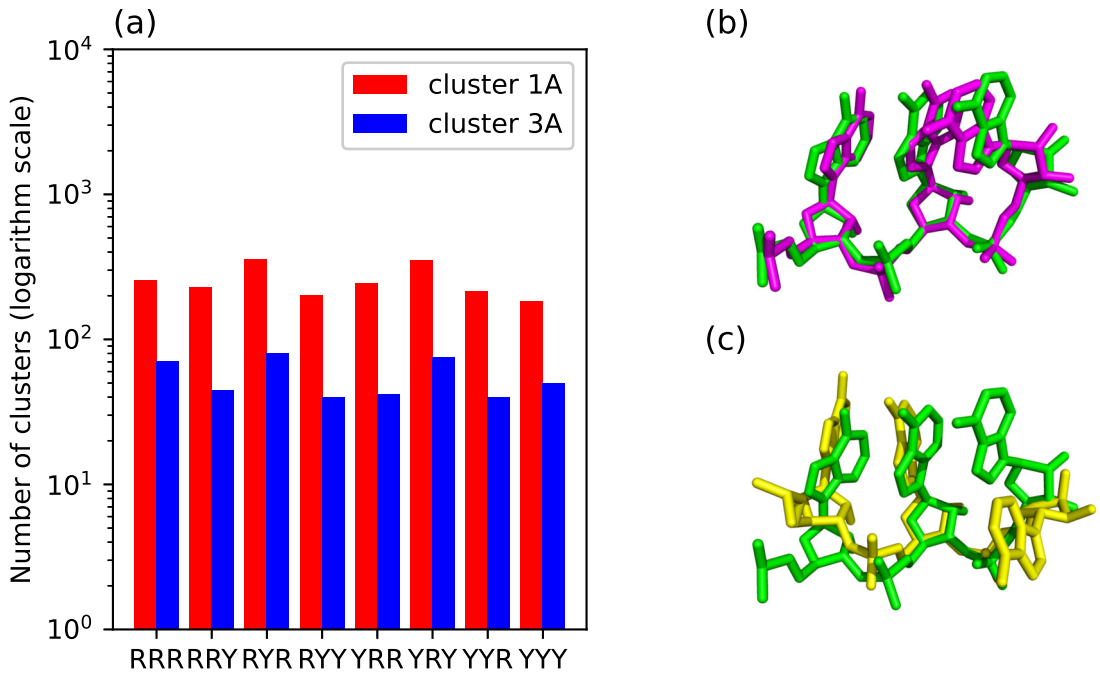


FIGURE 2.7 – Clustering avec différents seuils de RMSD pour l'ADN : (a) nombre de clusters en fonction de la largeur des clusters (1 ou 3 Å), en échelle logarithmique ; exemples aléatoires d'alignement de deux tri-nucléotides différents à 0,9 Å RMSD (b) ou 2,5 Å RMSD (c), pour illustrer la diversité conformationnelle à l'intérieur des clusters en fonction de leur largeur.

rappelons que nous enlevons les redondances à 0.2 Å, c'est pourquoi il y a si peu de membres dans les clusters.

Conformations locales de l'ADN induites par la liaison des protéines

Pour l'ADN, il semble que le contact avec la protéine induit des conformations différentes pour chacune des séquences. En effet, la moitié des clusters à 1 Å ne sont observés qu'en contact avec la protéine. Du fait du faible nombre de clusters à 3 Å, les résultats ne sont pas interprétés. De même pour la figure 2.9, le nombre de clusters simple brin est trop faible pour pouvoir faire une interprétation.

Taille des interfaces dans les boucles en épingle à cheveux d'ADN liées à des protéines

Il y a beaucoup moins d'épingles à cheveux pour l'ADN que pour l'ARN, cependant la distribution des tailles possibles est aussi large, les deux plus longues boucles faisant 16 nucléotides, comme pour l'ARN. Ces 2 boucles viennent de 2 structures différentes, 6ZVH qui est un ribosome

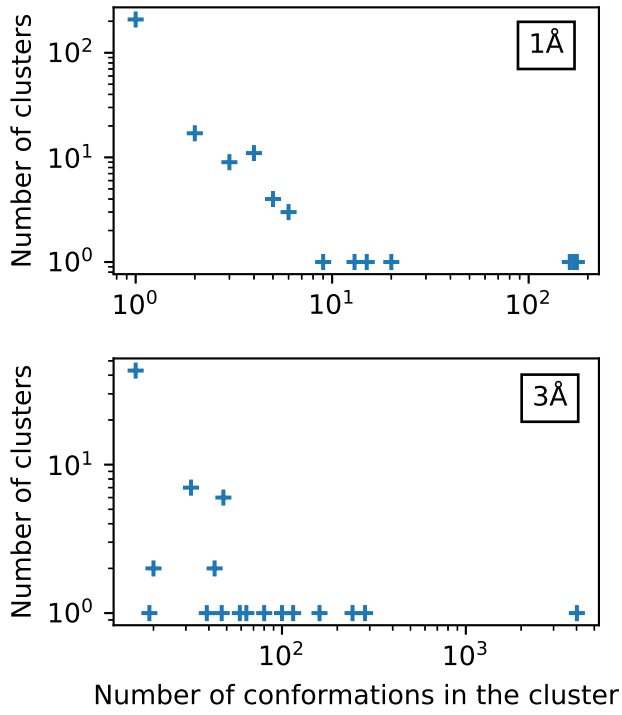


FIGURE 2.8 – Nombre de clusters en fonction du nombre d'individus par cluster après clustering à 1 Å (en haut) et 3 Å (en bas) pour l'ADN. Les deux axes sont en échelle logarithmique

et 6RTI qui est une hydrolase humaine en contact avec l'aptamère A9g. La taille des boucles en fonction de leurs contacts avec une protéine est montrée en figure 2.10.

2.3.6 Mémoire et complexité de calcul

Tous les temps de calcul ont été obtenus sur 8 CPU Intel(R) Xeon(R) Silver 4114 @ 2.20 GHz, avec 16Go de RAM. La création de la base de données n'est pas une étape parallélisée. Dans nos tests, le téléchargement des fichiers de complexes ARN-protéines de la PDB pour la création de la base de données prend entre 1 et 2 heures. L'ensemble des fichiers pdb téléchargés correspond à un volume de 1,7 GB. Enfin, la mise à jour de la base de données, c'est-à-dire l'ajout d'une nouvelle structure, prend quelques minutes en fonction de la taille de la structure, puis la création d'une bibliothèque de fragments tri-nucléotidiques à partir de la base de données précédente correspond à un calcul de 4 heures. À cette étape, la création des clusters est parallélisée : chaque séquence est associée à un *thread*. Enfin, le temps de calcul pour la sélection d'un sous-ensemble de structures dépend largement du filtre à appliquer. Pour les deux filtres présentés ici, le premier prend 30s et le second prend 2min.

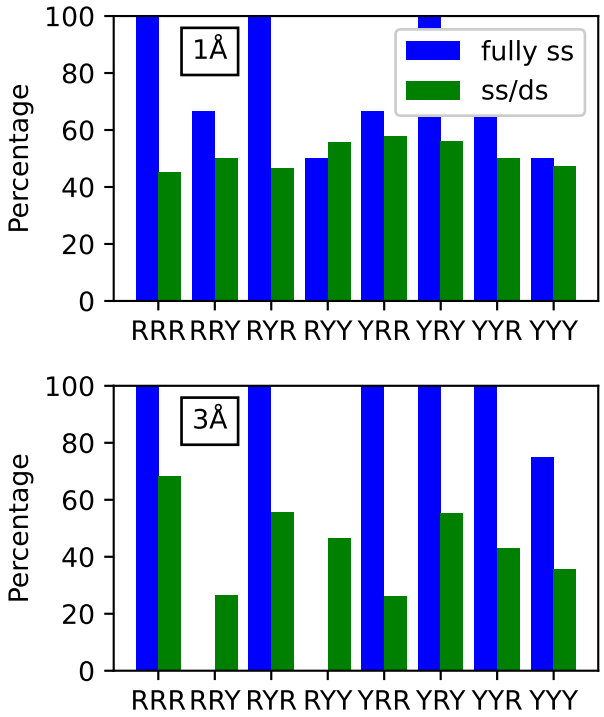


FIGURE 2.9 – Pourcentage de conformations entièrement simple brin induites par le contact avec la protéine et les conformations qui ne sont pas simple brin. En haut, les conformations sélectionnées à 1 Å, en bas, les conformations sélectionnées à 3 Å pour l'ADN.

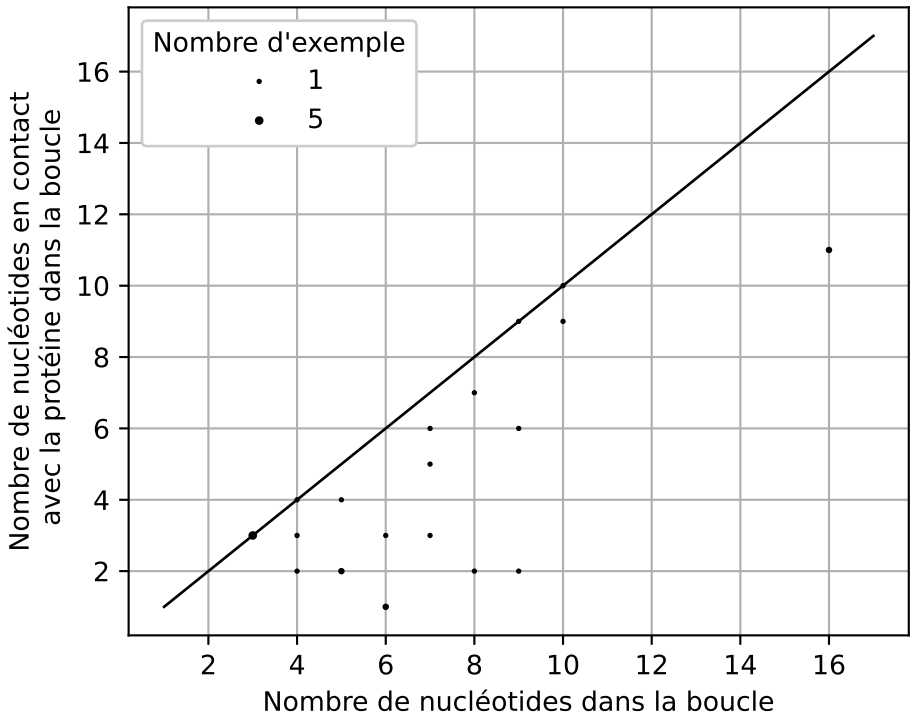


FIGURE 2.10 – Nombre de nucléotides de la boucle en épingle à cheveux d'ADN en contact avec la protéine en fonction du nombre de nucléotides de la boucle. La taille des points indique le nombre d'épingles à cheveux dans cette configuration, de 1 à 5.

RRR 1 Å	sb	mixte	db	total	RRR 3 Å	sb	mixte	db	total
lié	5(20)	7(48)	7(27)	19(95)	lié	1(3)	13(83)	0(0)	14(86)
mixte	0(0)	17(722)	0(0)	17(722)	mixte	0(0)	6(916)	0(0)	6(916)
non lié	0(0)	0(0)	0(0)	0(0)	non lié	0(0)	0(0)	0(0)	0(0)
total	5(20)	24(770)	7(27)	36(817)	total	1(3)	19(999)	0(0)	20(1002)
RRY 1 Å	sb	mixte	db	total	RRY 3 Å	sb	mixte	db	total
lié	2(6)	13(118)	4(15)	19(139)	lié	0(0)	4(29)	0(0)	4(29)
mixte	0(0)	17(825)	0(0)	17(825)	mixte	0(0)	11(1124)	0(0)	11(1124)
non lié	1(19)	0(0)	0(0)	1(19)	non lié	1(19)	0(0)	0(0)	1(19)
total	3(25)	30(943)	4(15)	37(983)	total	1(19)	15(1153)	0(0)	16(1172)
RYR 1 Å	sb	mixte	db	total	RYR 3 Å	sb	mixte	db	total
lié	4(24)	14(98)	12(81)	30(203)	lié	1(5)	14(116)	1(1)	16(122)
mixte	0(0)	29(564)	0(0)	29(564)	mixte	0(0)	12(916)	0(0)	12(916)
non lié	0(0)	1(6)	0(0)	1(6)	non lié	0(0)	0(0)	0(0)	0(0)
total	4(24)	44(668)	12(81)	60(773)	total	1(5)	26(1032)	1(1)	28(1038)
RYY 1 Å	sb	mixte	db	total	RYY 3 Å	sb	mixte	db	total
lié	2(7)	9(49)	6(38)	17(94)	lié	0(0)	6(45)	1(10)	7(55)
mixte	0(0)	12(930)	0(0)	12(930)	mixte	0(0)	8(1148)	0(0)	8(1148)
non lié	2(21)	0(0)	0(0)	2(21)	non lié	0(0)	0(0)	0(0)	0(0)
total	4(28)	21(979)	6(38)	31(1045)	total	0(0)	14(1193)	1(10)	15(1203)
YRR 1 Å	sb	mixte	db	total	YRR 3 Å	sb	mixte	db	total
lié	2(13)	11(79)	8(29)	21(121)	lié	2(7)	4(24)	1(6)	7(37)
mixte	0(0)	14(789)	0(0)	14(789)	mixte	0(0)	14(1098)	0(0)	14(1098)
non lié	1(20)	0(0)	0(0)	1(20)	non lié	0(0)	0(0)	0(0)	0(0)
total	3(33)	25(868)	8(29)	36(930)	total	2(7)	18(1122)	1(6)	21(1135)
YRY 1 Å	sb	mixte	db	total	YRY 3 Å	sb	mixte	db	total
lié	1(9)	14(97)	19(132)	34(238)	lié	2(8)	14(103)	2(47)	18(158)
mixte	0(0)	26(568)	0(0)	26(568)	mixte	0(0)	13(928)	0(0)	13(928)
non lié	0(0)	0(0)	0(0)	0(0)	non lié	0(0)	0(0)	0(0)	0(0)
total	1(9)	40(665)	19(132)	60(806)	total	2(8)	27(1031)	2(47)	31(1086)
YYR 1 Å	sb	mixte	db	total	YYR 3 Å	sb	mixte	db	total
lié	5(17)	8(55)	4(30)	17(102)	lié	1(3)	5(38)	1(11)	7(52)
mixte	0(0)	12(824)	0(0)	12(824)	mixte	0(0)	8(1074)	0(0)	8(1074)
non lié	1(20)	0(0)	0(0)	1(20)	non lié	0(0)	0(0)	0(0)	0(0)
total	6(37)	20(879)	4(30)	30(946)	total	1(3)	13(1112)	1(11)	15(1126)
YYY 1 Å	sb	mixte	db	total	YYY 3 Å	sb	mixte	db	total
lié	2(8)	5(29)	3(31)	10(68)	lié	3(16)	5(43)	0(0)	8(59)
mixte	0(0)	9(905)	0(0)	9(905)	mixte	0(0)	9(1098)	0(0)	9(1098)
non lié	2(38)	0(0)	0(0)	2(38)	non lié	1(3)	0(0)	0(0)	1(3)
total	4(46)	14(934)	3(31)	21(1011)	total	4(19)	14(1141)	0(0)	18(1160)

TABLE 2.2 – Cette table contient le nombre de clusters à 1 Å et à 3 Å (nombre de fragments) pour les différentes possibilités présentées dans la figure 2.4, dans le cadre de l'ADN.

2.4 Conclusions sur ProtNAff

Dans ce chapitre, nous avons présenté notre outil ProtNAff, que nous avons essayé de faire aussi généralisable que possible. Quelques exemples d'application ont aussi été présentés, montrant qu'il est possible de produire de très nombreux résultats statistiques divers sur les bases de données et les bibliothèques de fragments.

Nous avions vu à la section 1.4.4 l'importance des bibliothèques de fragments pour la méthode d'assemblage des fragments. Dans ce chapitre, nous avons présenté la méthode de clustering qui a été utilisé historiquement. Dans le chapitre suivant, nous présentons une méthode de clustering plus adaptée à notre utilisation des fragments.

Chapitre 3

Calcul d'ϵ-réseaux sur des ensembles finis

Nous avons vu au chapitre précédent que l'inférence d'ensembles de prototypes de conformations jouait un rôle central dans les méthodes de modélisation des complexes protéine-ARN par assemblage de fragments structuraux. Un ensemble de prototypes doit satisfaire deux contraintes contradictoires : être représentatif et de cardinalité aussi faible que possible.

Lorsque le critère de représentativité est une distance, le problème se réduit à celui de l'inférence d'un *ϵ-réseau* de cardinalité minimale. La condition nécessaire et suffisante de faisabilité est que le domaine soit totalement borné. Les fragments structuraux vivent dans un espace métrique et sont toujours de cardinalité finie. Dans ce chapitre, nous traitons donc la question de l'inférence d'ϵ-réseaux d'ensembles finis de points dans un espace métrique. Dans ce contexte, les méthodes de classification ascendante hiérarchique semblent fournir une solution satisfaisante. Nous introduisons une nouvelle méthode dédiée : *Radius*. Elle fournit les meilleurs résultats expérimentaux. Comme la question de l'inférence d'ϵ-réseaux sur des ensembles finis de points demeure ouverte, nous commençons par l'aborder dans toute sa généralité. Nous comparerons notre méthode dédiée à l'inférence faite par les ensembles dominants sur les représentations sous forme de graphe.

3.1 Caractérisation du problème

Le problème d'intérêt est un problème de géométrie computationnelle. Nous l'introduisons à présent.

3.1.1 Cadre théorique

Nous caractérisons d'abord les données. Elles constituent un sous-ensemble fini d'un espace métrique $(\mathcal{X}, \rho_{\mathcal{X}})$. Soit $s_{\mathcal{X}^n} = \{x_i : 1 \leqslant i \leqslant n\}$ ce sous-ensemble. Nous considérons que si l'on dispose de la métrique, il est toujours possible de calculer la distance entre deux points de $\mathcal{X}$ et donc de calculer la matrice de distance $D \in \mathcal{M}_{n,n}(\mathbb{R})$ de terme général $\rho_{\mathcal{X}}(x_i, x_j)$.

L'introduction du concept d'ϵ-réseau (d'ϵ-couverture et de nombre de couverture) est généralement attribuée à Kolmogorov et Tihomirov [45]. Dans ce qui suit, $(\mathcal{E}, \rho)$ désigne un espace métrique et $\mathcal{E}'$ est un sous-ensemble de $\mathcal{E}$ qui est totalement borné.

Définition 1 (ϵ-couverture, ϵ-réseau et nombre de couverture)**.** *Pour $\epsilon \in \mathbb{R}_+^*$, une* ϵ-couverture *de $\mathcal{E}'$ est une couverture de $\mathcal{E}'$ composée de boules ouvertes de rayon ϵ dont les centres appartiennent à $\mathcal{E}$. Ces centres forment un* ϵ-réseau *de $\mathcal{E}'$. Un* ϵ-réseau propre *de $\mathcal{E}'$ est un ϵ-réseau de $\mathcal{E}'$ inclus dans $\mathcal{E}'$. Le* nombre de couverture $\mathcal{N}(\epsilon, \mathcal{E}', \rho)$ *est la plus petite cardinalité des ϵ-réseaux de $\mathcal{E}'$:*

$$\mathcal{N}(\epsilon, \mathcal{E}', \rho) = \min\left\{\left|\mathcal{E}''\right| : \left(\mathcal{E}'' \subset \mathcal{E}\right) ET \left(\forall e \in \mathcal{E}',\ \rho\left(e, \mathcal{E}''\right) < \epsilon\right)\right\}.$$

$\mathcal{N}^{int}(\epsilon, \mathcal{E}', \rho)$ désigne un nombre de couverture de $\mathcal{E}'$ obtenu pour les ϵ-réseaux propres. Nous avons donc :

$$\mathcal{N}^{int}(\epsilon, \mathcal{E}', \rho) = \min\left\{\left|\mathcal{E}''\right| : \left(\mathcal{E}'' \subset \mathcal{E}'\right) ET \left(\forall e \in \mathcal{E}',\ \rho\left(e, \mathcal{E}''\right) < \epsilon\right)\right\}.$$

Le problème considéré est le suivant :

Problème 1. *Etant donnés $s_{\mathcal{X}^n}$ et $\epsilon \in \mathbb{R}_+^*$, trouver un ϵ-réseau $\mathcal{C}(\epsilon) = \{c_j(\epsilon) : 1 \leqslant j \leqslant |\mathcal{C}(\epsilon)|\}$ de $s_{\mathcal{X}^n}$ avec une cardinalité $|\mathcal{C}(\epsilon)|$ aussi proche que possible de $\mathcal{N}(\epsilon, s_{\mathcal{X}^n}, \rho_{\mathcal{X}})$.*

3.1.2 Problème d'optimisation à résoudre

Le problème 1 peut être reformulé comme un problème d'optimisation au moyen de la proposition suivante, dont la preuve est simple.

Proposition 1. *La valeur de $\mathcal{N}(\epsilon, s_{\mathcal{X}^n}, \rho_{\mathcal{X}})$ est $\mathcal{N}^*$ si et seulement si $\mathcal{N}^*$ est l'entier minimum tel qu'il existe une partition de $s_{\mathcal{X}^n}$ en $\mathcal{N}^*$ parties et pour chaque partie, le rayon de la plus petite boule englobante (PPBE) est strictement inférieur à ϵ.*

Dans la suite, $\mathcal{P} = \{p_k : 1 \leqslant k \leqslant |\mathcal{P}|\}$ désigne une partition de $s_{\mathcal{X}^n}$ en $|\mathcal{P}|$ parties p_k et nous nommons $|\mathcal{P}|$ sa cardinalité. $\mathcal{P}(s_{\mathcal{X}^n})$ désigne l'ensemble de toutes ces partitions. Pour tout $p_k \in \mathcal{P}$, $c(p_k)$ est le centre de sa PPBE. On définit ainsi $\mathcal{C}(\mathcal{P}) = \{c(p_k) : 1 \leqslant k \leqslant |\mathcal{P}|\}$. Cette

proposition et ces notations nous autorisent à définir le problème 2, dont la résolution fournit une solution optimale au problème 1 (c'est-à-dire une solution satisfaisant $|\mathcal{C}(\epsilon)| = \mathcal{N}(\epsilon, s_{\mathcal{X}^n}, \rho)$).

Problème 2.

$$\min_{\mathcal{P} \in \mathcal{P}(s_{\mathcal{X}^n})} |\mathcal{P}|$$

$$s.c.\ \ \forall k \in [\![1; |\mathcal{P}|]\!],\ \max_{\{i: x_i \in p_k\}} \rho_{\mathcal{X}}(x_i, c(p_k)) < \epsilon.$$

Soit $\mathcal{P}^*$ une solution optimale du problème 2. D'après la proposition 1, $\mathcal{C}(\mathcal{P}^*)$ est un ϵ-réseau de $s_{\mathcal{X}^n}$ de cardinalité minimale ($|\mathcal{P}^*| = \mathcal{N}(\epsilon, s_{\mathcal{X}^n}, \rho_{\mathcal{X}})$), de sorte qu'une solution optimale du problème 1 est obtenue en fixant $\mathcal{C}(\epsilon) = \mathcal{C}(\mathcal{P}^*)$. Malheureusement, le problème 2 est un problème d'optimisation combinatoire qui, à notre connaissance, ne peut être résolu qu'en explorant la combinatoire. Le nombre de partitions de $s_{\mathcal{X}^n}$, c'est-à-dire son *nombre de Bell* B_n, étant minoré par

$$B_n \geqslant \left(\frac{n-1}{2}\right)^{\frac{n-1}{4}},$$

cette exploration n'est pas envisageable. Cependant, l'étude du problème 2 est une source d'inspiration pour dériver une méthode produisant des solutions (sous-optimales) du problème 1. En effet, il existe une famille bien connue de méthodes de classification considérant initialement autant de clusters qu'il y a de points (solution réalisable triviale du problème 2) et qui fusionne itérativement les clusters (produisant des partitions de cardinalité de plus, en plus petite) en fonction d'une mesure de dissimilarité : la famille des Classifications Ascendantes Hiérarchiques (CAH) [27]. Les CAH sont des méthodes qui construisent un arbre binaire (un dendrogramme) où initialement chaque feuille est un *cluster* (singleton). Les clusters sont fusionnés deux par deux en clusters de plus en plus grands jusqu'à avoir un unique cluster (la racine). La fusion des clusters se fait selon un *linkage* (noté d_H dans ce qui suit). Dans un dendrogramme, la hauteur d'une fusion entre deux clusters donne la valeur correspondante du linkage. Ainsi un dendrogramme présente les clusters qui sont fusionnés et l'ordre dans lequel ils sont fusionnés (voir la figure 3.1 pour une illustration).

Dans le cadre de notre étude, c'est la métrique $\rho_{\mathcal{X}}$ qui sera utilisée comme mesure de dissimilarité (entre points). Les linkages s'appuieront sur cette mesure pour comparer deux clusters. Si l'on peut être satisfait avec une solution sous-optimale du problème 2, alors ces méthodes CAH transforment le problème 2 en un problème qui peut être résolu en pratique puisque le nombre de partitions considérées est maintenant borné par n. La question qui se pose alors est la suivante : quelle est la qualité de la meilleure solution, c.-à-d., la cardinalité de la plus petite partition (la plus proche de la racine) associée à une ϵ-couverture ?

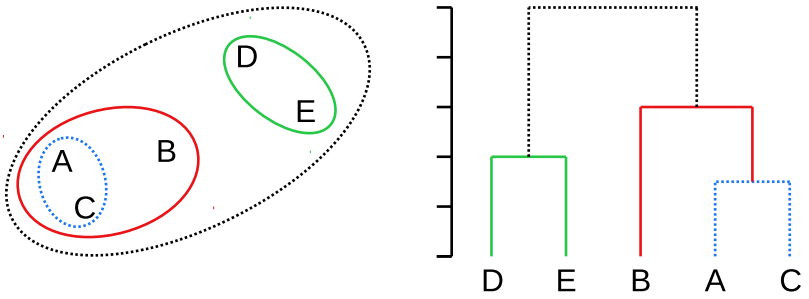


FIGURE 3.1 – Représentation d'une CAH. À gauche, les données avec en couleur les différents clusters produits. À droite, le dendrogramme associé.

La section suivante discute de l'inférence d'ϵ-réseaux à partir des méthodes de CAH.

3.2 Obtention d'ϵ-réseaux à partir d'une méthode de CAH

Dans cette section, $\mathcal{S}_1$ et $\mathcal{S}_2$ désignent deux sous-ensembles de $s_{\mathcal{X}^n}$ parmi les candidats à la fusion à une certaine itération d'une CAH. En tant que tels, ils sont des sous-ensembles particuliers de $s_{\mathcal{X}^n}$. Nous avons vu que même si les CAH ne sont pas conçues pour inférer des ϵ-réseaux, elles peuvent toutes être utilisées à cette fin.

Proposition 2. *À toute méthode de CAH de fonction de linkage d_H, on associe une méthode de calcul d'ϵ-réseaux de même nom consistant à calculer les PPBE sur les partitions qu'elle engendre.*

De plus, il existe des linkages qui présentent la propriété suivante.

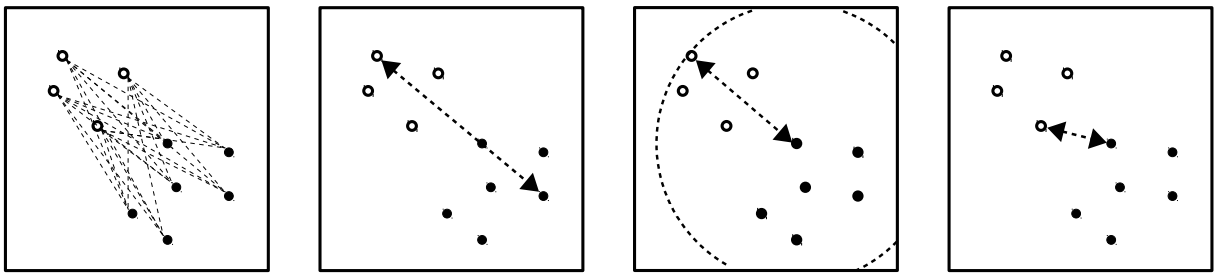
Propriété 1. *La fonction de linkage d_H croît à chaque fois que 2 clusters sont fusionnés et il existe une fonction f_H telle que, pour chaque paire $(\mathcal{S}_1, \mathcal{S}_2)$, $r\left(\mathcal{S}_1 \bigcup \mathcal{S}_2\right)$, le rayon de la plus petite boule englobant $\mathcal{S}_1 \bigcup \mathcal{S}_2$, est majoré par $f_H\left(d_H\left(\mathcal{S}_1, \mathcal{S}_2\right), |\mathcal{S}_1|, |\mathcal{S}_2|\right)$.*

La propriété 1 fournit une première partition non triviale qui donne un point de départ pour l'exploration du dendrogramme. En effet, elle permet d'obtenir un majorant de $\mathcal{N}\left(\epsilon, s_{\mathcal{X}^n}, \rho_{\mathcal{X}}\right)$, sans avoir à calculer de boule. Ainsi, le nombre total de boules à calculer est diminué du nombre de fusions.

Les propriétés intéressantes des fonctions f_H nous poussent à étudier les principaux linkages de la littérature (décrits par Bien et Tibshirani dans [6]). Ces linkages sont donnés dans le tableau 3.1, et nous déterminons ensuite les fonctions f_H correspondantes, telles que définies par la propriété 1.

TABLE 3.1 – Principaux linkages des CAH.

Méthode	Linkage d_H
Average	$d_A(\mathcal{S}_1, \mathcal{S}_2) = \frac{1}{\lvert\mathcal{S}_1\rvert\lvert\mathcal{S}_2\rvert} \sum_{(x,x') \in \mathcal{S}_1 \times \mathcal{S}_2} \rho_{\mathcal{X}}(x, x')$
Centroid	$d_{Ce}(\mathcal{S}_1, \mathcal{S}_2) = \rho_{\mathcal{X}}\left(\frac{1}{\lvert\mathcal{S}_1\rvert} \sum_{x \in \mathcal{S}_1} x, \frac{1}{\lvert\mathcal{S}_2\rvert} \sum_{x' \in \mathcal{S}_2} x'\right)$
Complete	$d_{Co}(\mathcal{S}_1, \mathcal{S}_2) = \max_{(x,x') \in \mathcal{S}_1 \times \mathcal{S}_2} \rho_{\mathcal{X}}(x, x')$
Minimax	$d_{mM}(\mathcal{S}_1, \mathcal{S}_2) = \min_{x \in \mathcal{S}_1 \bigcup \mathcal{S}_2} \max_{x' \in \mathcal{S}_1 \bigcup \mathcal{S}_2} \rho_{\mathcal{X}}(x, x')$
Single	$d_S(\mathcal{S}_1, \mathcal{S}_2) = \min_{(x,x') \in \mathcal{S}_1 \times \mathcal{S}_2} \rho_{\mathcal{X}}(x, x')$



Average Complete Minimax Single

FIGURE 3.2 – Application des principaux linkages de la littérature sur deux groupes de points.

Nous notons d'abord qu'une seule fonction d_H ne satisfait pas la condition de monotonicité de la propriété 1 : la fonction d_{Ce} (voir [46]). Nous limitons donc l'étude aux quatre autres linkages, dont des applications schématiques sont présentées en figure 3.2. Dans la suite, pour chaque sous-ensemble $\mathcal{S}'$ de $s_{\mathcal{X}^n}$, diam $(\mathcal{S}')$ désigne son diamètre (par rapport à la métrique $\rho_{\mathcal{X}}$). La fonction f_H la plus simple à exhiber est obtenue pour la liaison Minimax. En effet, puisque $\mathcal{S}_1 \bigcup \mathcal{S}_2 \subset \mathcal{X}$, c'est-à-dire que le rayon de la plus petite boule englobante telle que le centre est un point de $\mathcal{X}$ est inférieur au rayon de la plus petite boule englobante telle que le centre est un point de $(\mathcal{S}_1 \bigcup \mathcal{S}_2)$,

$$\begin{aligned} \forall \left(\mathcal{S}_1, \mathcal{S}_2\right),\ r\left(\mathcal{S}_1 \bigcup \mathcal{S}_2\right) &= \min_{x \in \mathcal{X}} \max_{x' \in \mathcal{S}_1 \bigcup \mathcal{S}_2} \rho_{\mathcal{X}}\left(x, x'\right) \\ &\leqslant d_{mM}\left(\mathcal{S}_1, \mathcal{S}_2\right). \end{aligned}$$

Ainsi nous pouvons écrire :

$$\forall \left(\mathcal{S}_1, \mathcal{S}_2\right),\ f_{mM}\left(d_{mM}\left(\mathcal{S}_1, \mathcal{S}_2\right), \left|\mathcal{S}_1\right|, \left|\mathcal{S}_2\right|\right) = d_{mM}\left(\mathcal{S}_1, \mathcal{S}_2\right).$$

Nous passons maintenant au linkage Complete.

$$\begin{aligned} \forall \left(\mathcal{S}_1, \mathcal{S}_2\right),\ r\left(\mathcal{S}_1 \bigcup \mathcal{S}_2\right) &\leqslant \operatorname{diam}\left(\mathcal{S}_1 \bigcup \mathcal{S}_2\right) \\ &= \max_{\{x, x'\} \in \mathcal{S}_1 \bigcup \mathcal{S}_2} \rho_{\mathcal{X}}\left(x, x'\right) \\ &= \max_{(x, x') \in \mathcal{S}_1 \times \mathcal{S}_2} \rho_{\mathcal{X}}\left(x, x'\right) \qquad (3.1) \\ &= d_{Co}\left(\mathcal{S}_1, \mathcal{S}_2\right). \end{aligned}$$

La transition vers (3.1) (qui serait fausse si $\mathcal{S}_1$ et $\mathcal{S}_2$ étaient des sous-ensembles arbitraires de $s_{\mathcal{X}^n}$), découle de la monotonicité de d_{Co}. En effet, par construction des clusters en utilisant le linkage Complete, le diamètre de $(\mathcal{S}_1 \bigcup \mathcal{S}_2)$ correspond obligatoirement à la distance entre un point de $\mathcal{S}_1$ et un point de $\mathcal{S}_2$. Par conséquent, nous pouvons définir

$$\forall \left(\mathcal{S}_1, \mathcal{S}_2\right),\ f_{Co}\left(d_{Co}\left(\mathcal{S}_1, \mathcal{S}_2\right), \left|\mathcal{S}_1\right|, \left|\mathcal{S}_2\right|\right) = d_{Co}\left(\mathcal{S}_1, \mathcal{S}_2\right).$$

De la même manière pour le linkage Average, puisque la somme correspondante peut être minorée par le maximum, en appliquant exactement le même raisonnement que pour Complete :

$$\forall \left(\mathcal{S}_1, \mathcal{S}_2\right),\ f_A\left(d_A\left(\mathcal{S}_1, \mathcal{S}_2\right), \left|\mathcal{S}_1\right|, \left|\mathcal{S}_2\right|\right) = \left|\mathcal{S}_1\right| \left|\mathcal{S}_2\right| d_A\left(\mathcal{S}_1, \mathcal{S}_2\right).$$

Méthode	Fonction $f_H\left(d_H\left(\mathcal{S}_1,\mathcal{S}_2\right),\left\|\mathcal{S}_1\right\|,\left\|\mathcal{S}_2\right\|\right)$
Average	$f_A = \theta\left\|\mathcal{S}_1\right\|\left\|\mathcal{S}_2\right\| d_A\left(\mathcal{S}_1,\mathcal{S}_2\right)$
Complete	$f_{Co} = \theta d_{Co}\left(\mathcal{S}_1,\mathcal{S}_2\right)$
Minimax	$f_{mM} = d_{mM}\left(\mathcal{S}_1,\mathcal{S}_2\right).$
Single	$f_S = \theta\left(\left\|\mathcal{S}_1 \bigcup \mathcal{S}_2\right\| - 1\right) d_S\left(\mathcal{S}_1,\mathcal{S}_2\right)$

Espace considéré		Valeur de θ
Espace Métrique		1
Espace de Hilbert	Dimension finie	$\sqrt{\frac{d}{2(d+1)}}$
	Dimension infinie	$\frac{1}{\sqrt{2}}$

TABLE 3.2 – Fonctions f_H définies par rapport à la structure de $\mathcal{X}$. θ varie en fonction de cette structure.

En ce qui concerne le linkage Single, l'inégalité triangulaire nous fournit

$$\forall\left(\mathcal{S}_1,\mathcal{S}_2\right),\ \operatorname{diam}\left(\mathcal{S}_1 \bigcup \mathcal{S}_2\right) \leqslant \operatorname{diam}\left(\mathcal{S}_1\right) + d_S\left(\mathcal{S}_1 \bigcup \mathcal{S}_2\right) + \operatorname{diam}\left(\mathcal{S}_2\right).$$

Par conséquent, on peut prouver par récurrence que

$$\operatorname{diam}\left(\mathcal{S}_1 \bigcup \mathcal{S}_2\right) \leqslant \left(\left|\mathcal{S}_1 \bigcup \mathcal{S}_2\right| - 1\right) d_S\left(\mathcal{S}_1,\mathcal{S}_2\right).$$

La fonction f_S est ainsi définie comme :

$$\forall\left(\mathcal{S}_1,\mathcal{S}_2\right),\ f_S\left(d_S\left(\mathcal{S}_1,\mathcal{S}_2\right),\left|\mathcal{S}_1\right|,\left|\mathcal{S}_2\right|\right) = \left(\left|\mathcal{S}_1 \bigcup \mathcal{S}_2\right| - 1\right) d_S\left(\mathcal{S}_1,\mathcal{S}_2\right).$$

3.2.1 Cas des espaces de Hilbert

Dans le cas particulier où l'espace métrique utilisé est un espace hilbertien, une relation fondamentale entre le diamètre et le rayon peut être utilisée.

Théorème 1 (D'après le théorème de Jung [43] et le théorème 1 de [64])**.** *Soit H un espace de Hilbert et d sa dimension. Soit $\overline{H}$ un sous-ensemble de H de diamètre $diam\left(\overline{H}\right)$. Alors le rayon $r\left(\overline{H}\right)$ de la plus petite boule englobant $\left(\overline{H}\right)$ vérifie : si $d < +\infty$,*

$$r\left(\mathcal{E}'\right) \leqslant \left(\sqrt{\frac{d}{2\left(d+1\right)}}\right) diam\left(\mathcal{E}'\right),$$

sinon,

$$r\left(\mathcal{E}'\right) \leqslant \left(\frac{1}{\sqrt{2}}\right) diam\left(\mathcal{E}'\right).$$

Cette relation peut être utilisée pour majorer de manière plus fine le rayon, et donc obtenir des fonctions f_H de meilleure qualité. Le tableau 3.2 reprend les différentes fonctions f_H (pour les linkages d_{Co}, d_A et d_S) en fonction de la structure de $\mathcal{X}$. Rappelons que tout espace de Hilbert de dimension finie est isomorphe à l'espace euclidien de même dimension.

3.2.2 Application des fonctions f_H

La figure 3.3a illustre les partitions résultant d'un arrêt au niveau retourné par f_H pour les quatre fonctions de linkage (d_{mM}, d_{Co}, d_A et d_S) sur un ensemble de 8 points de la droite réelle, si on exploite la dimension ($d = 1$). La solution de cardinalité minimale est obtenue avec les linkages Minimax, Complete et Single ($|\mathcal{P}| = 3$). Cette solution est optimale puisque 3 est la valeur du nombre de couverture. En effet, la distance maximale entre les points étant de $13,2$, et avec $\epsilon = 2,5$, il faut bien au moins 3 cercles pour couvrir tous les points.

La figure 3.3b représente quant à elle les résultats obtenus en poursuivant l'exploration du dendrogramme (vers la racine). Les ϵ-couvertures ainsi obtenues sont à présent toutes de cardinalité minimale. De plus, une seule de ces couvertures est différente des autres, celle du linkage Complete.

Il convient de remarquer que les ϵ-réseaux produits par le linkage Minimax sont des ϵ-réseaux propres. Cette propriété n'est pas requise dans le cadre théorique que nous avons défini. Inversement, elle pourrait avoir un effet négatif sur la cardinalité des solutions (en gardant en mémoire que $\mathcal{N}\left(\epsilon, \mathcal{E}', \rho\right) \leqslant \mathcal{N}^{\text{int}}\left(\epsilon, \mathcal{E}', \rho\right)$). Cette observation est à l'origine du principe de notre méthode qui est maintenant introduite.

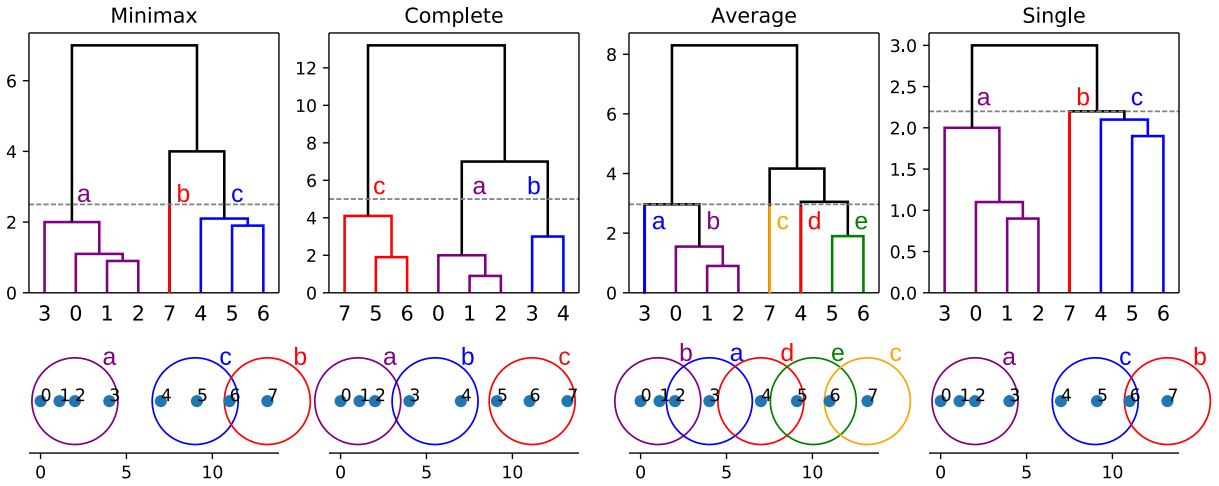
3.3 Méthode de CAH Radius

3.3.1 Définition

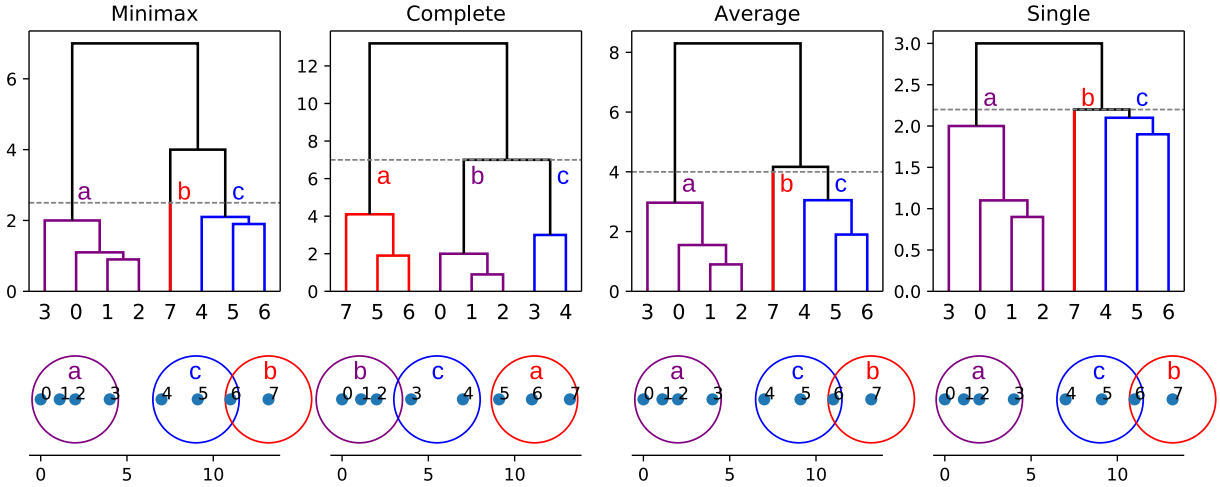
Notre méthode pour l'inférence empirique des ϵ-réseaux est essentiellement une CAH associée à la fonction de linkage donnée par la définition 2.

Définition 2. *Le linkage Radius (d_R) est donné par :*

$$\forall\left(\mathcal{S}_1, \mathcal{S}_2\right),\ d_R\left(\mathcal{S}_1, \mathcal{S}_2\right) = r\left(\mathcal{S}_1 \bigcup \mathcal{S}_2\right). \tag{3.2}$$



(a) ϵ-réseaux obtenus pour la partition associée à la hauteur fournie par la fonction f_H



(b) ϵ-réseaux obtenus pour la partition associée à la hauteur maximale fournie par exploration des dendrogrammes

FIGURE 3.3 – Dendrogrammes produits avec quatre linkages : d_{mM}, d_{Co}, d_A et d_S. Les coupes sont représentées par les lignes pointillées. Les ϵ-couvertures ($\epsilon = 2, 5$) résultantes sont données sous les dendrogrammes.

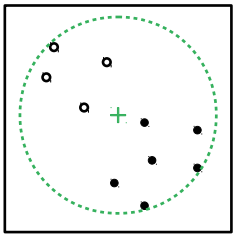


FIGURE 3.4 – Représentation du linkage Radius.

La figure 3.4 illustre une application du linkage Radius sur deux clusters.

Nous étudions la méthode Radius successivement comme méthode de CAH puis comme méthode d'inférence d'ϵ-réseaux.

Propriétés statistiques de la méthode de CAH

La méthode possède toutes les propriétés connues de la méthode Minimax (propriétés établies par Bien et Tibshirani).

Proposition 3. *La méthode Radius possède les cinq propriétés suivantes :*

- *elle ne produit pas d'inversion ;*
- *elle est admissible à la bonne structuration des k-groupes (well-structured k-group admissible) ;*
- *elle accepte les transformations monotones des distances (ces transformations ne changent pas la classification) ;*
- *elle accepte la duplication des points (cela ne change pas les clusters formés) ;*
- *elle satisfait la propriété de réductibilité.*

Les preuves étant simples, nous n'en donnons qu'une, l'admissibilité à la bonne structuration des k-groupes.

Définition 3. *Lorsqu'il existe une partition $\mathcal{P}$ en k clusters, dans laquelle toutes les distances intra-clusters sont plus petites que toutes les distances inter-clusters, alors une telle partition est obtenue après $n-k$ fusions.*

Démonstration. La preuve est adaptée de celle réalisée par Bien et Tibshirani. Supposons qu'il existe une partition des données $\mathcal{P}$ en k parties, telle que $\rho_{\mathcal{X}}(x, x') \leqslant a$ si $x, x' \in p_i$ et $\rho_{\mathcal{X}}(x, x') > a$ si $x \in p_i, x' \in p_j$ avec $i \neq j$. Il s'ensuit que pour tout $x \in p_i, r(\{x\} \bigcup p_i) \leqslant a$ et $r(\{x\} \bigcup p_j) > a$. Maintenant, si $\mathcal{S}_1, \mathcal{S}_2 \subset p_i$, alors :

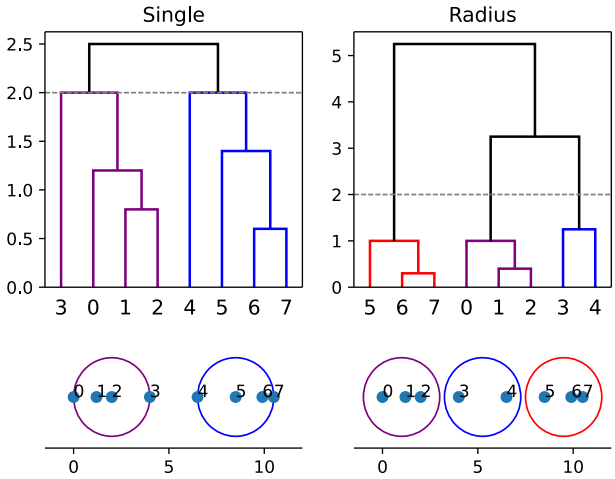


FIGURE 3.5 – Dendrogrammes produits par Single et Radius, sur un ensemble de 8 points de la droite des réels. Sous les dendrogrammes sont présentées des ϵ-couvertures associées.

$$d_R(\mathcal{S}_1, \mathcal{S}_2) \leqslant r(p_i) \leqslant a.$$

De plus, si $\mathcal{S}_1 \subset p_i$ et que $\mathcal{S}_2 \cap p_i = \emptyset$, alors :

$$d_R(\mathcal{S}_1, \mathcal{S}_2) = r\left(\mathcal{S}_1 \bigcup \mathcal{S}_2\right) \geqslant r\left(\{\, x, x'\} : (x, x') \in p_i \times p_j\right) > a.$$

Ainsi, le linkage Radius fusionnera toujours les clusters au sein de l'ensemble p_i avant de fusionner un cluster $\mathcal{S}_1 \subset p_i$ avec un sous-ensemble non contenu dans p_i. Ceci établit qu'à une certaine itération de l'algorithme, p_i est formé.

Or $d_R(p_i, p_j) > a$ et $r(p_i) \leqslant a$, donc couper à la hauteur a donne précisément le clustering $\mathcal{P}$. Comme il s'agit d'une solution de k cluster, cet état est atteint après $n - k$ fusions. □

Propriétés de la méthode d'inférence d'ϵ-réseaux

La méthode Radius étant particulièrement adapté pour inférer des ϵ-réseaux, il convient de souligner qu'il n'est cependant pas uniformément le meilleur. En effet, dans l'exemple donné en figure 3.5, le linkage Single donne un meilleur partitionnement pour un $\epsilon = 2$. Ce contre-exemple est très particulier. Nous n'avons pas trouvé de contre-exemple pour les linkages Complete et Minimax.

3.3.2 Calcul des plus petites boules englobantes

Dans cette partie nous discutons le calcul des PPBE en fonction du type d'espace.

Espace de Banach

Dans un Banach la réflexivité joue un rôle essentiel sur le calcul des PPBE.

Théorème 2 (D'après le théorème de [85] et le théorème 2.4 de [61])**.** *Soit* $(\mathcal{X}, \|\cdot\|)$ *un espace de Banach, si cet espace n'est pas réflexif :*

Pour chaque $\epsilon > 0$*,* $\mathcal{X}$ *admets une norme équivalente* $|\cdot|_\epsilon$ *telle que* $(1-\epsilon)\|\cdot\| \leqslant |\cdot|_\epsilon \leqslant \|\cdot\|$ *et trois points d'un sous-ensemble de* $\mathcal{X}$ *ne possèdent pas de centre de Chebyshev dans* $(\mathcal{X}, |\cdot|_\epsilon)$*.*

Si cet espace est réflexif :

la contrainte sur l'ensemble est faiblement fermée, alors le problème de la plus petite boule englobante admet une solution.

Espace de Hilbert

Dans le cas où l'espace est un espace de Hilbert, l'obtention de la PPBE est un problème de programmation convexe.

Le problème d'optimisation prend la forme suivante :

$$\min_{O} \max_{x_i} \rho^2_{\mathcal{X}}(O, x_i)$$

Le Lemme 3.6 de [63] établit la convexité de la fonction objectif.

Espace de Hilbert à noyau reproduisant

Dans le cas où l'espace de description est un espace Hilbertien à noyau reproduisant (*Reproducing Kernel Hilbert Space* : RKHS [4]), le calcul des boules s'effectue de la manière suivante.

Soit $s_{\mathcal{X}^m}$ un sous-ensemble de $s_{\mathcal{X}^n}$ de cardinalité m. Sans perte de généralité, ses points sont réindexés de telle sorte que l'on peut écrire $s_{\mathcal{X}^m} = \{x_i : 1 \leqslant i \leqslant m\}$. $K_m \in \mathcal{M}_{m,m}(\mathbb{R})$ est la matrice de Gram correspondante et $\boldsymbol{\kappa}_m \in \mathbb{R}^m_+$ est sa première diagonale. Le problème de programmation quadratique est alors le suivant :

Problème 3.

$$\min_{O, R_2 \in \mathbb{R}_+} R_2$$

$$s.t.\ \forall i \in [\![1; m]\!],\ \rho^2_{\mathcal{X}}(O, x_i) \leqslant R_2.$$

Le problème 3 peut être difficile à résoudre directement, surtout si l'espace est de dimension infinie. Cependant, la manière classique d'inférer les valeurs des paramètres d'une méthode à noyau, l'application de la dualité lagrangienne, est disponible.

Le lagrangien prend la forme suivante, avec les multiplicateurs de Lagrange notés α_i :

$$\mathcal{L}(O, R, \alpha) = R_2 + \sum_{i=1}^{m} \alpha_i \left(\|O - x_i\|^2 - R_2 \right). \tag{3.3}$$

Le calcul des gradients du lagrangien nous donne :

$$\nabla_{R_2} \mathcal{L} = 1 - \sum_{i=1}^{m} \alpha_i$$

$$\nabla_O \mathcal{L} = 2 \sum_{i=1}^{m} \alpha_i O - 2 \sum_{i=1}^{m} \alpha_i x_i.$$

En conséquence, à l'optimum (c.-à-d., au point-selle) :

$$\begin{cases} {R_2}^* \Rightarrow \sum_{i=1}^{m} \alpha_i^* = 1 \\ O^* = \frac{1}{\sum_{i=1}^{m} \alpha_i} \sum_{i=1}^{m} \alpha_i x_i \Rightarrow O^* = \sum_{i=1}^{m} \alpha_i x_i \end{cases} \tag{3.4}$$

Par substitution dans le lagrangien :

$$\begin{aligned}
\mathcal{L}(O^*, {R_2}^*, \alpha^*) = & \sum_{i=1}^{m} \alpha_i^* \left(\|O^* - x_i\|^2 \right) && (3.5) \\
= & \sum_{i=1}^{m} \alpha_i^* \|O^*\|^2 + \sum_{i=1}^{m} \alpha_i^* \|x_i\|^2 - 2 \sum_{i=1}^{m} \alpha_i^* \langle O^*, x_i \rangle && (3.6) \\
= & \langle O^*, O^* \rangle \sum_{i=1}^{m} \alpha_i^* + \sum_{i=1}^{m} \alpha_i^* \|x_i\|^2 - 2 \sum_{i=1}^{m} \alpha_i^* \langle O^*, x_i \rangle && (3.7) \\
= & \left\langle \sum_{i=1}^{m} \alpha_i^* x_i, \sum_{j=1}^{m} \alpha_j^* x_j \right\rangle + \sum_{i=1}^{m} \alpha_i^* \|x_i\|^2 - 2 \sum_{i=1}^{m} \alpha_i^* \left\langle \sum_{j=1}^{m} \alpha_j^* x_j, x_i \right\rangle && (3.8) \\
= & -\sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i^* \alpha_j^* \langle x_i, x_j \rangle + \sum_{i=1}^{m} \alpha_i^* \langle x_i, x_i \rangle && (3.9)
\end{aligned}$$

Le problème dual du problème 3 donc la forme suivante (puisque l'on minimise le Lagrangien) :

Problème 4.

$$\min_{\alpha \in \mathbb{R}^m} \left\{ \alpha^T K_m \alpha - \boldsymbol{\kappa}_m^T \alpha \right\}$$

$$s.t. \begin{cases} \forall i \in [\![1; m]\!],\ \alpha_i \geqslant 0 \\ \mathbf{1}_m^T \alpha = 1 \end{cases}.$$

En notant (O^*, R_2^*) la solution du problème primaire et α^* la solution du problème secondaire

la solution de de son dual, nous déduisons des conditions de Kuhn-Tucker (KT) l'expression analytique du prototype de $s_{\mathcal{X}^m}$:

$$O^* = \sum_{i=1}^{m} \alpha_i^* x_i. \tag{3.10}$$

Précisément, l'équation (3.10) fournit la localisation du centre de la PPBE de $s_{\mathcal{X}^m}$ dans sa coque convexe. Avec cette formule en main, la distance carrée entre un prototype et tout point dans $\mathbf{H}$ (qu'il soit dans $s_{\mathcal{X}^n}$ ou non) est

$$\begin{aligned}\rho^2_{\mathcal{X}}(O^*, x) = &\ \|O^*\|^2 + \|x\|^2 - 2\langle O^*, x\rangle \\ = &\ \sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i^*\alpha_j^*\langle x_i, x_j\rangle + \langle x, x\rangle - 2\sum_{i=1}^{m}\alpha_i^*\langle x_i, x\rangle.\end{aligned}$$

Quant à la valeur du rayon $\sqrt{R_2^*}$, puisque $\mathbf{1}_m^T\alpha^* = 1$, les conditions complémentaires KT nous fournissent :

$$R_2^* = \sum_{i=1}^{m} \alpha_i^* \rho^2_{\mathcal{X}}(O^*, x_i).$$

3.3.3 Implémentation algorithmique

Notre problème est d'une forme particulière, il s'agit de la minimisation d'une forme quadratique sur le simplex standard/probabilité. Cette propriété peut être exploitée par une méthode classique de programmation quadratique, l'algorithme de Frank-Wolfe [23], qui se trouve être parmi les meilleurs "algorithmes d'approximation" de l'état de l'art pour le calcul des PPBE et nous montrons la pertinence de ce choix. Cet algorithme décompose le problème quadratique en une série de problèmes linéaires. En notant notre fonction objectif $\mathcal{D} = \alpha^T K_m \alpha - \boldsymbol{\kappa}_m^T \alpha$, nous avons donc $\nabla\mathcal{D} = 2K_m\alpha - \boldsymbol{\kappa}_m$. Le problème linéaire prend la forme suivante :

$$\min_v \nabla\mathcal{D} v$$

$$\text{s.t.} \quad \begin{cases} \forall i \in [\![1; m]\!],\ v_i \geqslant 0 \\ \mathbf{1}_m^T v = 1 \end{cases}$$

Les contraintes de positivité ($v_i \geqslant 0$) et la somme égale à 1 ($\mathbf{1}_m^T v = 1$) impliquent qu'une solution réalisable doit résider dans le simplex standard. Il est connu que ce problème linéaire possède une solution analytique. Cette solution est sur le sommet du simplexe qui minimise le gradient. Elle est nommé v^*.

En outre, comme notre problème est un problème quadratique, il existe une expression analytique pour le pas de gradient. Ainsi la mise à jour de α à l'itération $t+1$ se fait avec le pas de gradient nomme θ suivant :

$$\theta = \frac{1}{2} \frac{\left(\nabla \mathcal{D}^T(\alpha - v^*)\right)}{(\alpha - v^*)^T K (\alpha - v^*)}.$$

Ce qui donne :

$$\alpha(t+1) = (1-\theta)\alpha(t) + \theta v^*$$

Cette expression analytique permet de réaliser la descente de gradient de manière à toujours rester dans le simplexe, et donc à garder des solutions réalisables.

Notre implémentation de la méthode fait usage de ces deux caractéristiques. En pratique, pour une valeur de l'écart de dualité égale à 0.99, la complexité temporelle observée est de $O\left(m^2\right)$.

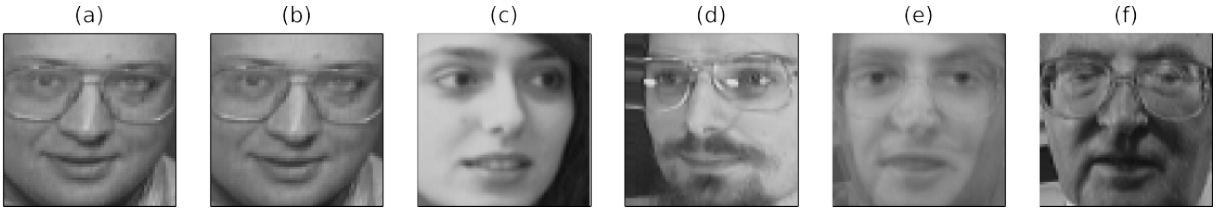
3.4 Résultats expérimentaux

Notre méthode est évaluée dans le cadre d'une étude comparative. Le critère d'évaluation est la cardinalité des ϵ-réseaux générés. Les méthodes de référence sont les quatre linkages de la section 3.2 qui possèdent la Propriété 1. Ainsi, l'étude compare l'efficacité (pour produire de petits ϵ-réseaux) des fonctions de linkages suivantes : d_R (donnée par l'équation 3.2), d_A, d_{Co}, d_{mM} et d_S (rassemblés dans le Tableau 3.1). Notre méthode est implémentée comme décrite dans la section 3.3. En ce qui concerne les algorithmes de la littérature, les ϵ-réseaux sont obtenus soit en s'arrêtant aux points de départ définis par les fonctions f_H correspondantes, mais aussi en explorant les dendrogrammes.

L'ensemble de test choisi est l'un de ceux utilisés par Bien et Tibshirani dans [6] : the Olivetti faces dataset [1]. Il est composé des images des visages de 40 individus différents. Pour chaque individu, 10 images sont fournies. Il s'agit d'images en niveaux de gris (8 bits) de taille 64×64. Les données sont donc des vecteurs de $\mathbb{R}^{64 \cdot 64}$. Cet espace est doté de sa structure canonique d'espace euclidien. Pour résumer, les algorithmes opèrent sur la matrice de Gram de $s_{\mathcal{X}^n} = \{x_i : 1 \leqslant i \leqslant n\}$ où $n = 400$. Il est possible de déduire de cette matrice des éléments notables : les deux images les plus proches, les deux images les plus éloignées, le centre de la PPBE, et les images sur sa surface. Ces éléments sont représentés dans la Fig 3.6.

1. https://cs.nyu.edu/∼roweis/data.html

FIGURE 3.6 – Jeu de données Olivetti Faces. (a) et (b) : les deux images les plus proches ; (c) et (d) : les deux images les plus éloignées ; (e) : centre de la PPBE ; (f) : une image sur la surface de la PPBE.



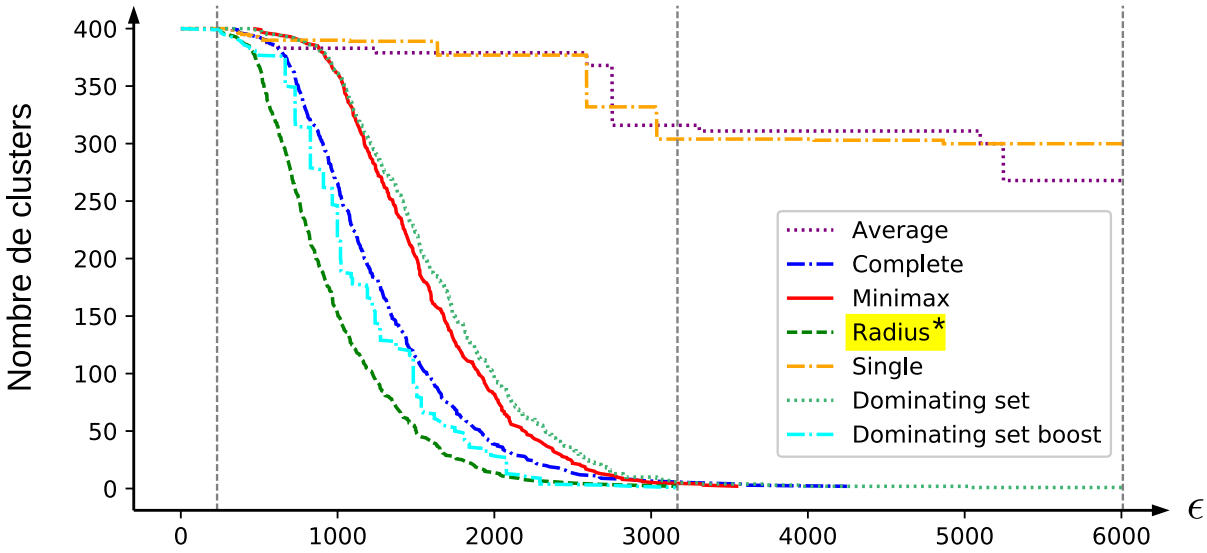
Soit $d_{2,\min} = \min_{1 \leqslant i < j \leqslant n} d_2(x_i, x_j)$. Alors, $d_{2,\min}$, $r\left(s_{\mathcal{X}^n}\right)$, et $\operatorname{diam}\left(s_{\mathcal{X}^n}\right)$ peuvent être utilisés pour restreindre l'étude aux valeurs d'ϵ. pour lesquelles la dérivation d'un ϵ-réseau est non triviale. En effet, si $\epsilon < \frac{1}{2} d_{2,\min}$, alors $\mathcal{N}\left(\epsilon, s_{\mathcal{X}^n}, d_2\right) = n$ et nous pouvons définir $\mathcal{C}\left(\epsilon\right) = s_{\mathcal{X}^n}$. Inversement, si $\epsilon > r\left(s_{\mathcal{X}^n}\right)$, alors $\mathcal{N}\left(\epsilon, s_{\mathcal{X}^n}, d_2\right) = 1$. et nous pouvons définir $\mathcal{C}\left(\epsilon\right) = c\left(s_{\mathcal{X}^n}\right)$ (centre de la PPBE de $s_{\mathcal{X}^n}$). Enfin, si $\epsilon > \operatorname{diam}\left(s_{\mathcal{X}^n}\right)$, alors $\mathcal{N}^{int}\left(\epsilon, s_{\mathcal{X}^n}, d_2\right) = 1$ et nous pouvons définir $\mathcal{C}\left(\epsilon\right)$ égal à tout singleton.

Par conséquent, les cardinalités des ϵ-réseaux produits par les cinq algorithmes sont calculés pour ϵ dans l'intervalle $(0, \operatorname{diam}\left(s_{\mathcal{X}^n}\right)]$. Les courbes correspondantes sont rassemblées dans la Figure 3.7.

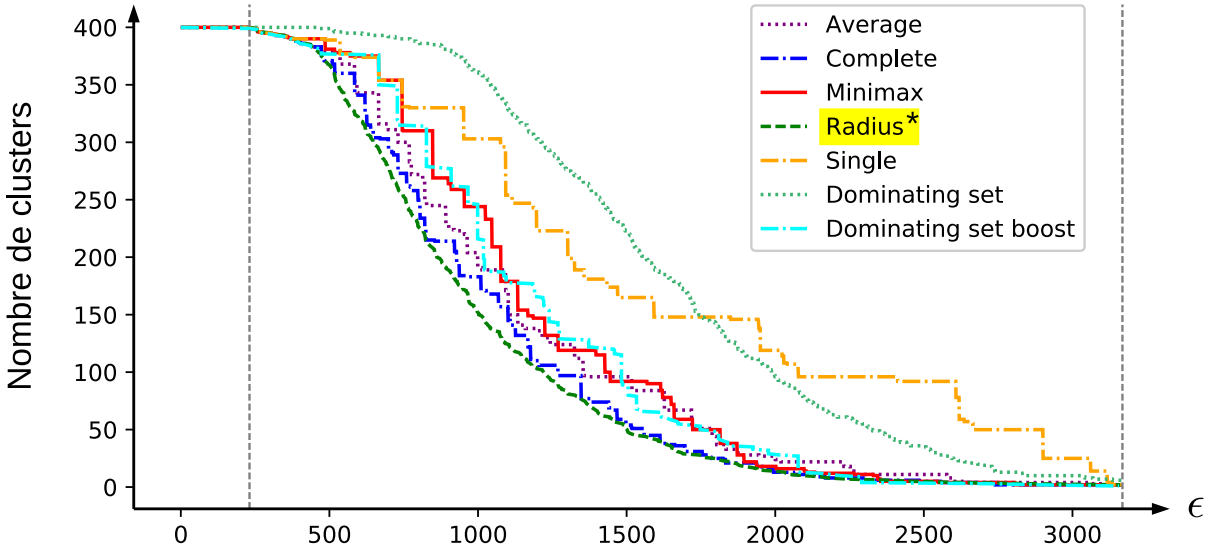
La nouvelle méthode est uniformément (pour toutes les valeurs de ϵ) supérieure aux quatre autres. En se concentrant sur les algorithmes CAH de la littérature, il est intéressant de noter que contrairement à notre intuition, le linkage Complete semble être plus efficace que le linkage Minimax. Comme prévu, ni d_A ni d_S n'apparaissent comme un choix approprié, s'il l'on considère le point de départ. Cependant, l'exploration des dendrogrammes donne des résultats surprenamment bons pour Average.

3.5 Applications aux bibliothèques de fragments

Le principal problème lié à la création des bibliothèques de fragments est la mesure utilisée. En effet, la RMSD après superposition (cRMSD) n'est pas une distance mathématique. Nous avons montré cela en calculant la matrice de cRMSD deux à deux sur l'ensemble des tri-nucléotides. Nous avons ensuite calculé le spectre de cette matrice, et une partie des valeurs du spectre sont négatives. Cette matrice des RMSD après superposition deux à deux n'est donc pas utilisable pour les méthodes classiques de clustering, ni même pour notre méthode telle quelle.



(a) Points de départ définis par les fonctions f_H.



(b) Meilleures solutions possibles en explorant les dendrogrammes.

FIGURE 3.7 – Cardinalité des ϵ-réseaux produits par les algorithmes CAH en fonction de ϵ

3.5.1 Modification de la méthode pour l'application aux bibliothèques de fragments

Pour pallier ce problème, nous avons choisi d'utiliser le RMSD après superposition de toutes les structures sur une seule. La matrice de RMSD obtenue est une matrice de distance, mais qui donne un majorant de la cRMSD réelle. La méthode de clustering utilisée précédemment pour la création des bibliothèques faisait la même chose, mais avec une unique superposition. Pour limiter l'approximation, notre méthode applique de nombreuses superpositions sur des structures différentes.

Approximation de la cRMSD par superposition sur une référence

Pour essayer de comprendre à quel point l'approximation de la cRMSD par la superposition sur une référence est juste, nous avons comparé les deux sur un jeu de donnée de 17500 tri-nucléotides. Pour ces tri-nucléotides nous avons réalisé deux matrices de RMSD, une qui est la cRMSD, et une qui est la RMSD après superposition sur la première structure du jeu de données. Les résultats sont présentés sous forme de boîtes à moustache en figure 3.8.

Nous avons regroupé les cRMSD en classes, puis regardé la distribution des RMSD pour chacune des classes. La première chose intéressante est la distance maximum observée. Il semble que pour certaines structures la cRMSD peut être de moins de 1 Å, mais après superposition sur une référence la RMSD entre ces structures est de presque 10 Å. Il semble que l'approximation que représente la superposition sur une référence soit une approximation très grossière quand la cRMSD entre deux structures dépasse 3 Å. Cela renforce la nécessité d'une méthode qui approxime au mieux la cRMSD.

Changements sur notre méthode d'inférence des prototypes

Les changements appliqués à notre méthode sont les suivants :

— lorsqu'un calcul de boule est à effectuer, tous les fragments impliqués sont superposés sur une structure de référence ;

— pour savoir quelles boules sont vraiment utiles à calculer, des bornes sont utilisées (pour limiter les superpositions).

La structure de référence sur laquelle les fragments sont superposés est la structure moyenne des deux centres de boule que l'on veut fusionner.

Les critères de calcul des boules sont basés sur la distance entre les centres des boules. Par construction, la distance entre deux centres est plus grande que chacun des rayons de leurs

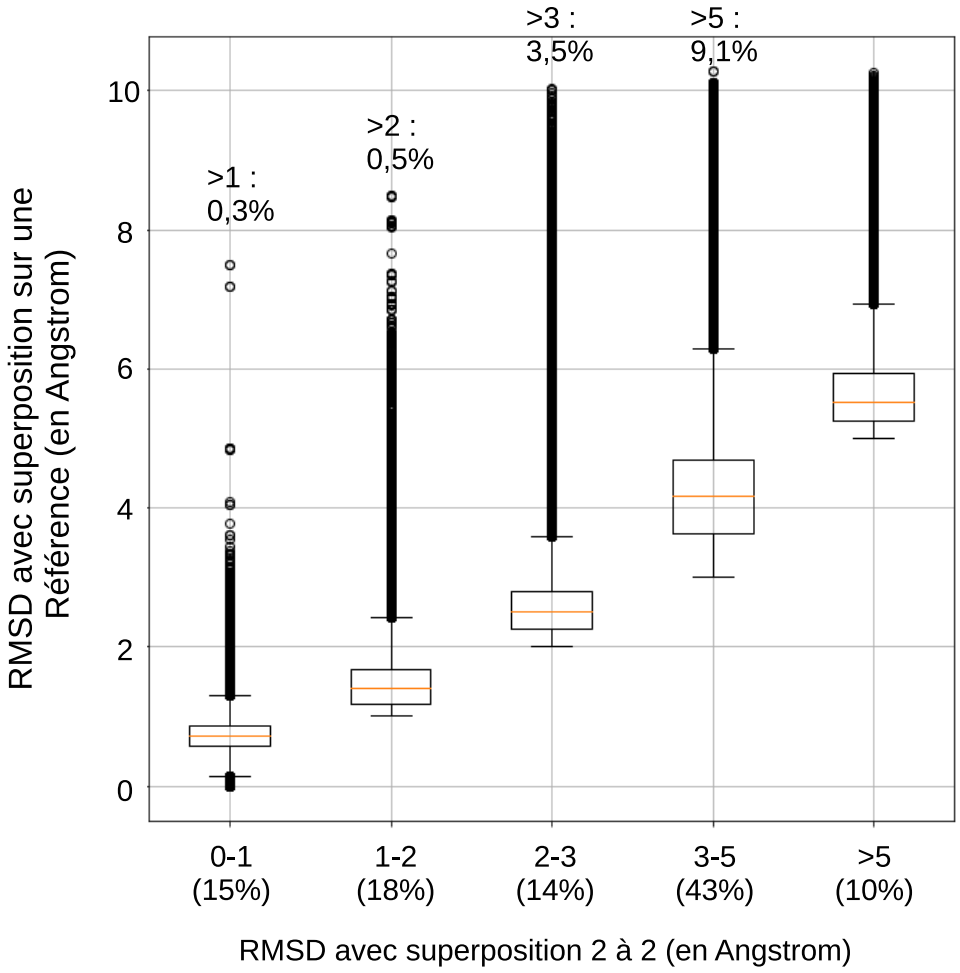


FIGURE 3.8 – Boîtes à moustaches représentant les comparaisons entre cRMSD et RMSD après superposition sur une référence. Les pourcentages sont ceux sur le nombre de comparaisons dans la classe comparé à l'ensemble des comparaisons (153 116 250).

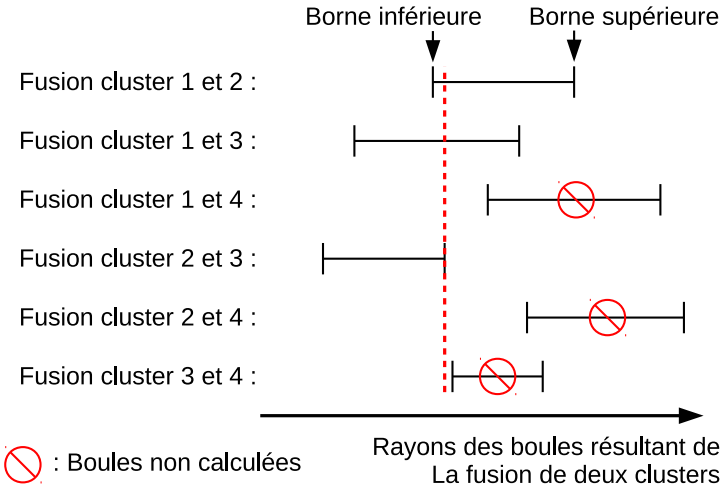


FIGURE 3.9 – Dans cette représentation schématique, nous avons 4 clusters. Nous représentons les différentes longueurs possibles des rayons des plus petites boules englobants ces clusters. Pour chaque fusion nous avons une borne supérieure et une borne inférieure. La plus petite borne supérieure nous permet de savoir quelles fusions sont à envisager. Dans cet exemple il faudra calculer les PPBE des fusions entre le cluster 1 et 2, 2 et 3, et 1 et 3.

	RRR	RRY	RYR	RYY	YRR	YRY	YYR	YYY
Ancienne méthode	4771	3372	3115	2533	3467	2412	2808	3037
Nouvelle méthode	1012	716	647	661	782	566	703	839

TABLE 3.3 – Nombre de représentants à 1 Åde résolution pour les différentes séquences.

boules. Le rayon résultant de la fusion sera au minimum la distance entre les deux centres (borne inférieure). De même, le rayon maximum sera la distance entre les deux centres plus les deux rayons (borne supérieure).

La boule donnant la plus petite valeur de la borne supérieure contraint le calcul des boules à celles dont la borne inférieure est plus petite que cette borne supérieure. La figure 3.9 donne un exemple d'application de cette règle.

Ces modifications permettent d'appliquer notre méthode aux structures 3D et donc de calculer des bibliothèques de fragments.

3.5.2 Résultats sur les bibliothèques de tri-nucléotides

La table 3.3 donne le nombre de prototypes obtenus pour les différentes séquences mutées de tri-nucléotides (motifs purines/pyrimidines).

On observe une diminution de la taille des bibliothèques par un facteur d'environ 4 quelle que soit la séquence. Cela implique un nombre de poses de docking à échantillonner plus faible avec ces bibliothèques, pour des résultats similaires.

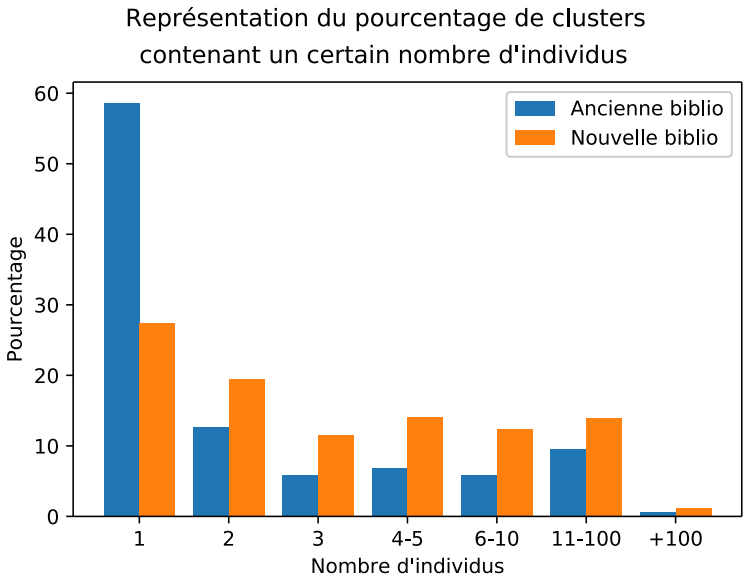


FIGURE 3.10 – Nombre de membres par cluster pour la bibliothèque Y-Y-Y à 1 Å.

Statistique sur les nouveaux clusters

Nous avons cherché à savoir en quoi ces bibliothèques sont différentes des anciennes. Pour cela, nous avons regardé sur la bibliothèque Y-Y-Y la proportion de clusters possédant 1, 2, 3, 4-5, 6-10, 11-100 membres ou plus (voir figure 3.10).

On remarque une diminution du nombre de singletons. Cela permet d'augmenter la fiabilité de nos prototypes, en diminuant la probabilité que cette conformation n'existe que dans une seule structure et soit donc très particulière, voire fausse (erreur dans la résolution de la structure). Il nous faudrait refaire les statistiques présentées dans le chapitre 2. En effet, pour les statistiques dans ce chapitre, nous enlevons les clusters contenant moins de 3 membres, ce qui signifie que nous avons enlevé 71% de nos clusters. Avec la nouvelle méthode, nous n'aurions à enlever "que" 47% des clusters. Ce nombre reste élevé, et signifie qu'il y a beaucoup de clusters peu fiables.

Validation pour le docking

Pour valider l'utilité de la nouvelle bibliothèque, nous avons réalisé des docking avec l'ancienne et la nouvelle bibliothèque et nous avons comparé les résultats.

Ce sont 34 docking de fragments répartis sur 22 complexes ARN-protéine qui ont servi de benchmark pour ce test. Nous avons comparé la distribution du nombre de poses quasi-natives (RMSD inférieur à 3 Å avec la pose native) dans le top 10000, 100000 et 1 million. Le test statistique utilisé pour la comparaison est un test des deux échantillons de Kolmogorov-Smirnov.

La première comparaison est faite sur 30 millions de poses entre la nouvelle et l'ancienne bibliothèque. Pour les trois tops choisis, la statistique de test à 95% considère que les répartitions sont les mêmes pour les deux bibliothèques, et donc que les résultats de docking sont similaires. Cela veut dire que la diminution de la taille de la bibliothèque n'influe pas sur les résultats de docking. Nous pensions que dans le cas où la taille de la bibliothèque diminue, et que l'on garde le même nombre de poses, les résultats seraient meilleurs. En effet, le nombre de poses pour une conformation donnée augmente, et il y a donc plus de chance d'obtenir un meilleur échantillonnage.

La deuxième comparaison des résultats de docking a été réalisée sur le nombre de positions initiales. Maintenant que nous savons qu'à positions initiales équivalentes les résultats sont les mêmes, nous nous intéressons au cas où l'on diminue le nombre de positions initiales.

Pour cela, nous avons pris comme facteur la diminution de la taille de la bibliothèque de fragments. On obtient ainsi un nombre de positions initiales de 7 millions au lieu de 30 millions. Cela diminue du même facteur la durée du docking complet (de $\approx 60h$ contre $\approx 12h$, ce qui est conforme au facteur).

Nous avons observé des statistiques de test un peu moins bonnes, puisque la p-value est de 0.93. Cela signifie quand même qu'en prenant une incertitude de 10%, on peut considérer que les résultats sont similaires, et donc que l'on peut utiliser la nouvelle bibliothèque pour réaliser des docking avec un nombre de positions initiales plus faible.

Nous avons une hypothèse pour expliquer ces résultats. Dans l'ancienne bibliothèque, il y avait bien plus de singletons, et donc la structure native peut être présente dans la bibliothèque, ce qui rend le docking beaucoup plus simple. Or, dans notre nouvelle bibliothèque, les prototypes ne sont pas des fragments, sauf pour les singletons. Cela implique qu'il est très peu probable d'avoir la structure native dans la bibliothèque. Nous aurions dû vérifier la présence des structures natives et les enlever le cas échéant.

Un petit exemple est proposé pour le complexe 1CVJ (protéine liant des poly-A en complexe avec un poly-A), qui est composé d'un ARN poly-A d'une taille de 9 nucléotides, dont 8 sont en contact avec la protéine. Nous considérons donc 6 tri-nucléotides pour ce complexe. Sur ce complexe nous donnons les valeurs du rang de la première occurrence d'une pose ayant un LRMSD inférieur à 3 Å, et le RMSD des 10 meilleures poses (voir tableau 3.4).

3.5.3 Utilisation de coordonnées internes

Pour contourner le problème lié à la superposition des fragments pour le calcul du cRMSD, nous avons essayé de définir des coordonnées internes (par opposition aux coordonnées carté-

Fragment1										
Nouvelle	1.5	1.6	2.1	2.3	2.3	2.5	2.6	2.6	2.7	2.8
Ancienne	0.7	0.9	0.9	1.2	2.5	2.8	2.8	2.8	3.0	3.0
Fragment2										
Nouvelle	0.8	1.2	2.7	2.7	2.8	2.8	2.8	2.9	2.9	3.0
Ancienne	0.7	1.6	1.9	1.9	1.9	2.0	2.0	2.0	2.0	2.1
Fragment3										
Nouvelle	1.4	1.5	1.7	1.8	1.9	2.1	2.1	2.2	2.3	2.4
Ancienne	1.1	1.1	1.3	1.5	1.6	1.6	1.6	1.6	1.6	1.6
Fragment4										
Nouvelle	1.4	1.8	1.8	1.9	1.9	1.9	2.1	2.1	2.2	2.2
Ancienne	1.2	2.1	2.1	2.1	2.3	2.4	2.4	2.4	2.4	2.5
Fragment5										
Nouvelle	1.7	1.7	1.8	2.5	2.5	2.6	2.8	2.8	2.8	2.8
Ancienne	2.9	3.0	3.1	3.1	3.1	3.2	3.2	3.2	3.6	3.6
Fragment6										
Nouvelle	2.0	3.3	3.3	3.4	3.7	3.7	3.8	3.9	3.9	4.0
Ancienne	1.4	1.6	2.5	2.8	2.9	2.9	2.9	3.0	3.1	3.1

Fragment1	
Nouvelle	13358
Ancienne	1422
Fragment2	
Nouvelle	875
Ancienne	131
Fragment3	
Nouvelle	471
Ancienne	4
Fragment4	
Nouvelle	10
Ancienne	47
Fragment5	
Nouvelle	77
Ancienne	43501
Fragment6	
Nouvelle	1016
Ancienne	7516

TABLE 3.4 – À gauche, le tableau présentant les 10 meilleures poses en termes de LRMSD parmi les 10 millions. À droite, les rangs des premières poses quasi-natives (en dessous d'un seuil de 3 Å par rapport à la structure native).

siennes) des fragments pour les comparer sans avoir à les superposer, permettant ainsi d'avoir une matrice de distance plus juste que celle obtenue en superposant tous les fragments sur un seul.

Cette idée a d'abord été explorée par une stagiaire de M2, Alix Delannoy, et a abouti à une présentation orale à JOBIM en 2021 [18] (hal-03540927). J'ai co-encadré le travail qui est présenté ici.

Présentation du choix des coordonnées internes

Tout d'abord nous travaillons avec tous les tri-nucléotides de la PDB, que nous représentons avec le gros-grain ATTRACT (voir figure 3.11). Pour définir les coordonnées internes pertinentes, nous nous sommes inspirés de méthodes existantes [84, 96]. Nous avons sélectionné et calculé 6 distances et 9 angles dièdres :

— les 3 distances entre les bases, en utilisant pour chaque base le pseudo-atome le plus éloigné du squelette (GA4, GG4, GC3, GU3) ;
— la distance entre GS2 (sucre du premier nucléotide) et GA4 (base du dernier nucléotide) ;
— la distance entre GA4 (base du premier nucléotide) et GS2 (sucre du dernier nucléotide) ;
— la longueur du squelette, de GP (phosphate du premier nucléotide) à GS1 (sucre du dernier

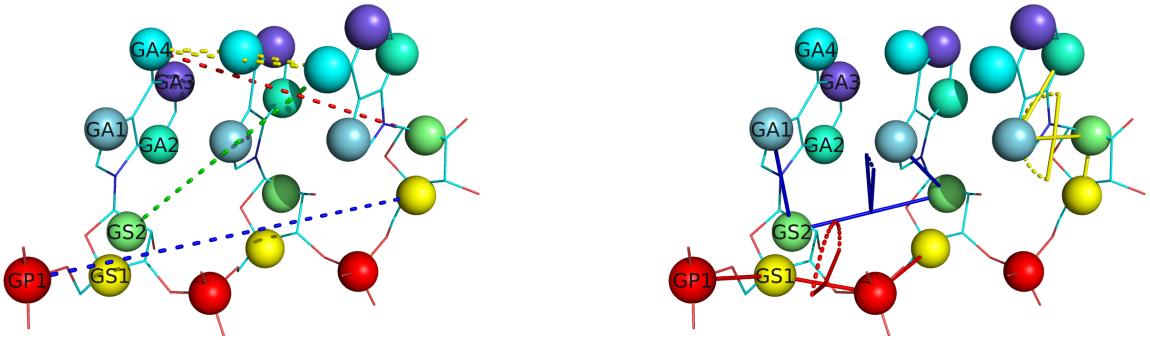


FIGURE 3.11 – Coordonnées internes sélectionnées : distances (à gauche) et angles (à droite) sur un tri-nucléotide dans les représentations tout-atome (bâtons) et gros-grain (boules), avec le nom des pseudo-atomes sur le premier nucléotide.

nucléotide) ;

— les 3 angles du squelette entre les pseudo-atomes GP et GS1 (nucléotides consécutifs) ;

— les 3 angles μ entre le sucre et la base de chaque nucléotide, en utilisant les pseudo-atomes GS1-GS2-GA1-GA2 ;

— les 3 angles χ entre l'axe GS2-GA1.

Les distributions des coordonnées internes sont présentées en figure 3.12.

Comme attendu, parmi les distances, toutes les distances bases-bases présentent une grande variance, tandis que toutes les distances sucres-bases sont plus conservées parmi les fragments. Parmi les angles, les angles χ représentant l'orientation relative de deux nucléotides présentent une grande variance, tandis que les angles μ entre le sucre et la base de chaque nucléotide sont beaucoup plus conservés.

Lien entre les coordonnées internes et la RMSD

Nous avons analysé la distribution des valeurs de coordonnées internes et de cRMSD parmi les fragments, et évalué comment relier les différences entre deux fragments. Nous avons sélectionné quatre fois un échantillon aléatoire de 10% de l'ensemble des fragments, et calculé pour toutes les paires de fragments (i) la cRMSD par paire après superposition (ii) la différence entre chaque coordonnée interne, et (iii) la somme des différences sur les distances ou les angles internes. Nous avons calculé, pour chacune des 15 coordonnées et des 2 sommes de coordonnées, la valeur seuil au-dessus de laquelle toutes les paires de fragments ont une cRMSD supérieure à 1 Å. En pratique, pour rendre le filtrage plus efficace, nous avons autorisé 1 faux négatif pour 1000 positifs (ce qui signifie que 1/1000 paire avec cRMSD $<$ 1 Å sont au-dessus de ce seuil). Nous avons ensuite calculé les moyennes pour les 17 valeurs seuils sur les 4 échantillons aléatoires.

Nous avons appliqué ces seuils sur l'ensemble complet de 39431 fragments obtenus en récu-

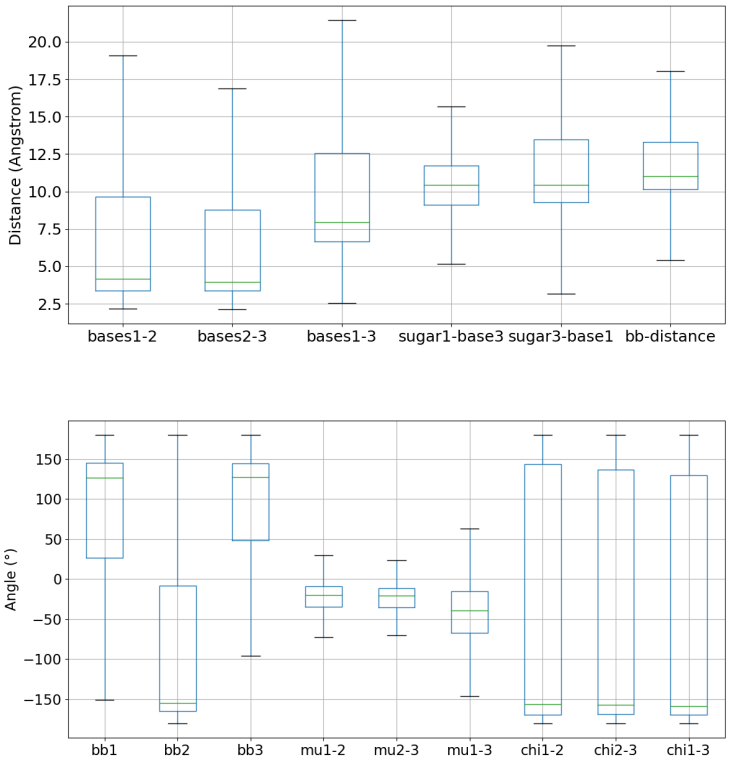


FIGURE 3.12 – Distribution des coordonnées internes obtenues pour les 39431 fragments.

base 1-2	base 2-3	base 1-3	sucre 1 - base 3	sucre 3 - base 1	squelette	somme
43%	47%	43%	57%	49%	52%	23%

TABLE 3.5 – Pourcentage de paires qui sont sous le seuil retenant 99,9% des paires compatibles, pour chaque distance, pour les 39431 fragments.

squelette 1	squelette 2	squelette 3	μ 1-2	μ 2-3	μ 1-3	χ 1	χ 2	χ 3	somme
76%	67%	70%	49%	52%	45%	80%	77%	89%	28%

TABLE 3.6 – Pourcentage de paires qui sont sous le seuil retenant 99,9% des paires compatibles, pour chaque angle, pour les 39431 fragments.

pérant tous les fragments de la PDB et en retirant les redondances à 0,2 Å RMSD. Nous avons calculé toutes les coordonnées internes et leurs différences par paire, puis nous avons sélectionné les paires dont les 17 valeurs étaient inférieures au seuil correspondant. La superposition paire à paire et le calcul de la cRMSD ont été effectués uniquement sur ce sous-ensemble. Pour les autres paires, la cRMSD a été considérée comme supérieure à 1 Å.

Nous avons testé la combinaison des 17 seuils sur les quatre échantillons aléatoires. Le pourcentage réel de valeurs cRMSD inférieures à 1 Å est compris entre 8,6 et 9,7% (moyenne de 9,2%) dans chaque échantillon, et on suppose qu'il se situe dans la même fourchette pour l'ensemble des fragments. Nous avons constaté que la proportion de paires pour lesquelles toutes les valeurs sont inférieures aux 17 seuils se situe dans une fourchette de 14 à 16% dans les échantillons, ce qui signifie que nous pouvons réduire le nombre de paires à aligner à seulement environ 15% de toutes les paires. Parmi les paires conservées, 54 à 62% sont des vrais positifs. Cet ensemble de seuils a ensuite été appliqué à l'ensemble des 39431 fragments. Comme prévu, 15% des $1,6 \cdot 10^9$ paires ont été identifiées comme potentiellement inférieures à 1 Å cRMSD. Celles-ci ont été sélectionnées pour un alignement par paire et un calcul de cRMSD. Pour les autres paires, la cRMSD a été considérée comme supérieure à 1 Å.

En examinant chaque seuil individuel, le filtrage le plus efficace est fourni par la somme des distances, la somme des angles et les distances base-base, tandis que les angles χ donnent les filtres les moins efficaces.

Ce travail de stage a abouti à un filtre qui permet de ne réaliser qu'environ 15% de toutes les superpositions 2 à 2 pour l'utilisation des méthodes de clustering des fragments.

Pour estimer le gain du pré-filtrage avec les coordonnées internes en termes de temps CPU, nous avons utilisé nos calculs de la matrice de cRMSD complète pour les 4 échantillons, avec ou sans pré-filtrage, sur 1 CPU. Le calcul des coordonnées internes et de leurs différences par paires prend moins de 1 seconde par échantillon. Le calcul de la matrice de cRMSD pour 4773 fragments AAA prend environ 23min pour toutes les paires, et moins de 5min pour les paires pré-filtrées.

Sur l'ensemble des fragments, le calcul des coordonnées internes et de leurs différences par paires prend respectivement 2 secondes et 4min.

3.5.4 Améliorations possibles des coordonnées internes

Suite à ce travail, nous avons exploré un peu plus l'utilisation de coordonnées internes, l'objectif étant de pouvoir passer d'une représentation interne à une représentation cartésienne et inversement.

Pour cela, les nouvelles coordonnées internes définies sont les distances entre pseudo-atomes consécutifs, les angles entre 3 pseudo-atomes consécutifs et les angles dièdres entre 4 atomes consécutifs. Cela donne entre 48 et 57 coordonnées internes permettant la transition des coordonnées internes à cartésiennes (voir figure 3.13).

L'idée d'avoir des coordonnées internes est de pouvoir transformer 3 pseudo-atomes, plus une distance, plus un angle, plus un angle dièdre en coordonnées cartésiennes d'un quatrième pseudo-atome.

Statistiques sur ces coordonnées internes

Pour savoir si ces coordonnées internes peuvent être intéressantes pour le clustering, nous avons récupéré tous les fragments du motif Y-Y-Y dans les structures ayant une résolution d'au moins 3 Å. Cela représente 11814 tri-nucléotides après avoir enlevé les redondances à 0,2 Å RMSD. Pour chacun de ces tri-nucléotides, l'ensemble des 87 coordonnées internes est calculé.

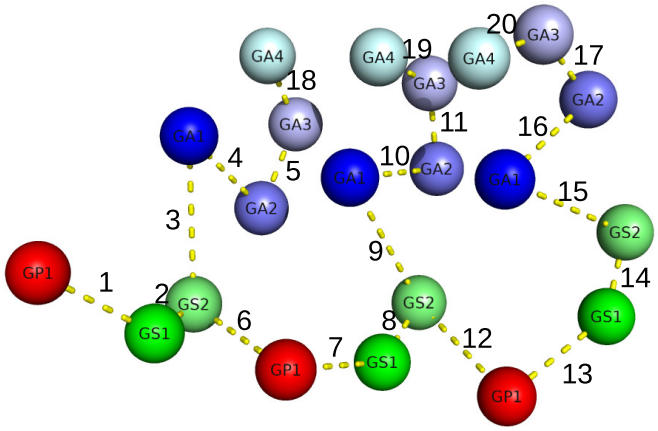
L'étude de ces coordonnées internes laisse supposer que seuls les angles dièdres sont importants, la variabilité des distances et des angles étant presque nulle (voir figure 3.14).

L'absence de variabilité parmi les distances et les angles était attendue, puisque la longueur des liaisons covalentes entre atomes est fixe, de même que les angles. La faible variabilité que l'on observe peut être due à la résolution des structures expérimentales.

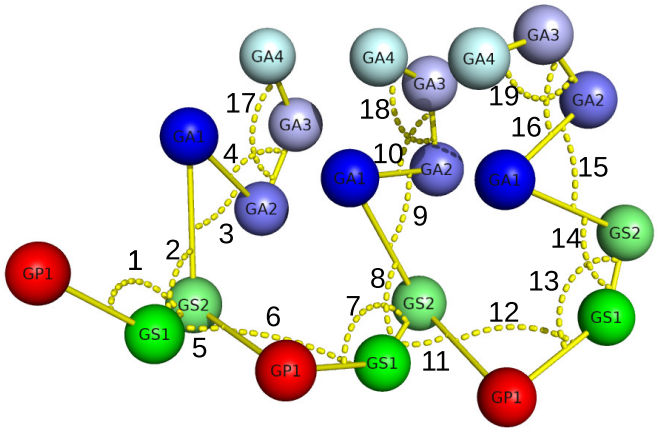
Pour les angles dièdres, les angles 2, 5, 8, 11 et 14 semblent les plus variables. Cela semble logique puisque les angles 2, 8 et 14 représentent les rotations des 3 bases azotées autour de leur liaison au sucre. Les angles 5 et 11 quant à eux représentent la liaison phosphate-sucre entre les nucléotides. L'ensemble de ces 5 liaisons semble donner la forme globale d'un tri-nucléotide, ce qui est exactement ce que nous avions imaginé.

3.5.5 Utilisation de ces coordonnées internes

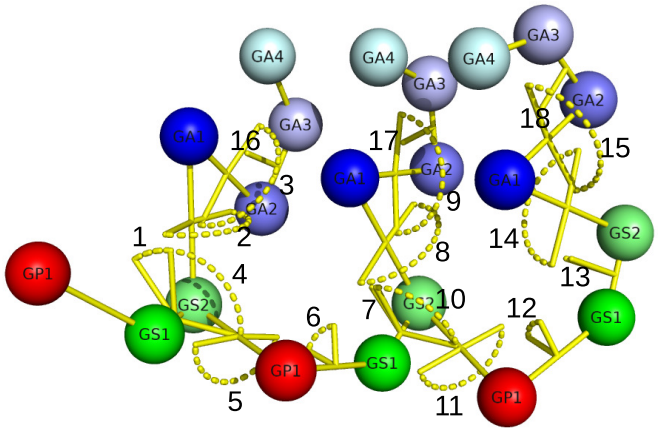
L'objectif de ce deuxième jeu de coordonnées internes est de ne plus avoir à réaliser de superposition du tout, contrairement au premier jeu de coordonnées internes qui a eu pour



(a) Présentation des distances retenues.



(b) Présentation des angles retenus.



(c) Présentation des angles diédraux retenus.

FIGURE 3.13 – Les différentes coordonnées internes retenues.

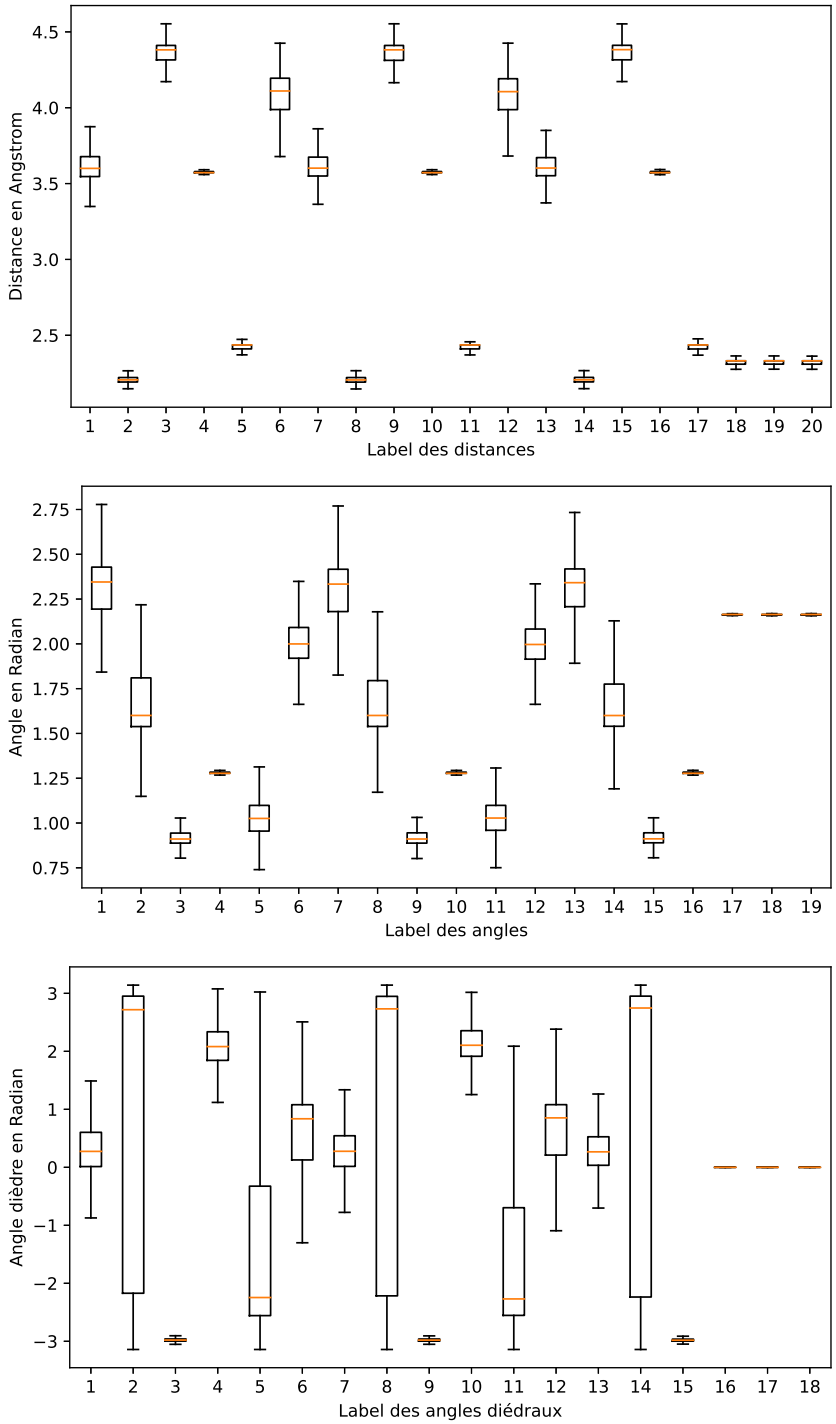


FIGURE 3.14 – Boxplots montrant les distributions des différentes coordonnées internes.

utilisation de limiter le nombre de superpositions à réaliser.

Pour cela, on considère le passage aux coordonnées internes comme un noyau. On se retrouve donc avec un produit scalaire dans notre nouvel espace de représentation qui est celui des coordonnées internes. Il faut ensuite relier les valeurs de ce produit scalaire à celles que nous voulons estimer, les valeurs de la cRMSD. Pour réaliser cela, une idée est d'utiliser les *Kernel Target Alignment* (KTA) [10]. Par manque de temps, cette approche n'a pas été appliquée, mais reste une piste possible pour supprimer la nécessité de la superposition entre les structures.

Une autre application de ces coordonnées internes est de les utiliser pour créer des bibliothèques de fragments *de novo*. Pour cela, on pourrait essayer de discrétiser les rotations des angles dièdres et ainsi créer des bibliothèques artificielles.

3.6 Conclusion sur l'inférence des ϵ-réseaux

Dans ce chapitre nous avons présenté une méthode originale pour l'inférence des ϵ-réseaux. Cette méthode repose sur la CAH qui produit un partitionnement spécifique des données. Le linkage Radius que nous avons introduit dans ce chapitre, propose la meilleure solution (c.-à-d., les ϵ-réseaux de plus faible cardinalité) sur l'ensemble de test. Et ce notamment face aux inférences basées sur les ensembles dominants de graphes. Ces résultats peuvent être expliqués de plusieurs manières, tout d'abord les solutions proposées par ces approches sur les graphes produisent des ϵ-réseaux propres. De plus, l'application de cette méthode s'est faite à travers un package Python qui précise que les solutions obtenus ne sont pas optimales. Il faudrait appliquer les algorithmes présentés dans [49].

Cette méthode d'inférence a été appliqué pour la création des bibliothèques de fragments. Nous avons observés une diminution de la cardinalité de nos bibliothèques. Cependant cela n'a pas diminué le rang des premières poses quasi-natives. Le gain se fait donc sur le nombre de positions de docking initiales et donc sur le temps de calcul.

Enfin, nous avons commencé l'exploration de l'utilisation des coordonnées internes pour représenter nos tri-nucléotides. L'idée serait de se passer de la superposition pour la comparaison de nos fragments. Ce travail a été fait en deux temps. Premièrement l'utilisation de coordonnées internes pour savoir s'il est nécessaire de calculer le RMSD après superposition. Et deuxièmement, des coordonnées internes pour décrire complètement les fragments, et donc faire une approximation beaucoup plus fine du RMSD après superposition. La suite de ce travail pourrait s'appuyer

sur d'autres travaux tel que [1], qui propose une réduction de la dimension des données pour une meilleure visualisation et comparaison.

Chapitre 4

Docking des épingles à cheveux d'ARN

Ce chapitre expose la contribution au sujet initial de la thèse. En effet ce travail porte sur l'appariement d'une épingle à cheveux, une structure d'ARN 2D comportant une partie simple brin et une partie double brin. L'objectif est de réaliser le docking de la structure complète en contact avec la protéine au moins par la partie simple brin.

Pour cela, nous prenons en compte que la partie double brin est moins flexible que la partie simple brin. La stratégie que nous adoptons est de couper la structure au niveau de la boucle simple brin. Le docking de la boucle simple brin en contact avec la protéine se fait en utilisant l'approche par fragment décrite dans la section 1.4.4. L'étude du docking de la partie simple brin a fait l'objet d'une présentation orale à la conférence JOBIM en 2020 [60] (hal-02927185). La partie double brin étant moins flexible, le docking se fera d'un bloc.

Le chapitre va présenter d'abord le docking de la boucle simple brin puis les possibilités de docking des hélices double brin.

4.1 Méthode d'application du docking par fragment aux boucles des épingles à cheveux

Pour le docking de la boucle simple brin des épingles à cheveux, nous avons adapté la méthode présentée dans la section 1.4.4. L'obtention de modèles complets à haute précision d'un ARN simple brin lié par assemblage de fragments est possible en utilisant des points d'ancrage connus sur la protéine. En revanche, le graphe d'assemblage reliant les poses compatibles décrit dans la section 1.4.4 est bien trop grand si l'on n'applique pas une telle contrainte. Une caractéristique utile des épingles à cheveux pour la modélisation de leurs interactions avec les protéines est que la distance entre les nucléotides aux extrémités de la boucle est contrainte [9]. Notre

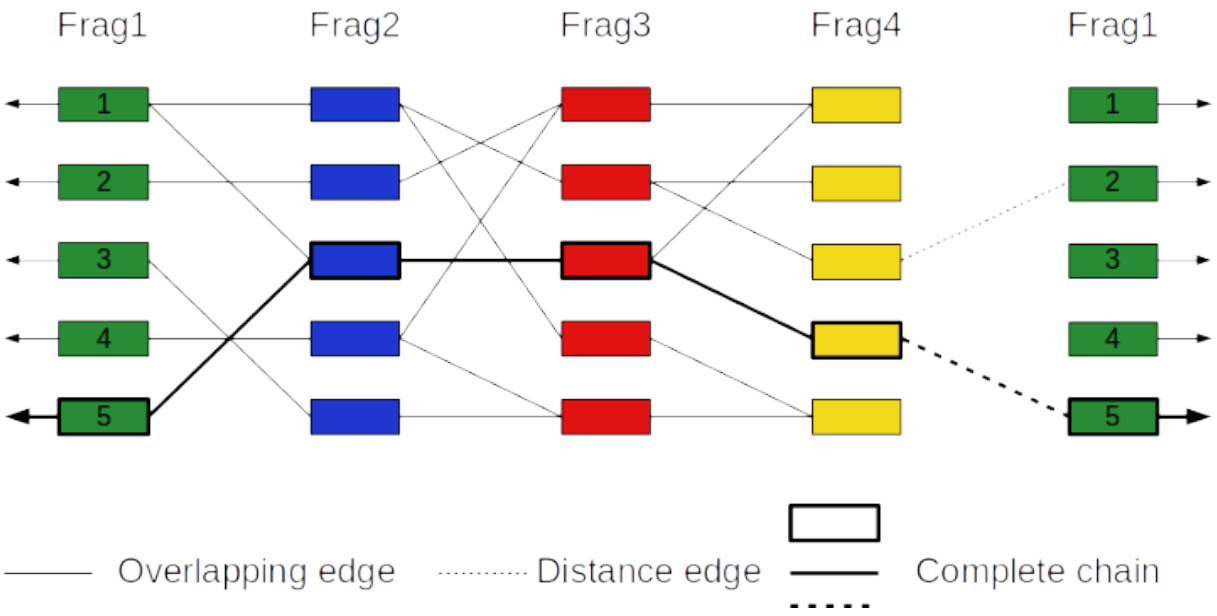


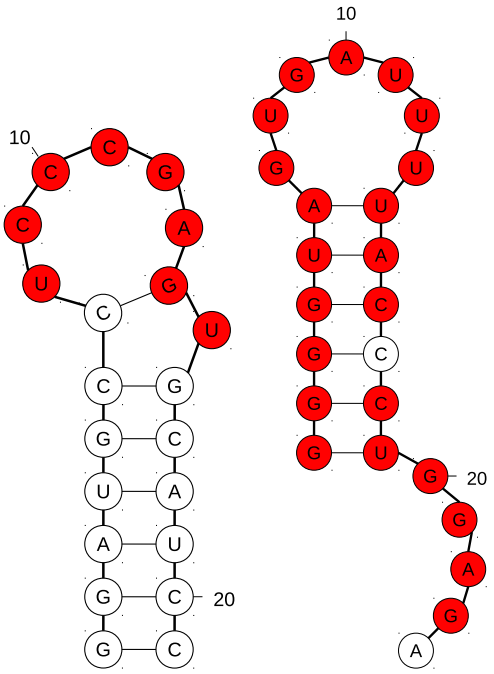
FIGURE 4.1 – Graphe dédié pour les épingles à cheveux décrit dans la section 1.4.4 et avec l'ajout d'un nouveau type d'arête.

méthode est basée sur la conjecture qu'une exploitation appropriée de cette caractéristique peut s'avérer suffisante pour contraindre l'assemblage de manière à relâcher le besoin initial de points d'ancrage. Un nouveau type d'arête est introduit, pour relier les poses du dernier et du premier fragment. Ainsi, aucune information de contact ARN-protéine ou d'interface sur la protéine n'est utilisée. L'information utilisée est l'identification (la position dans la séquence) des nucléotides de la fermeture de l'épingle à cheveux, qui peuvent être obtenus par prédiction de la structure secondaire. Cette connexion est ajoutée dans le graphe lorsque la distance euclidienne entre le phosphate du premier nucléotide du premier fragment et le phosphate du dernier nucléotide du dernier fragment appartient à l'intervalle 11,8-24,7 Å (intervalle observé dans le benchmark présenté en section 4.2). Le nouveau graphe est représenté sur la figure 4.1. Par rapport au cas général (donc sans les points d'ancrage), l'ensemble des chaînes considérées ici est plus petit, puisqu'il ne retient que celles incluses dans un cycle du nouveau graphe.

Pour la validation de la méthode, un benchmark a été produit en utilisant protNAff (voir chapitre 2).

4.2 Création du benchmark

La création du benchmark est composée de deux étapes principales. La première étape prend en entrée la liste de codes PDB de toutes les structures expérimentales disponibles de complexes



(a) Structure secondaire de l'ARN du complexe 1RKJ.

(b) Structure secondaire de l'ARN du complexe 5UDZ.

FIGURE 4.2 – Les nucléotides en rouge sont en contact (à moins de 5Å) de leur protéine respective.

épingle à cheveux-protéine pour lesquelles la partie simple brin est en contact avec la protéine (les structures redondantes sont supprimées). Cet ensemble initial contient 19 complexes. Pour chaque épingle à cheveux, le docking de tous les conformères de tous les motifs présents dans sa boucle est effectué sur la protéine liée correspondante en utilisant ATTRACT. Pour ce docking les bibliothèques de fragments utilisées sont les bibliothèques initiales (donc ce ne sont pas celles améliorées au chapitre précédent). Le nombre de positions initiales est de 30 millions, avec environ $10 \cdot 10^3$ poses par conformère puisqu'il y a en moyenne 3000 conformères par motif. Nous ne retenons que les complexes pour lesquels tous les fragments ont au moins une pose quasi-native. Une pose quasi-native est une pose dont la RMSD avec la position native (expérimentale) est inférieure à 3 Å. À ce niveau, seuls 2 complexes sont sélectionnés : 5UDZ [87] et 1RKJ [39] . Ces complexes sont respectivement un RRM en lien avec un microARN et un complexe formé par un domaine RBP situé dans la nucléoline et un pré-ARNr (voir la figure 4.2 pour les structures secondaires et les contacts avec les protéines).

Pour ces deux complexes, au moins un fragment a la pose quasi-native la mieux classée à un rang supérieur à 10^6. Ceci explique, au moins pour notre jeu de données, la nécessité de la deuxième étape, pour garder une chance d'obtenir un assemblage pertinent.

La deuxième étape est un affinement qui consiste à refaire le docking avec un sous-ensemble de conformères structurellement assez proches de la position native. Cela correspond à l'élimination des conformères qui sont structurellement trop différents du fragment natif pour constituer une pose proche de la position native. Pour chaque motif, un sous-ensemble de conformères est créé qui contient seulement les dix plus proches (selon la cRMSD) de la forme liée. Le nombre de poses par conformère est augmenté à $50 \cdot 10^3$, de sorte que le nouveau nombre de poses par fragment soit de $500 \cdot 10^3$. À ce niveau, le critère pour retenir un complexe est le suivant : pour chaque fragment, le rang du score ATTRACT de la pose quasi-native la mieux classée doit être inférieur à un seuil. Évidemment, cette étape, qui fait intervenir des informations en principe inconnues, fait de notre approche expérimentale une preuve de concept.

Le tableau 4.1 présente l'ensemble des épingles à cheveux sélectionnées lors de la première étape, avec les résultats de docking correspondants.

		frag1	frag2	frag3	frag4	frag5
5UDZ	premier docking	5743309	423163	15403	27578	2138595
	deuxième docking	90424	2114	285	1334	34291
1RKJ	premier docking	1382237	1110963	13187	599859	
	deuxième docking	7056	14949	191	1275	

TABLE 4.1 – Premier rang d'une pose quasi-native pour les deux épingles à cheveux.

Les chiffres du tableau 4.1 établissent que l'obtention d'une chaîne de poses quasi-natives pour 5UDZ et 1RKJ nécessite de considérer un maximum de $6 \cdot 10^{24}$ et $5 \cdot 10^{16}$ chaînes possibles respectivement. Ces chiffres sont basés sur le calcul suivant : pour chaque fragment de chaque séquence, le nombre de poses considérées est environ les rangs des poses quasi-natives classées en premier (ici $1 \cdot 10^5$ pour 5UDZ et $15 \cdot 10^3$ pour 1RKJ). Ce choix arbitraire correspond à une hypothèse raisonnable sur les informations qui pourraient être déduites de données expérimentales telles que des informations d'homologies. Dans la suite, les résultats sont fournis pour 1RKJ uniquement.

4.2.1 Résultats pour l'épingle à cheveux 1RKJ

La séquence de la boucle de l'épingle à cheveux est UCCCGA (donc quatre fragments). Les trois paramètres à régler pour dériver les chaînes sont :

— l'intervalle dans lequel contraindre la distance entre les deux nucléotides aux extrémités de la boucle ;

— le nombre de poses considérées par fragment ;

— le seuil sur la RMSD minimal entre les nucléotides communs des fragments qui se chevauchent pour considérer deux poses comme connectées.

Nous avons défini la valeur de la distance entre les deux nucléotides aux extrémités comme étant dans l'intervalle observé de 11,8-24,7 Å. Le nombre de poses considérées est ici de $15 \cdot 10^3$. Quant au troisième paramètre, la valeur retenue est la valeur minimale permettant d'assurer de générer une chaîne reliant les poses quasi-natives. Cette valeur est obtenue en testant de ne relier que ces poses de façon itérative pour des valeurs croissantes. Elle est de 2,6 Å.

Pour ces différentes valeurs, le nombre de chaînes est de 47617288. Les valeurs de RMSD utilisées pour qualifier les chaînes sont les RMSD moyens des poses avec les fragments natifs correspondants, ce qui signifie que les nucléotides au milieu de la chaîne ont un poids plus important, car ils sont comptés trois fois. La meilleure solution et le modèle expérimental sont représentés sur la figure 4.3. Dans l'ensemble des chaînes obtenues, nous obtenons 732 solutions acceptables (avec une RMSD inférieure à 5 Å), ce qui fait $0,015\%$¸ et 122 bonnes solutions (RMSD inférieure à 3 Å), avec le meilleur modèle à 1,9 Å.

La section 1.4.4 a présenté les deux méthodes mises en œuvre pour trier les chaînes. La première méthode se base sur la moyenne géométrique des rangs des poses d'une chaîne. La deuxième méthode s'appuie sur le nombre de chaînes dans lesquelles les poses sont impliquées.

Nous discutons maintenant de leur efficacité. En utilisant comme critère la moyenne géométrique des rangs des poses, le plus petit rang d'une solution parmi les 732 acceptables est de 4783648. Ceci est proche du top 10%, mais laisse trop de faux positifs au-dessus. De plus, si les 732 solutions acceptables sont réparties de manière uniforme sur l'ensemble des chaînes, la première occurrence devrait avoir lieu autour du rang 65000. Ainsi ce critère ne semble pas efficace pour discriminer les solutions acceptables.

	frag1	frag2	frag3	frag4
Toutes les poses	206742	502600	654520	1105597
Poses dans une solution	5701	16849	366	1672
Moyenne pour toutes les poses	11985	15830	24865	20151

TABLE 4.2 – Plus grand nombre de chaînes dans lesquelles une pose des 47617288 chaînes (toutes les poses) et une pose des 732 solutions (poses dans une solution) sont impliquées. La dernière ligne est le nombre moyen de chaînes dans lesquelles est impliquée une pose.

Le deuxième critère est le nombre de chaînes dans lesquelles les poses sont impliquées (voir tableau 4.2). Nous avons calculé que les poses qui sont impliquées dans de bonnes chaînes ne sont pas impliquées dans plus de chaînes que le reste des poses.

Afin de diminuer le nombre de chaînes retenues, nous avons testé d'abaisser à 10-15 Å l'in-

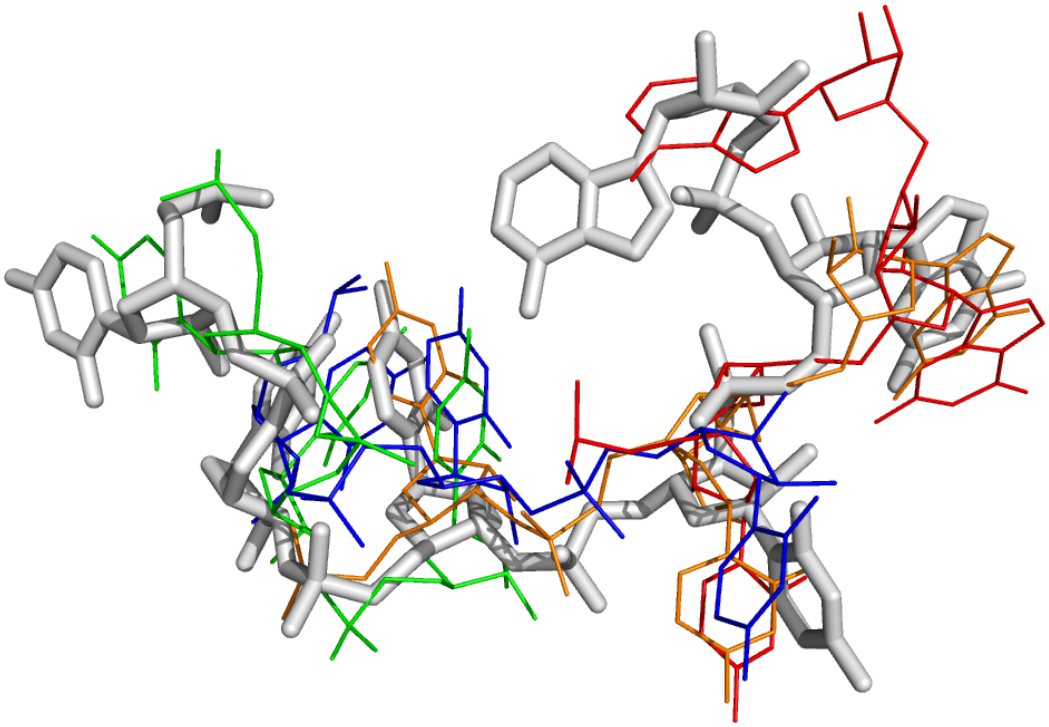


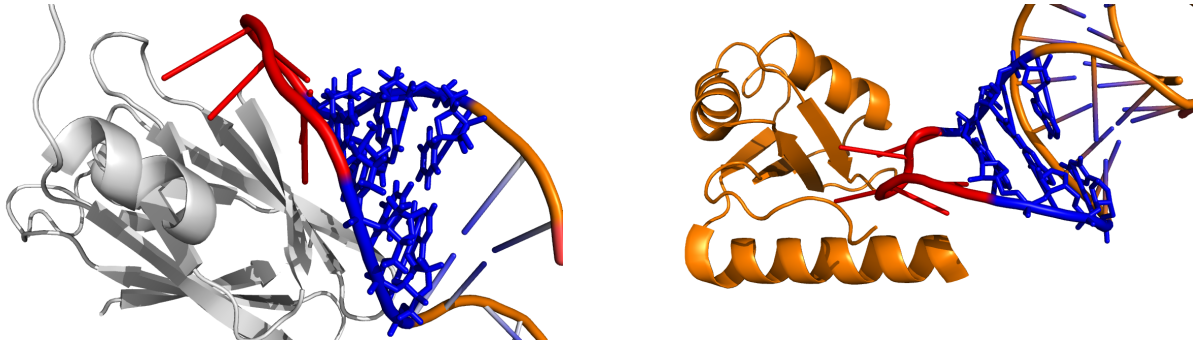
FIGURE 4.3 – Les poses composant la meilleure solution possible considérant le RMSD avec la chaîne native (en blanc). Fragment 1 en vert, fragment 2 en bleu, fragment 3 en orange et fragment 4 en rouge.

tervalle pour la distance entre le phosphate du premier nucléotide du premier fragment et le phosphate du dernier nucléotide du dernier fragment (qui était initialement fixé à 11,8-24,7 Å), puisque cette distance est de 13,2 Å dans 1RJK. Cet ajustement a permis de diminuer la cardinalité de plus de 30%, en obtenant 32765494 chaînes, mais a aussi diminué le nombre de bonnes solutions de 732 à 517. Ce changement de distance n'a pas augmenté le pourcentage de solutions acceptables dans les chaînes.

4.3 Possibles améliorations

Compte tenu de la difficulté de la tâche abordée, nos résultats semblent prometteurs pour l'échantillonnage, mais restent à améliorer du côté de l'évaluation (voir annexe B). La difficulté principale que rencontrent ces méthodes d'assemblage de fragments est que, pour de nombreux complexes, le docking des fragments n'est pas suffisamment bon. En effet, de nombreux fragments n'obtiennent même pas une pose quasi-native, ce qui rend l'assemblage impossible. La connaissance de la structure secondaire semble être suffisamment pertinente pour remplacer la connaissance des points d'ancrage en ce qui concerne la diminution du nombre de chaînes possibles. Les résultats étaient cependant meilleurs avec les points d'ancrage, avec un pourcentage plus élevé de chaînes correctes et une meilleure chaîne avec un plus petit RMSD dans la plupart des cas [15]. Ceci était prévisible puisque la position des nucléotides (et donc aussi la distance entre ces nucléotides) constitue une contrainte plus forte que la distance seule. La connaissance de la distance seule a permis d'assembler et d'évaluer toutes les chaînes possibles, sans le précédent pré-filtrage heuristique des poses les plus connectées [37], et d'obtenir un modèle plus précis que ceux obtenus avec ce pré-filtrage (1,9 Å, au lieu de 3,6-5,7 Å). En revanche, l'hypothèse d'une fermeture de boucle étant plus faible que celle de la position exacte des extrémités de la chaîne sur la protéine, l'approche actuelle retient plus de faux positifs que le docking ancré (0,015% de modèles corrects dans les chaînes assemblées contre 2-3%). De plus, la contrainte de fermeture de boucle n'est appliquée qu'au moment de l'assemblage, et ne change rien au docking des fragments. Ceci ne réduit donc pas la taille du graphe mais juste son exploration.

Une amélioration de notre méthode pourrait résulter d'un changement de cible pour la distance. La distance entre le phosphate du premier nucléotide et le sucre du dernier nucléotide semble être une contrainte plus forte que la distance entre les phosphates aux extrémités. L'intervalle de distance phosphate-sucre observé dans l'ensemble des structures du benchmark est de 13,7-21,6 Å, ce qui est plus resserré que l'intervalle phosphate-phosphate (11,8-24,7 Å)). De plus, la variance de la distance phosphate-sucre (2,4) est plus petite que celle de la distance



(a) Hélice en contact avec la protéine. (b) Hélice sans aucun contact avec la protéine.

FIGURE 4.4 – Exemple de deux épingles à cheveux, une avec contact de l'hélice et l'autre sans.

phosphate-phosphate (5,9). Le fait d'avoir une plus petite variance permet d'être plus strict sur l'intervalle à considérer lors de l'assemblage, et donc de diminuer le nombre de faux positifs.

Enfin, une amélioration de la méthode pourrait aussi être obtenue en considérant des informations biologiques supplémentaires lors du docking. Ces informations pourraient être sur des contacts, ou sur des interfaces sur la protéine. Cela représente une contrainte très forte et pourrait réduire drastiquement le nombre de faux positifs.

4.4 Assemblages des épingles à cheveux

Maintenant que nous avons vu le docking de la partie simple brin des épingles à cheveux d'ARN, nous allons voir les solutions que nous proposons pour le reste de la structure. Ces solutions n'ont pas pu être implémentées du fait d'un manque de temps.

4.4.1 Docking des hélices

Pour le docking des hélices, nous avons créé des bibliothèques de fragments d'hélices. L'objectif de ces bibliothèques est de poser des hélices, puis de les lier à la partie simple brin obtenue précédemment. Pour cela, deux solutions ont été pensées dépendantes des contacts entre l'hélice et la protéine (voir figure 4.4). Les fragments choisis font six nucléotides : trois paires appariées de manière canonique. Cela fait donc 64 séquences possibles (comme pour les tri-nucléotides), puisque l'on considère uniquement les appariements canoniques. J'ai créé ces bibliothèques de la même manière que présentée dans le chapitre 2. Le nombre de représentants obtenus pour une précision de 1 Å de cRMSD varie entre 26 et 67 selon les séquences.

La première méthode suppose que l'hélice interagit avec la protéine. Dans ce cas, les hélices peuvent être amarrées sur la protéine, puis reliées selon un critère de distance et d'angles entre

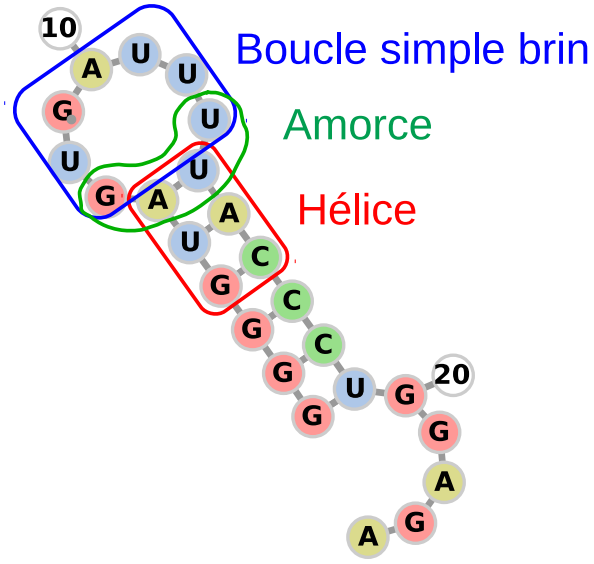


FIGURE 4.5 – Schéma d'une amorce sur une épingle à cheveux.

les nucléotides de l'hélice et ceux de la boucle.

La deuxième méthode aurait pour objectif d'utiliser des structures de fermeture de boucle (dites amorces, présentées en figure 4.5), pour contourner le fait que l'hélice n'interagit pas ou très peu avec la protéine. Ces structures contiennent quatre nucléotides, avec deux nucléotides en simple brin et deux en double brin. Il existe ainsi 16 séquences possibles d'amorces. En superposant les deux nucléotides simple brin sur les deux nucléotides de fermeture de la boucle simple brin préalablement amarrée, il est possible d'obtenir la transition avec la partie double brin. La partie double brin pourrait ensuite servir à superposer toutes les structures de la bibliothèque sur cette amorce pour garder la meilleure superposition. Ces amorces ont été obtenues, mais n'ont pas été utilisées. Pour ces structures nous avons choisi de muter les nucléotides de la même manière que dans la section 2.2.3. Cela permet de passer de 16 séquences à 4 séquences, ce qui augmente le nombre de structures. Ainsi pour la séquence AA, 70 structures sont obtenues, 49 pour AC, 21 pour CA, et 52 pour CC.

Une étude de la diversité de ces structures a été faite (comparaison des cRMSD) en gros-grain. La distance maximum observée au sein de chaque séquence est d'environ 1.6 Å. Nous avons conclu qu'utiliser une seule structure moyenne par séquence serait suffisant.

4.4.2 Perspectives liées au docking complet

Le fait de réaliser le docking des hélices permettrait aussi de diminuer le nombre de chaînes simple brin à envisager. En effet, en vérifiant si les nucléotides fermant la boucle ont une géométrie compatible avec l'appariement des bases des nucléotides voisins, il est possible de supprimer certaines chaînes obtenues. Cela consistera à évaluer si au moins un conformère de la bibliothèque d'amorces peut être ajusté sur les deux nucléotides terminaux de façon à établir un appariement en base avec au moins un conformère ajusté de l'autre côté de la boucle.

4.5 Conclusions sur le docking des épingles à cheveux

Ce chapitre présente des résultats de docking des parties simple brin d'épingles à cheveux sur une protéine, ainsi que des idées pour le docking des parties double brin.

Nous avons vu que le docking des parties simple brin reste très complexe si l'on relâche les contraintes des points d'ancrage. La distance entre nucléotides aux extrémités est une contrainte intéressante pour réduire le nombre de chaînes à considérer, mais ne suffit pas actuellement à garantir des résultats utilisables, du fait notamment de la mauvaise évaluation des poses (voir annexe B).

Bien entendu des connaissances biologiques supplémentaires amélioreraient ces résultats.

Enfin, une extension naturelle de ce travail consiste à appliquer son principe à différentes structures secondaires d'ARN (connues ou prédites), dans le but de réaliser un docking global du complexe. À terme, la contrainte ne devrait pas nécessairement impliquer la connaissance de la structure secondaire, mais pourrait bénéficier de toute connaissance de la distance entre deux nucléotides.

Conclusions

L'objectif de la thèse était de répondre au problème de combinatoire défini par l'assemblage de fragments d'ARN amarrés sur une protéine. Pour cela nous avons traité deux sous-problèmes : le premier est l'utilisation de contraintes de distances pour réduire le graphe à explorer, le deuxième est de réduire la taille des bibliothèques de fragments. Pour réaliser la diminution de la cardinalité de ces bibliothèques, nous nous sommes intéressés aux calculs des ϵ-réseaux à partir des méthodes de classification ascendante hiérarchique. Les contributions présentées dans ce document ont répondu à notre objectif via ces deux approches, tout en essayant de généraliser au maximum ces résultats pour la communauté scientifique.

La première contribution présentée au chapitre 2 répond à des besoins de création de benchmarks ou de bibliothèques structurales dans des contextes très spécifiques. L'outil propose en particulier la création de bibliothèques de fragments tri-nucleotidiques, qui nous sont essentielles pour la méthode de docking utilisée. Mais ces bibliothèques peuvent aussi être utilisées par d'autres personnes pour du design par exemple.

Le travail sur ces bibliothèques de fragments structuraux a soulevé le problème de l'inférence de prototypes selon des contraintes particulières. Il se trouve que ces objets ne possèdent pas de méthodes dédiées, nous avons donc répondu à ce besoin et proposé une solution qui est l'inférence d'ϵ-réseaux. Le chapitre 3 propose une application particulière aux classifications ascendantes hiérarchiques, ainsi qu'un nouveau linkage pour le calcul des ϵ-réseaux. Dans ce chapitre aussi, nous avons traité le problème de la manière la plus générale possible, puis nous l'avons adaptée pour calculer nos bibliothèques de fragments structuraux.

Enfin, la dernière contribution est centrée sur le docking d'ARN structurés, et appliquée à une structure particulière : les épingles à cheveux. Cette contribution présentée dans le chapitre 4 s'appuie sur l'introduction d'une nouvelle contrainte plus faible que les contraintes précédemment établies comme nécessaires au bon fonctionnement du docking par fragment : une contrainte de distance sur la fermeture de la boucle simple brin des épingles à cheveux au lieu de la connaissance de points d'ancrage sur la protéine. Sur un exemple choisi, nous avons donné une preuve de

concept que l'ajout de cette contrainte est suffisant pour assembler toutes les chaînes possibles lors du docking des boucles d'épingles à cheveux, si des conditions particulières sont atteintes à l'étape du docking des fragments. Nous donnons ci-après des pistes pour atteindre ces conditions.

Perspectives

Ce travail a ouvert de nombreuses perspectives dont certaines sont présentées ici. Une des perspectives très importantes sur le côté bio-informatique est la possibilité de considérer des fragments autres que les tri-nucléotides pour la méthode de docking par fragment. En effet, la nouvelle méthode d'inférence des prototypes permet d'avoir une cardinalité plus faible que la méthode utilisée précédemment. Cela remet en cause l'utilisation des tri-nucléotides qui historiquement étaient utilisés, car des tetra-nucléotides produisent un trop grand nombre de prototypes. La possibilité de couvrir l'espace conformationnel de manière plus efficace pourrait permettre l'utilisation de tetra-nucléotides.

Une autre perspective importante est l'application de la contribution à des structures de l'ARN autres que les épingles à cheveux. Cependant, il faudra tout d'abord régler en partie la question du docking des fragments avant de se lancer sur cet autre défi. En effet, les problèmes d'échantillonnages liés aux faibles contacts de certains fragments, ainsi que d'évaluation de ces fragments, sont très complexes. Une des principales raisons est sans doute que, dans les structures expérimentales, certains fragments insérés la chaîne d'ARN sont à une position sous-optimale sur la protéine par rapport à la position optimale qu'aurait ce fragment isolé. Cela est nécessairement le cas pour certains fragments dans une chaîne composée d'une seule base, donc de même séquence pour chaque fragment : un seul fragment peut être à la position optimale de cette séquence sur cette protéine.

De très nombreuses discussions sur des améliorations possibles du docking par fragments ont eu lieu au cours de la thèse. Actuellement nous utilisons la même fonction d'énergie pour les étapes d'échantillonnage et d'évaluation, pour des raisons pratiques. Une solution pourrait être d'optimiser la fonction d'énergie d'interaction pour l'échantillonnage d'un côté et d'optimiser une fonction de score de l'autre. Cela relâcherait la contrainte de dérivabilité de la fonction de score (liée à l'échantillonnage par descente de gradient), permettant un plus large choix de fonctions et de méthodes d'optimisation. Ce travail fait l'objet d'une autre thèse l'équipe.

Pour moi, une autre idée qui pourrait grandement améliorer la méthode serait de considérer

qu'un des fragments est à une position optimale (un hot-spot) et de reconstruire la chaîne à partir de ce fragment. À partir de cette idée, on pourrait régler soit le problème d'échantillonnage, soit le problème d'évaluation. Pour régler le problème d'échantillonnage, il serait possible de considérer itérativement chaque fragment comme à une position optimale, et de conserver un petit nombre de poses pour le fragment considéré, puis de réaliser un docking ancré à partir de ces poses. Pour cela, il faudrait réduire le nombre de poses pour ce fragment, avec par exemple des informations biologiques supplémentaires, ou une nouvelle fonction d'énergie dédiée aux hot-spots. Cette solution amène cependant des problèmes de temps de calcul du fait du grand nombre de docking ancrés que cela implique. Pour régler les problèmes d'évaluation, il serait possible d'utiliser l'hypothèse d'un hot-spot pour limiter le nombre de poses à considérer pour ce fragment, et donc construire un graphe beaucoup plus petit tout en gardant plus de poses pour les autres fragments. Cela nécessite néanmoins que l'échantillonnage ait produit au moins une bonne pose par fragment.

Les discussions sur le cadre d'application de notre méthode d'inférence d'ϵ-réseaux ont aussi été très nombreuses. Nous avons proposé un calcul des PPBE dans un RKHS, permettant d'appliquer la méthode à cet espace, mais nous avons aussi montré que dans un Hilbert le calcul de la PPBE se réduit à un problème d'optimisation convexe. Ainsi proposer une solution à ce problème permettrait d'étendre notre méthode aux Hilbert en général. D'autre part, il serait intéressant de définir les fonctions f_H permettant de donner un point de départ pour l'exploration du dendrogramme pour d'autres fonctions de linkage telles que Ward qui est aussi très utilisée dans la communauté des biologistes.

Enfin, l'application des coordonnées internes n'a pas pu être poussée à fond. L'idée d'utiliser le *Kernel Target Alignment* pour optimiser le produit scalaire défini entre les coordonnées internes au vu de la matrice de cRMSD est très intéressante. Dans le cas où l'approximation est bonne, cela autoriserait à comparer les fragments sans les superposer, ce qui gagnerait du temps de calcul, tout en étant plus précis que la superposition sur une structure moyenne.

Annexe A

Première annexe : les acides aminés

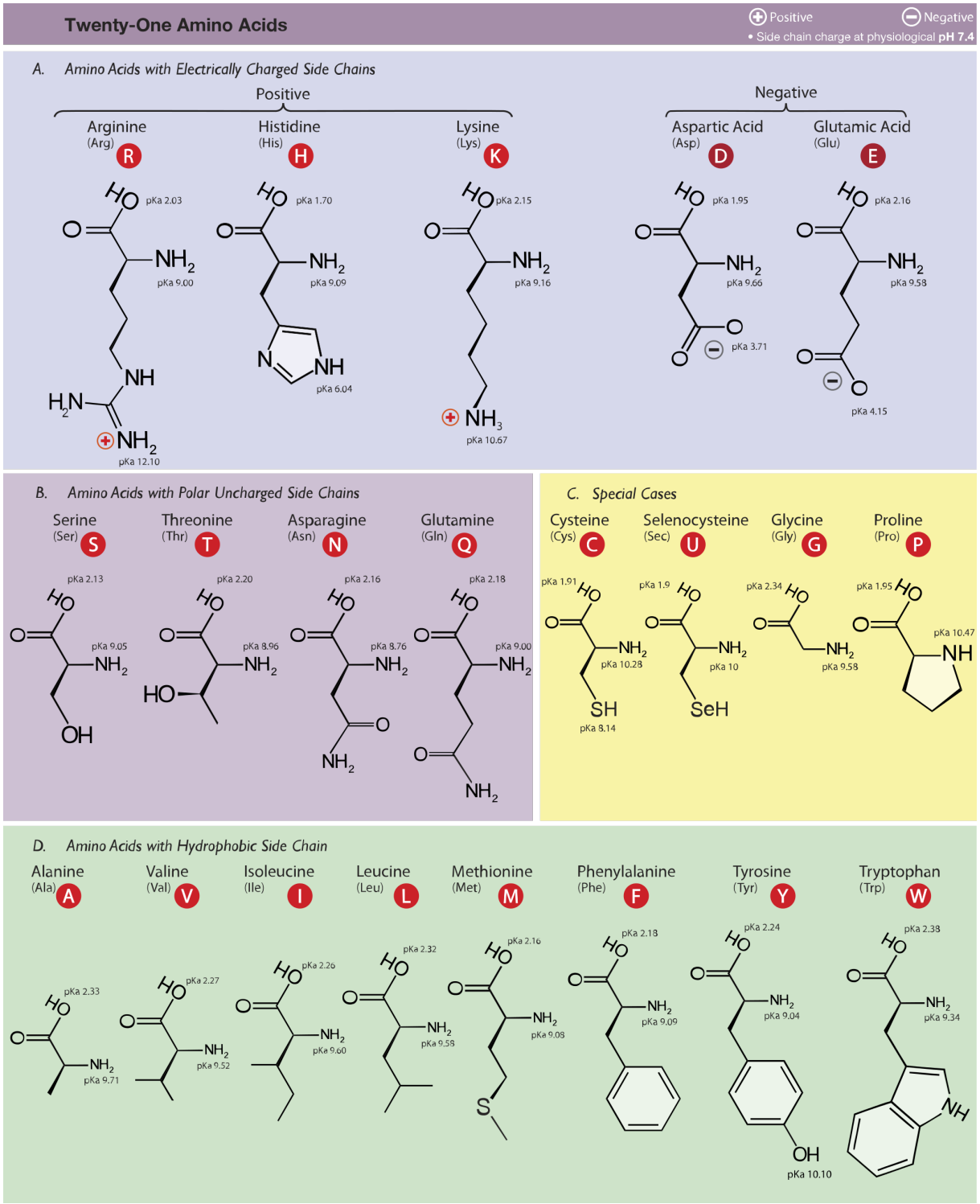


FIGURE A.1 – Les 21 acides aminés.

Annexe B

Deuxième annexe : optimisation du docking en général

Cette annexe présente une partie du travail que j'ai réalisé pendant ma thèse. Ce travail est plus lié à des observations sur les outils que j'ai utilisés pour réaliser les docking.

B.1 L'évaluation pour les tri-nucléotides

En biologie structurale, une hypothèse forte est qu'une structure d'un complexe obtenue expérimentalement correspond au minimum d'énergie d'interaction. La plupart des fonctions de score de docking approximent cette énergie d'interaction, c'est le cas de celle de ATTRACT. Dans le cadre du docking des fragments, l'hypothèse est juste pour la chaîne totale, mais n'est plus vraie si l'on regarde localement au niveau des fragments. Il est cependant possible que certains fragments interagissent très fortement avec le récepteur, et que l'on ait donc un hot-spot.

Dans le cas des fragments on pose donc une nouvelle hypothèse qui est que les fragments natifs sont à des positions sous-optimales. Nous essayons de développer une fonction d'énergie utilisable pour l'échantillonnage et l'évaluation qui serait capable de discriminer les différentes positions sous-optimales possibles. Lors de l'évaluation, et en prévision de l'assemblage, l'objectif est d'obtenir les positions sous-optimales de tous les fragments tout en réduisant au maximum le nombre de poses à garder pour les obtenir.

La fonction de score de ATTRACT est optimisée pour des structures complètes et n'est donc pas totalement optimale pour le docking des fragments. De plus, le set d'entraînement contient peu de complexe protéine/ARN simple brin [80], ce qui fait que les paramètres sont mal optimisés pour certaines interactions bases azotées/protéines.

J'ai étudié plus en détail les principaux fragments dont l'évaluation posait problème. Pour cela, j'ai créé un benchmark (grâce à protNAff) contenant les complexes dont une partie d'ARN simple brin est en contact avec la protéine, d'une longueur d'au moins 5 nucléotides. Cela représente environ 6500 tri-nucléotides.

La fonction d'énergie peut ne pas fonctionner pour deux raisons :

1. l'échantillonnage des fragments ne donne pas de bonne pose du fait de manques de contacts ;
2. l'évaluation ne donne pas un rang suffisamment faible pour certains fragments.

Pour distinguer chacun de ces deux problèmes, nous avons regardé les données plus en détail.

Pour distinguer les points (1) et (2), j'ai d'abord regardé le score des fragments natifs en utilisant la fonction de score de ATTRACT. J'ai remarqué que certains scores sont fortement positifs, ce qui indique un problème de paramètres.

Pour comprendre d'où viennent ces scores positifs, j'ai regardé les scores de tous les tri-nucléotides et j'ai gardé ceux qui ont un score positif. Il y a 450 tri-nucléotides avec un score positif (sur les 6500 du benchmark, soit environ 7%).

Ensuite j'ai identifié chaque nucléotide puis chaque pseudo-atome ayant un score positif. Pour chacun de ces pseudo-atomes j'ai calculé les scores d'interaction avec chacun des pseudo-atomes de la protéine avec qui ils sont en contact. On considère un contact si les deux pseudo-atomes sont à une distance de moins de 5 Å. J'ai ainsi identifié 1130 couples de pseudo-atomes ayant un score positif. Cela représente 169 assemblages biologiques de structures PDB.

Les scores allaient de valeurs très petites de l'ordre de 10^{-3} à des valeurs plus grandes de l'ordre de 10^4, et les distances de 1,14 Å à 5 Å (la limite maximum de contact, il est donc possible que certains couples à des distances plus élevées aient aussi un score positif).

Nous avons trouvé plusieurs possibilités pour expliquer ces scores positifs :

— les interpénétrations après ajout d'atomes manquants ;
— les artefacts de cristallographie ;
— les paramètres de la fonction de score de ATTRACT.

B.1.1 Problème d'interpénétration après ajout d'atomes manquants

Lors de l'ajout des atomes manquants d'un nucléotide dans la structure PDB par ProtNAff, on ne prenait d'abord pas en compte les structures à proximité. Il est donc possible d'ajouter un atome du nucléotide qui va interpénétrer un autre atome dans l'ARN ou la protéine. C'est le cas de 1Q2R où c'est toute une base azotée qu'il manque.

FIGURE B.1 – Représentation du problème dans 1Q2R. En vert, l'ARN du fichier PDB de référence, avec la base manquante. En rose, le fichier pdb produit par aareduce, avec la base ajoutée. En violet, la protéine. Les étoiles blanches sont les pseudo-atomes de l'ARN, et l'étoile orange le pseudo-atome de la protéine qu'il interpénètre.

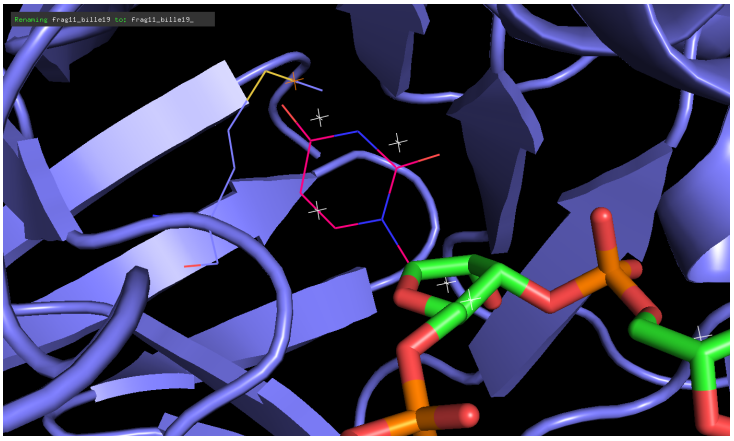
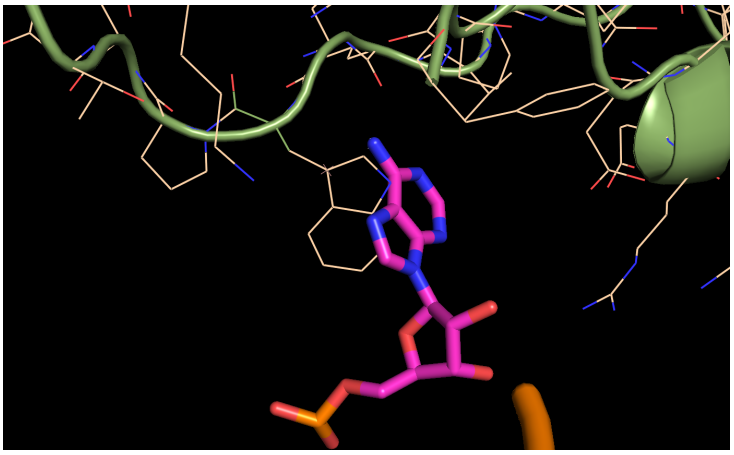


FIGURE B.2 – Représentation du problème dans 4B3O. En vert, la protéine (cristallographie), en beige après aareduce, une chaîne latérale a été ajoutée. En rose l'ARN.



Même si ce cas est plutôt rare, il survient plus régulièrement lors de l'ajout du Phosphate terminal, souvent manquant dans les chaînes d'ARN.

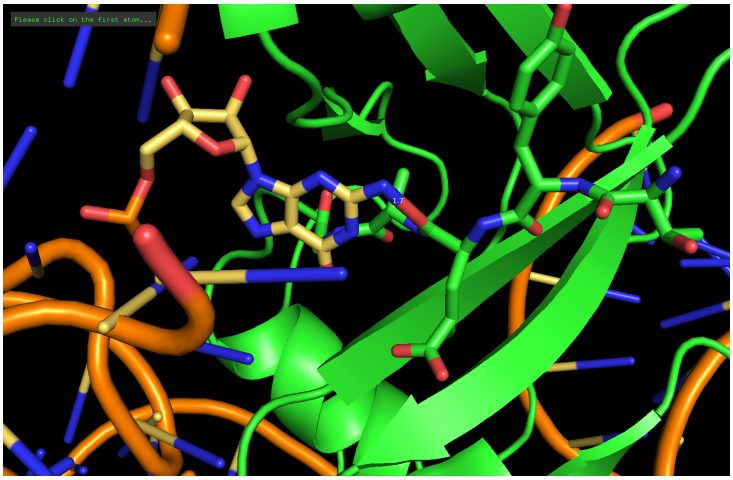
C'est aussi possible dans le cas de l'ajout d'une chaîne latérale de la protéine à l'aide du logiciel pdb2pqr.

Une solution a été proposée pour ce problème, en ajoutant une étape qe calcul des interpénétrations lors de l'ajout des atomes manquants par `aareduce.py` d'ATTRACT.

B.1.2 Artefact de cristallographie

Dans l'ensemble du benchmark nous n'avons repéré qu'un seul cas de ce type, où deux atomes non liés sont à moins de 2 Å dans la structure cristallographique.

FIGURE B.3 – Représentation du problème dans 2CZJ. En jaune l'ARN, en vert la protéine, la liaison au centre marquée de la distance 1.7 Å ne devrait pas exister.



Ce problème est probablement dû à la résolution de la structure qui est de 3.01 Å.

B.1.3 Problème d'évaluation lié aux paramètres ATTRACT

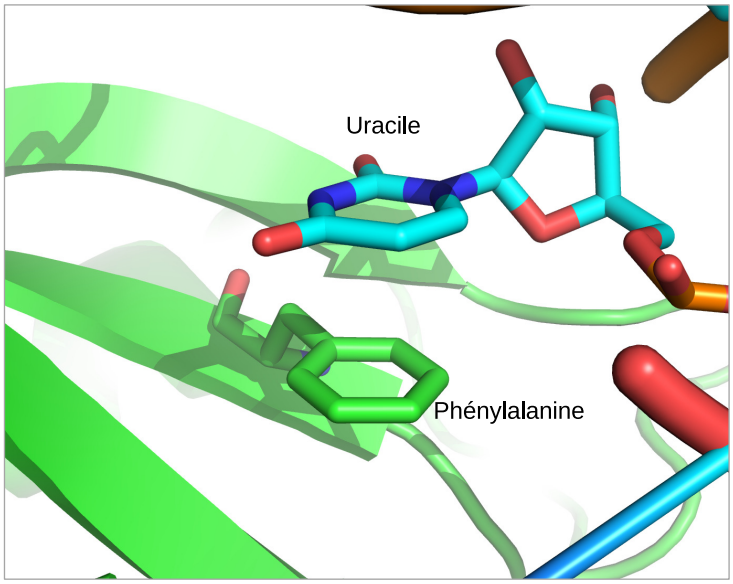
Certaines interactions de pseudo-atomes donnent souvent des scores positifs dans les structures natives. Nous avons recensé 3 cas :

— l'interaction Tryptophane (chaîne latérale) - Cystéine (base azotée) ;
— l'interaction Tryptophane (chaîne latérale) - Uracile (base azotée) ;
— l'interaction Phénylalanine (chaîne latérale) - Guanine (base azotée).

Ces 3 interactions sont des interactions de types empilements, ces interactions sont particulièrement fortes et impliquent 2 cycles aromatiques qui se stabilisent par échange d'électrons. Elles sont courantes au sein de l'ARN où elles stabilisent les hélices, et représentent une interaction forte avec les protéines. Cependant, lors de l'apprentissage de la fonction de score, la majorité des structures d'ARN étaient double brins, nous supposons que c'est pour cela que l'apprentissage a mal fonctionné pour ces interactions précises.

Ce problème est complexe à résoudre, car il implique de revoir les paramètres de la fonction d'énergie de ATTRACT. Actuellement c'est la même fonction d'énergie qui est utilisée pour l'échantillonnage et pour l'évaluation, cela pourrait changer pour avoir une fonction de score différente. C'est ce sur quoi travaille une doctorante de l'équipe, elle essaye de créer et d'optimiser une fonction de score qui améliorerait les docking et enlèverait ces erreurs de scores positifs.

FIGURE B.4 – Représentation d'une interaction d'empilement. En bleu l'ARN et en vert la protéine. Les deux cycles sont parallèles ce qui permet le transfert d'électrons entre les couches π.



B.2 Prise en compte des empilements dans ATTRACT

Pour l'empilement, des solutions à base de calcul d'angles dièdres sont possibles au moins entre les différentes bases azotées. En effet, comme elles possèdent au moins 3 pseudo-atomes nous conservons la planarité des bases. Cependant, pour les cycles aromatiques des chaînes latérales des protéines, ce n'est pas le cas. Ce sont deux pseudo-atomes qui représentent les cycles aromatiques. Cependant, à partir de deux points il serait toujours possible de définir un empilement en calculant l'angle entre cette droite (représenté par 2 points) et le plan (représenté par les 3 points des bases azotées), sans en prendre en compte la rotation possible du cycle selon la droite. Il serait aussi possible de repasser en tout-atome pour effectuer une courte minimisation qui prendrait en compte les empilements. Une autre solution proposée est d'ajouter des billes fantômes au niveau des cycles aromatiques des chaînes latérales des protéines. Ces billes fantômes serviraient à définir l'orientation du plan défini par le cycle aromatique.

Cette question des empilements a été maintes fois discutée au cours de ma thèse, elle joue un rôle important dans les interactions ARN-protéines, et plus particulièrement dans le cas des RRM, structures qui ont été utilisées pour définir l'approche du docking ancré.

Annexe C

Troisième annexe : Bibliothèque de fragments non spécifiques

L'idée de départ est de réaliser le docking d'un seul type de fragment et de réaliser l'assemblage à partir d'un unique docking (comme si l'ARN est un poly-A par exemple).

Nous avons créé une bibliothèque de tri-nucléotides non spécifiques en termes de séquence. Pour cela, nous avons simplifié le gros-grain des bases azotées pour ne conserver que 3 pseudo-atomes (de manière à conserver l'information sur l'orientation de la base).

Le clustering est appliqué sur les centres des clusters à 1 Å des tri-nucléotides normaux, ce clustering est réalisé pour 1 Å. Le résultat est un ensemble de 1002 représentants.

Ces structures sont intéressantes pour regarder les conformations possibles adoptées, quelle que soit la séquence. Cette bibliothèque pourrait être utilisée pour complémenter les résultats obtenus sur la spécificité des séquences ainsi que les conformations due aux contacts avec la protéine.

Il n'y a pas eu d'applications directes de cette bibliothèque, mais si elle est faite proprement, c.-à-d., que le clustering est fait sur l'ensemble des fragments provenant de la PDB, on pourrait réaliser le docking d'une unique bibliothèque. Cela réduirait grandement le temps de calcul des docking, permettant de passer plus de temps sur l'assemblage.

Glossaire

Ce glossaire reprend les abréviations présentes dans la thèse.

ADN : *Acide DésoxyriboNucléique*, p. 6

AN : *Acide Nucléique*, p.6

ARN : *Acide RiboNucléique*, p.5

ARNlnc : *ARN long non-codant*, p.6

ARNm : *ARN messager*, p.6

ARNr : *ARN ribosomique*, p.6

ARNt : *ARN de transfert*, p.6

CAH : *Classification Ascendante Hiérarchique*, p. 69

CAPRI : *Critical Assessment of PRediction of Interactions*, p. 27

CASP : *Critical Assessment of Structure Prediction*, p. 22

CNN : *Convolutional Neural Network*, p.22

cRMSD : *RMSD conformationnel*, p.21

Cryo-ME : *Cryo-Microscopie Électronique*, p. 20

FBDD : *Fragment-Based Drug Design*, p.34

FFT : *Fast Fourier Transform*, p.25

Fnat : *fraction de contact natif*, p.28

iRMSD : *RMSD de l'interface*, p.28

KT : *conditions de Kuhn-Tucker*, p.80

KTA : *Kernal Target Alignment*, p.94

LRMSD : *RMSD du ligand*, p.28

miARN : *micro-ARN long non-codant*, p.6

PDB : *Protein Data Bank*, p.12

PPBE : *Plus Petite Boule Englobante*, p.68

R : *purine*, p.51

RBPs : *RNA-Binding Proteins*, p.6

RKHS : *Reproducing Kernel Hilbert Space*, p. 78

RMN : *Résonance Magnétique Nucléaire*, p. 20

RMSD : *Root Mean Squarred Deviation*, p.21

RRM : *RNA Recognition Motif*, p.14

SAXS : *Small X-Ray Scattering*, p.31

Y : *pyrimidine*, p.51

Bibliographie

[1] H. Alawieh, N. Wicker, and C. Biernacki. Projection under pairwise distance control. In *Communications in Statistics - Theory and Methods*, pages 1–29, 2020.

[2] M. Antczak, T. Zok, M. Osowiecki, M. Popenda, R.W. Adamiak, and M. Szachniuk. RNA-fitme : a webserver for modeling nucleobase and nucleoside residue conformation in fixed-backbone RNA structures. *BMC Bioinformatics*, 19, 2018.

[3] L. Becquey, E. Angel, and F. Tahi. RNANet : an automatically built dual-source dataset integrating homologous sequences and RNA structures. *Bioinformatics*, 37 :1218 – 1224, 2021.

[4] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, Boston, 2004.

[5] D. Bhattacharya, B. Adhikari, J. Li, and J. Cheng. FRAGSION : ultra-fast protein fragment library generation by IOHMM sampling. *Bioinformatics*, 32(13), 2016.

[6] J. Bien and R. Tibshirani. Hierarchical clustering with prototypes via minimax linkage. *Journal of the American Statistical Association*, 106(495) :1075–1084, 2011.

[7] R. S. Bohacek, C. McMartin, and W. C. Guida. The art and practice of structure-based drug design : a molecular modeling perspective. *Med Res Rev.*, 16(1) :3–50, 1996.

[8] P. Bryant, G. Pozzati, and A. Elofsson. Improved prediction of protein-protein interactions using AlphaFold2. *Nature Communications*, 13(1265), 2022.

[9] P. Clote, Y. Ponty, and JM. Steyaert. Expected distance between terminal nucleotides of RNA secondary structures. *Journal of Mathematical Biology*, 65(3) :581–599, 2011.

[10] N. Cristianini, J. Kandola, A. Elisseeff, and J. Shawe-Taylor. *On Kernel Target Alignment*, pages 205–256. Springer Berlin Heidelberg, 2006.

[11] J.A. Cruz, M.F. Blanchet, M. Boniecki, J.M. Bujnicki, S.J. Chen, S. Cao, R. Das, F. Ding, N.V. Dokholyan, S.C. Flores, L. Huang, C.A. Lavender, V. Lisi, F. Major, K. Mikolajczak,

D.J. Patel, A. Philips, T. Puton, J. Santalucia, F. Sijenyi, T. Hermann, K. Rother, M. Rother, A. Serganov, M. Skorupski, T. Soltysinski, P. Sripakdeevong, I. Tuszynska, K.M. Week, C. Waldsich, M. Wildauer, N.B. Leontis, and E. Westhof. RNA-puzzles : a CASP-like evaluation of RNA three-dimensional structure prediction. *RNA*, 18 :610–625, 2012.

[12] R. Das, I. André, Y. Shen, Y. Wu, A. Lemak, S. Bansal, C.H. Arrowsmith, T. Szyperski, and D. Baker. Simultaneous prediction of protein folding and docking at high resolution. *Proceedings of the National Academy of Sciences of the USA*, 106 :18978–18983, 2009.

[13] R. Das and D. Baker. Automated de novo prediction of native-like RNA tertiary structures. *Proceedings of the National Academy of Sciences of the USA*, 104 :14664–14669, 2007.

[14] R. Das, J. Karanicolas, and D. Baker. Atomic accuracy in predicting and designing noncanonical RNA structure. *Nature Metjods*, 7 :291–294, 2010.

[15] I. Chauvot de Beauchêne, S. de Vries, and M. Zacharias. Fragment-based modelling of single stranded RNA bound to RNA recognition motif containing proteins. *Nucleic Acids Research*, 44(10) :4565–4580, 2016.

[16] S. de Vries, C. Schindler, I. Chauvot de Beauchêne, and M. Zacharias. A web interface for easy flexible protein-protein docking with ATTRACT. *Biophysical journal*, 3 :462–465, 2015.

[17] S. de Vries and M. Zacharias. ATTRACT-EM : A new method for the computational assembly of large molecular machines using cryo-EM maps. *PLoS ONE*, 7, 2012.

[18] A. Delannoy, A. Moniot, Y. Guermeur, and I. Chauvot de Beauchêne. Feature extraction for the clustering of small 3D structures : application to RNA fragments. In *JOBIM'21*, 2021.

[19] E.J. Denning, U Deva Priyakumar, L. Nilsson, and A.D. Mackerell Jr. Impact of 2'-hydroxyl sampling on the conformational properties of RNA : Update of the CHARMM all-atom additive force field for RNA. *Journal of Computational Chemistry*, 32 :1929–1943, 2011.

[20] T.J. Dolinsky, P. Czodrowski, H. Li, J.E. Nielsen, J.H. Jensen, G. Klebe, and N.A. Baker. PDB2PQR : expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Research*, 35 :522–525, 2007.

[21] R.L. Dunbrack. Rotamer libraries in the 21st century. *Current Opinion in Structural Biology*, 12(4) :431–440, 2002.

[22] A. Fernández, R. Salthú, and H. Cendra. Discretized torsional dynamics and the folding of an RNA chain. *Phys. Rev. E*, 60 :2105–2119, 1999.

[23] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2) :95–110, 1956.

[24] D. Gautheret, F. Major, and R. Cedergren. Modeling the three-dimensional structure of RNA using discrete nucleotide conformational sets. *Journal of Molecular Biology*, 229(4) :1049–1064, 1993.

[25] S. Gerstberger, M. Hafner, and T. Tuschl. A census of human RNA-binding proteins. *Nature Reviews Genetics*, 15 :829–845, 2014.

[26] R. Gonzalez-Aleman, N. Chevrollier, M. Simoes, L. Montero-Cabrera, and F. Leclerc. MCSS-based predictions of binding mode and selectivity of nucleotide ligands. *Journal of Chemical Theory and Computation*, 17 :2599–2618, 2021.

[27] John C. Gower and Gavin J. S. Ross. Minimum spanning trees and single linkage cluster analysis. *Journal of The Royal Statistical Society Series C-applied Statistics*, 18 :54–64, 1969.

[28] J.J. Gray, S. Moughon, C. Wang, O. Schueler-Furman, B. Kuhlman, C.A. Rohl, and D. Baker. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *Journal of Molecular Biology*, 331 :281–299, 2003.

[29] D. Hall, S. Li, K. Yamashita, R. Azuma, J.A. Carver, and D.M. Standley. RNA-LIM : a novel procedure for analyzing protein/single-stranded RNA propensity data with concomitant estimation of interface structure. *Analytical Biochemistry*, 472 :52–61, 2015.

[30] K.F. Han and D. Baker. Recurring local sequence motifs in proteins. *Journal of Molecular Biology*, 251 :176–187, 1995.

[31] K. Hanssens, F. Brenet, J. Agopian, S. Georgin-Lavialle, G. Damaj, L. Cabaret, M.O. Chandesris, P. de Sepulveda, O. Hermine, P. Dubreuil, and E. Soucie. SRSF2-p95 hotspot mutation is highly associated with advanced forms of mastocytosis and mutations in epigenetic regulator genes. *Haematologica*, 99 :830–835, 2014.

[32] C. R. Harris, K. Jarrod Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. Fernández del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585(7825) :357–362, September 2020.

[33] J. Hennig, S.J. de Vries, K.DM. Hennig, L. Randles, K.J. Walters, M. Sunnerhagen, and A.MJJ. Bonvin. MTMDAT-HADDOCK : High-throughput, protein complex structure mo-

deling based on limited proteolysis and mass spectrometry. *BMC Structural Biology*, 12, 2012.

[34] X. Huang, Q. Lü, D. Miao, and P. Qian. Assessment of rotamer library for protein structure prediction. In *The 2nd International Conference on Information Science and Engineering*, pages 152–155, 2010.

[35] X. Huang, R. Pearce, and Y. Zhang. EvoEF2 : accurate and fast energy function for computational protein design. *Bioinformatics*, 36(4) :1135–1142, 2019.

[36] R.G. Huber, M.A. Margreiter, J.E. Fuchs, S. von Grafenstein, C.S. Tautermann, K.R. Liedl, and T. Fox. Heteroaromatic pi-stacking energy landscapes. *Journal of chemical information and modeling*, 54 :1371–1379, 2014.

[37] Chauvot de Beauchene I, S. de Vries, and M. Zacharias. Binding site identification and flexible docking of single stranded RNA to proteins using a fragment-based approach. *PLoS Computational Biology*, 12(1), 2016.

[38] W. P. Jencks. On the attribution and additivity of binding energies. *Proc Natl Acad Sci U S A*, 78(7) :4046–4050, 1981.

[39] C. Johansson, L.D. Finger, L. Trantirek, T.D. Mueller, S. Kim, I.A. Laird-Offringa, and J. Feigon. Solution structure of the complex formed by the two N-terminal RNA-binding domains of nucleolin and a pre-rRNA target. *Journal of Molecular Biology*, 337(4) :799–816, 2004.

[40] C.P. Jones and A.R. Ferré-D'Amaré. RNA quaternary structure and global symmetry. *Trends Biochem Sci.*, 40 :211–220, 2015.

[41] S. Jones. Protein–RNA interactions : structural biology and computational modeling techniques. *Biophysical Reviews*, 8(4) :359–367, 2016.

[42] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. Kohl, A. Ballard, A. Cowie, B.R. Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, D. Silver, O. Vinyals, A. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. Applying and improving alphafold at CASP14. *Proteins*, 89 :1711–1721, 2021.

[43] H.W.E. Jung. Über die kleinste kugel, die eine räumliche figur einschliesst. *J. Reine Angew. Math.*, 123 :241–257, 1901.

[44] Y. Karami, F. Guyon, S.J. de Vries, and P. Tufféry. DaReUS-Loop : accurate loop modeling using fragments from remote or unrelated proteins. *Scientic Reports*, 8, 2018.

[45] A.N. Kolmogorov and V.M. Tihomirov. ϵ-entropy and ϵ-capacity of sets in functional spaces. *American Mathematical Society Translations, series 2*, 17 :277–364, 1961.

[46] B. Kopp. Hierarchical classification III : Average-linkage, median, centroid, WARD, flexible strategy. *Biometrical Journal*, 20 :703–711, 1978.

[47] M.F. Lensink, N. Nadzirin, S. Velankar, and S.J. Wodak. Special issue : Critical assessment of predicted interactions (CAPRI). *Proteins : Structure, Function, and Bioinformatics*, 88(8) :911–1129, 2020.

[48] A.M. Lesk. Casp2 : report on ab initio predictions. *Proteins*, Suppl 1 :151–166, 1997.

[49] J.S. Li, R. Potru, and F. Shahrokhi. A performance study of some approximation algorithms for computing a small dominating set in a graph. *Algorithms*, 13 :339, 2020.

[50] 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*, 6, 2011.

[51] X.J. Lu and W.K. Olson. 3DNA : a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Research*, 31 :5108–5121, 2003.

[52] S. Sarkhel M. Egli. Lone pair-aromatic interactions : To stabilize or not to stabilize. *Accounts of Chemical Research*, 40 :197–205, 2007.

[53] M.M. Maljković, N.S. Mitić, and A.G. de Brevern. Prediction of structural alphabet protein blocks using data mining. *Biochimie*, 197 :74–85, 2022.

[54] V. Mallet, M. Nilges, and G. Bouvier. quicksom : Self-organizing maps on GPUs for clustering of molecular dynamics trajectories. *Bioinformatics*, 37 :2064–2065, 2021.

[55] C. Maris and F. H.-T. Allain C. Dominguez. The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *The FEBS Journal*, 272 :2118–2131, 2005.

[56] M. Marz and P.F. Stadler. *RNA Interactions*, pages 20–38. Springer New York, 2011.

[57] D. Mias-Lucquin and I. Chauvot de Beauchene. Conformational variability in proteins bound to single-stranded DNA : A new benchmark for new docking perspectives. *Proteins*, 2021.

[58] K. Moldave. RNA and protein synthesis. Academic Press, 1981.

[59] A. Moniot, Y. Guermeur, S. de Vries, and I. Chauvot de Beauchene. ProtNAff : Protein-bound nucleic acid filters and fragment libraries. *Bioinformatics*, 38 :3911–3917, 2022.

[60] A. Moniot, R. Roy, Y. Guermeur, and I. Chauvot de Beauchêne. Docking of RNA Hairpin on Protein Using a Fragment-Based Method. In *JOBIM 2020 - Journées Ouvertes en Biologie, Informatique et Mathématiques*, June 2020.

[61] B. Mordukhovich, M. Nguyen, and C. Villalobos. The smallest enclosing ball problem and the smallest intersecting ball problem : Existence and uniqueness of solutions. *Optimization Letters*, 7, 2011.

[62] C. E. Nakamura and R. H. Abeles. Mode of interaction of beta-hydroxy-beta-methylglutaryl coenzyme A reductase with strong binding inhibitors : compactin and related compounds. *Biochemistry*, 24(6) :1364–1376, 1985.

[63] N.M. Nam, N.T. An, and J. Salinas. Applications of convex analysis to the smallest intersecting ball problem. *Journal of Convex Analysis*, 19, 2011.

[64] V. Nguen-Khac and K. Nguen-Khac. An infinite-dimensional generalization of the jung theorem. *Math Notes*, 80(2) :224–232, 2006.

[65] W.K. Olson, S. Li, T. Kaukonen, A.V. Colasanti, Y. Xin, and X.-J. Lu. Effects of noncanonical base pairing on RNA folding : Structural context and spatial arrangements of G·A pairs. *Biochemistry*, 58(20) :2474–2487, 2019.

[66] S. Páll, A. Zhmurov, P. Bauer, M. Abraham, M. Lundborg, A. Gray, B. Hess, and E. Lindahl. Heterogeneous parallelization and acceleration of molecular dynamics simulations in GROMACS. *J. Chem. Phys.*, 153, 2020.

[67] M. Parisien and F. Major. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, 452 :51–55, 2008.

[68] D.A. Pearlman, D.A. Case, J.W. Caldwell, W.S. Ross, T.E. Cheatham, S. DeBolt, D. Ferguson, G. Seibel, and P. Kollman. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications*, 91(1) :1–41, 1995.

[69] B. Pereira, M. Billaud, and R. Almeida. RNA-binding proteins in cancer : Old players and new actors. *Trends in Cancer*, 3(7) :506–528, 2017.

[70] X. Periole and S.J. Marrink. The martini coarse-grained force field. *Methods in molecular biology*, 924 :533–565, 2013.

[71] 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–e112, 2012.

[72] M. Popenda, M. Szachniuk, M. Blazewicz, S. Wasik, E.K. Burke, J. Blazewicz, and R.W. Adamiak. RNA FRABASE 2.0 : an advanced web-accessible database with the capacity to search the three-dimensional fragments within RNA structures. *BMC Bioinformatics*, 11, 2010.

[73] P.D. Renfrew, T.W. Craven, G.L. Butterfoss, K. Kirshenbaum, and R. Bonneau. A rotamer library to enable modeling and design of peptoid foldamers. *Journal of the American Chemical Society*, 136(24) :8772–8782, 2014.

[74] K.E. Richardson, C.C. Kirkpatrick, and B.M. Znosko. RNA CoSSMos 2.0 : an improved searchable database of secondary structure motifs in rna three-dimensional structures. *Database (Oxford)*, 2020.

[75] C.A. Rohl, C.E.M. Strauss, D. Chivian, and D. Baker. Modeling structurally variable regions in homologous proteins with rosetta. *Proteins*, 55 :656–677, 2004.

[76] C.A. Rohl, C.E.M. Strauss, K.M.S. Misura, and D. Baker. Protein structure prediction using Rosetta. In *Numerical Computer Methods, Part D*, volume 383 of *Methods in Enzymology*, pages 66–93. Academic Press, 2004.

[77] M. Rother, K. Rother, T. Puton, and J.M. Bujnicki. ModeRNA : a tool for comparative modeling of RNA 3d structure. *Nucleic Acids Research*, 39(10) :4007–4022, 2011.

[78] J. M. Sagendorf, N. Markarian, H. M. Berman, and R. Rohs. DNAproDB : an expanded database and web-based tool for structural analysis of DNA–protein complexes. *Nucleic Acids Research*, 48(1), 2020.

[79] C.E.M. Schindler, S. de Vries, A. Sasse, and M. Zacharias. SAXS data alone can generate high-quality models of protein-protein complexes. *Structure*, 24(8) :1387–1397, 2016.

[80] P. Setny and M. Zacharias. A coarse-grained force field for Protein-RNA docking. *Nucleic Acids Research*, 39(21) :9118–9129, 2011.

[81] M.P. Thelen and M.J. Kye. The role of RNA binding proteins for local mRNA translation : Implications in neurological disorders. *Frontiers in Molecular Biosciences*, 6(161), 2020.

[82] R.J.L. Townshend, S. Eismann, A.M. Watkins, R. Rangan, M. Karelina, R. Das, and R.O. Dror. Geometric deep learning of RNA structure. *Science*, 373 :1047–1051, 2021.

[83] G.C.P van Zundert, J.P.G.L.M. Rodrigues, M. Trellet, C. Schmitz, P.L. Kastritis, E. Karaca, A.S.J. Melquiond, M. van Dijk, S.J. de Vries, and A.M.J.J. Bonvin. The HADDOCK2.2 webserver : User-friendly integrative modeling of biomolecular complexes. *Journal of Molecular Biology*, 428 :720–725, 2016.

[84] J. Černý, P. Božíková, J. Svoboda, and B. Schneider. A unified dinucleotide alphabet describing both RNA and DNA structures. *Nucleic Acids Research*, 48(11), 2020.

[85] L. Veselý. For a dense set of equivalent norms, a non-reflexive banach space contains a triangle with no chebyshev center. *Comment.Math.Univ.Carolin*, 42(1) :153–158, 2001.

[86] R.R. Walia, C. Caragea, B.A. Lewis, F. Towfic, M. Terribilini, Y. El-Manzalawy, D. Dobbs, and V. Honavar. Protein-RNA interface residue prediction using machine learning : an assessment of the state of the art. *BMC Bioinformatics*, 13, 2012.

[87] L. Wang, Y. Nam, A.K. Lee, C. Yu, K. Roth, C. Chen, E.M. Ransey, and P. Sliz. LIN28 zinc knuckle domain is required and sufficient to induce let-7 oligouridylation. *Cell Reports*, 18(11) :2664–2675, 2017.

[88] X. Wang, J. McLachlan, P.D. Zamore, and T.M. Hall. Modular recognition of RNA by a human pumilio-homology domain. *Cell*, 110 :501–512, 2002.

[89] A.M. Watkins, R. Rangan, and R. Das. Farfar2 : Improved de novo rosetta prediction of complex global RNA folds. *Strcuture*, 28 :963–976, 2020.

[90] J.D. Watson and F. Crick. A structure for deoxyribose nucleic acid. *Nature*, 171 :737–738, 1953.

[91] J. Xu. Rapid protein side-chain packing via tree decomposition. In *Research in Computational Molecular Biology*, pages 423–439. Springer Berlin Heidelberg, 2005.

[92] J. Zhang, Y. Bian, H. Lin, and W. Wang. RNA fragment modeling with a nucleobase discrete-state model. *Phys. Rev. E*, 85, 2012.

[93] Y. Zhang. I-TASSER server for protein 3d structure prediction. *BMC Bioinformatics*, 9, 2008.

[94] Q. Zhao, Z. Zhao, X. Fan, Z. Yuan, Q. Mao, and Y. Yao. Review of machine learning methods for RNA secondary structure prediction. *PLoS computational biology*, 17(8), 2021.

[95] G. Zheng, A. V. Colasanti, X.J. Lu, and W. K. Olson. 3DNALandscapes : a database for exploring the conformational features of DNA. *Nucleic Acids Research*, 38(1), 2010.

[96] T. Zok, M. Antczak, M. Riedel, D. Nebel, T. Villmann, P. Lukasiak, J. Blazewicz, and M. Szachniuk. Building the library of RNA 3D nuclotide conformations using the clustering approach. *International Journal of Applied Mathematics and Computer Science*, 25 :689–700, 2015.

Résumé : La caractérisation des complexes ARN-protéine à l'échelle atomique nous permet de mieux comprendre les fonctions de ces complexes, et de définir des cibles thérapeutiques pour réguler les phénomènes biologiques auxquels ils participent. L'objet de cette thèse est de développer des outils permettant de prédire la structure d'un complexe protéine-ARN lorsque l'on connaît une structure 3D de la protéine ainsi que la structure secondaire de la partie d'ARN en interaction. Nous nous concentrons sur le cas où l'ARN est principalement sous forme simple brin (nucléotides non appariés), posant la difficulté de sa flexibilité.
Une méthode de *docking* développée dans l'équipe CAPSID repose sur l'utilisation de fragments structuraux d'ARN simple brin. Le travail de cette thèse s'est appuyé sur cette méthode pour réaliser le docking de structures secondaires de l'ARN. Nous avons d'abord évalué l'apport d'une contrainte de fermeture de boucle pour le docking de la boucle simple brin d'une structure en épingle, puis abordé le docking des éléments double brin de ces structures, ouvrant la voie à l'assemblage du complexe entier.
Cette méthode de docking est dépendante de l'utilisation de bibliothèques de fragments structuraux. Ces bibliothèques sont composées de prototypes qui représentent le paysage conformationnel observé expérimentalement dans les structures d'ARN liés à des protéines. Une large partie du travail de thèse a consisté en la création et l'optimisation de telles bibliothèques de fragments.
Nous avons créé l'outil ProtNAff qui permet d'extraire de la PDB des sous-ensembles de structures et de créer des bibliothèques de fragments d'acides nucléiques, suivant des combinaisons complexes de critères. Il a été conçu de façon à dépasser nos besoins, afin d'être adopté par la communauté pour le traitement de problèmes variés.
Nous avons développé une nouvelle approche pour l'inférence de prototypes représentatifs d'un ensemble de conformations. L'ensemble de prototypes doit satisfaire deux contraintes contradictoires : être représentatif (au sens de la métrique) et de cardinalité aussi petite que possible. Le problème se réduit donc à celui de l'inférence d'un epsilon-réseau de cardinalité minimale. Nous le traitons dans toute sa généralité en discutant des ensembles sur lesquels sont définies les données. Notre méthode se base sur la classification ascendante hiérarchique avec comme linkage le rayon des plus petites boules englobant les points de chaque sous-ensemble. Appliquée à nos bibliothèques, cette approche a permis de réduire d'un facteur 4 leur taille, et d'autant nos temps de calcul de docking, tout en améliorant leur fiabilité.
Enfin, pour pallier le problème posé par les superpositions de structures deux à deux, nous avons utilisé une représentation des fragments en coordonnées internes permettant de réduire encore les temps de calcul de création des bibliothèques.

Mots clés : Bio-informatique structurale, Epsilon-réseaux, Combinatoire

Abstract : The characterization of RNA-protein complexes at the atomic scale allows us to better understand the biological functions of these complexes, and to define therapeutic targets to regulate the biological phenomena in which they participate. The aim of this thesis is to develop tools to predict the structure of a protein-RNA complex when a 3D structure of the protein is known as well as the secondary structure of the interacting RNA part. We focus on the case where RNA is mainly in single-stranded form (unpaired nucleotides), raising the difficulty of its flexibility.
A docking method developed in the CAPSID team is based on the use of structural fragments of single-stranded RNA. The work of this thesis builds on this method to perform docking of RNA secondary structures. We first evaluated the contribution of a loop closure constraint for docking the single-stranded loop of a hairpin structure, and then addressed the docking of the double-stranded elements of these structures, paving the way for the assembly of the entire complex.
This fragment-based docking method is dependent on the use of structural fragment libraries. These libraries are composed of prototypes that represent the conformational landscape experimentally observed in protein-bound RNA structures. A large part of the thesis work consisted in the creation and optimization of such fragment libraries.
We created the ProtNAff tool that allows to extract subsets of structures from the PDB and to create libraries of nucleic acid fragments, following complex combinations of criteria. It has been designed to exceed our needs, so that it can be adopted by the community for the treatment of various problems.
We have developed a new approach for inferring prototypes of a set of conformations. The set of prototypes must satisfy two contradictory constraints : to be representative (in the sense of the metric) and of cardinality as small as possible. The problem thus reduces to that of inferring an epsilon-network of minimal cardinality. We treat it in all its generality by discussing the spaces on which the data are defined. Our method is based on hierarchical agglomerative classification with as linkage the radius of the minimum balls enclosing the points of each subset. Applied to our libraries, this approach reduced their size by a factor of 4, and our docking computation time by the same amount, while improving their reliability.
Finally, to overcome the problem posed by the pairwise superimposition of structures, we used a representation of the fragments in internal coordinates, allowing to reduce further the computation time for the creation of libraries.

Keywords : Structural bio-informatics, Epsilon-nets, Combinatorics