



**HAL**  
open science

# Développement et application d'une approche de docking par fragments pour modéliser les interactions entre protéines et ARN simple-brin

Nicolas Chevrollier

► **To cite this version:**

Nicolas Chevrollier. Développement et application d'une approche de docking par fragments pour modéliser les interactions entre protéines et ARN simple-brin. Biologie structurale [q-bio.BM]. Université Paris-Saclay, 2019. Français. NNT : 2019SACLS106 . tel-02436914

**HAL Id: tel-02436914**

**<https://theses.hal.science/tel-02436914>**

Submitted on 13 Jan 2020

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

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

# Développement et application d'une approche de docking par fragments pour modéliser les interactions entre protéines et ARN simple-brin

Thèse de doctorat de l'Université Paris-Saclay  
Préparée à l'Université Paris-Sud

École doctorale n°577 : Structure et Dynamique des Systèmes  
Vivants  
Spécialité de doctorat: Sciences de la vie et de la santé

Thèse présentée et soutenue à Orsay, le 9 mai 2019, par

**Nicolas Chevrollier**

## Composition du Jury :

M. Stéphane Bressanelli Directeur de Recherche, CNRS (- UMR 9198, I2BC)	Président
Mme Samuela Pasquali Professeur, Université Paris-Descartes (- UMR 8038-CNRS, CiTCoM)	Rapporteur
Mme Annick Dejaegere Professeur, Ecole Supérieure de Biotechnologie de Strasbourg (- UMR 7104, IGBMC)	Rapporteur
Mme Isaure Chauvot de Beauchêne Chargée de Recherche, CNRS (- UMR 7503, LORIA)	Examinatrice
M. Alexandre de Brevern Directeur de Recherche, INSERM, Université Paris-Diderot (- UMR_S 1134)	Examineur
M. Fabrice Leclerc Chargé de Recherche, CNRS (- UMR 9198, I2BC)	Directeur de thèse



# Table des matières

<b>Introduction.....</b>	<b>5</b>
<b>I Importance fonctionnelle des protéines de liaison à l'ARN.....</b>	<b>6</b>
1 Maturation de l'ARN pré-messager précurseur.....	7
1.1 La famille des protéines nucléaires hétérogènes.....	7
1.2 L'épissage.....	8
1.3 La polyadénylation.....	8
1.4 L'édition.....	9
2 Export de l'ARNm.....	10
3 Adressage subcellulaire.....	10
4 La traduction.....	11
5 Dégradation et stabilité des ARNm.....	11
<b>II Structures des protéines de liaison à l'ARN.....</b>	<b>13</b>
1 Les principaux domaines de liaison à l'ARN.....	13
1.1 Le motif de reconnaissance de l'ARN.....	13
1.2 Le domaine à doigts de zinc.....	16
1.3 Le domaine KH.....	17
1.4 Le domaine de liaison à l'ARN double-brin.....	19
2 Arrangement modulaire des protéines de liaison à l'ARN.....	20
2.1 Résidus connectant différents domaines.....	21
3 Méthodes d'étude des interactions protéine-ARN.....	23
3.1 Identification des RBPs.....	24
3.2 Caractérisation des acides aminés interagissant avec l'ARN.....	29
3.3 Caractérisation des séquences nucléotidiques reconnues par les RBPs.....	29
3.4 Caractérisation tridimensionnelle des complexes protéine-ARN.....	34
3.5 Code de reconnaissance protéine-ARN ?.....	35
<b>III Les principes du Docking.....</b>	<b>38</b>
1 L'étape d'échantillonnage.....	38
1.1 Traitement de la flexibilité du ligand.....	39
1.2 Traitement de la flexibilité de la protéine.....	43
2 Les fonctions de score.....	45
2.1 Principes physico-chimiques de la reconnaissance protéine-ligand.....	45
2.2 Fonctions de score basées sur champ de force.....	47
2.3 Fonctions de score empiriques.....	49
2.4 Fonctions de score à potentiels statistiques.....	50
2.5 Fonctions de score par méthodes d'apprentissage.....	51
2.6 Fonctions de score consensus.....	52
3 Approches par fragment.....	53
3.1 Principes des approches expérimentales et in silico.....	53
3.2 MCSS et méthodes complémentaires associées.....	56
<b>IV Modélisation des interactions protéine-ARNsb : état de l'art.....</b>	<b>61</b>
1 Du 2D à la 3D.....	61
2 Le docking protéine-ARN.....	61



3	Approches par fragment pour modéliser les interactions entre protéines et ARN liés sous forme simple-brin non structurée.....	62
3.1	RNA-LIM.....	62
3.2	Approche basée sur le programme de docking ATTRACT.....	64
3.3	RNP- <i>denovo</i> .....	66

## **Travaux de thèse.....69**

<b>I</b>	<b>Objectifs.....</b>	<b>69</b>
<b>II</b>	<b>Méthodes générales.....</b>	<b>71</b>
1	Simulations de docking avec MCSS.....	71
1.1	Procédure d'échantillonnage et paramètres utilisés.....	71
1.2	Préparation des protéines.....	72
1.3	Fragment nucléotidique utilisé comme ligand.....	72
1.4	Définition de l'espace d'échantillonnage.....	76
1.5	Evaluation de l'énergie d'interaction.....	77
2	Mesures de la déviation quadratique moyenne (RMSD).....	80
3	Procédure de regroupement des poses (clustering).....	80
<b>III</b>	<b>Développement de l'approche FBDRNA lorsque la séquence ARN est connue.....</b>	<b>81</b>
1	Introduction.....	81
2	Matériels et méthodes.....	85
2.1	Sélection du jeu de données.....	85
2.2	Simulations de docking.....	87
2.3	Molpy.....	88
2.4	Analyse de la conformation des nucléotides.....	90
3	Résultats.....	90
3.1	Analyse des performances de l'étape de docking.....	90
3.2	Mise en place d'une stratégie de sélection.....	99
4	Discussion.....	114
<b>IV</b>	<b>Développement de l'approche FBDRNA sans <i>a priori</i> sur la séquence ARN.....</b>	<b>121</b>
1	Introduction.....	121
2	Matériels et méthodes.....	122
2.1	Jeu de données.....	122
2.2	Simulations de docking.....	122
2.3	Recherche de chaînes.....	123
3	Résultats.....	123
3.1	Augmentation de la combinatoire.....	123
3.2	Comparaison de l'énergie d'interaction entre poses natives et native-like.....	124
3.3	Adaptation de la stratégie de sélection des poses "diviser pour mieux régner".....	127
3.4	Recherche de chaînes à partir des poses retenues par la procédure de sélection adaptée "diviser pour mieux régner".....	131
3.5	Analyse de la composition nucléotidique des chaînes natives et native-like.....	133
4	Discussion.....	134
<b>V</b>	<b>Etude de l'influence de la structure du nucléotide sur les performances de docking.....</b>	<b>138</b>
1	Introduction.....	138
2	Matériels et méthodes.....	140
2.1	Sélection du jeu de données protéine-nucléotide.....	140

2.2	Simulations de docking.....	142
3	Résultats.....	143
3.1	Echantillonnage et scoring des cinq structures de ligand.....	143
4	Discussion.....	149
<b>VI</b>	<b>Comparaison de cinq fonctions de score dans leur capacité à discriminer les poses natives.....</b>	<b>152</b>
1	Introduction.....	152
2	Matériels et méthodes.....	153
2.1	Données.....	153
2.2	Fonctions de scores comparées.....	153
2.3	Adaptation du format des fichiers de coordonnées.....	155
3	Résultats.....	155
4	Discussion.....	157
<b>VII</b>	<b>Conclusions générales et perspectives.....</b>	<b>160</b>
<b>VIII</b>	<b>Annexes.....</b>	<b>164</b>
1	Définition d'un seuil de clustering adapté pour la stratégie de sélection "diviser pour mieux régner".....	164
2	Contacts cristallins du nucléotide U1 de 5ELH.....	165
3	Comparaison de l'énergie d'interaction des poses puriques et pyrimidiques.....	166
4	Sélection des deux nucléotides de plus basse énergie dans les clusters à 2 Å.....	166
5	Description du jeu de données protéine-nucléotide non-redondant.....	168
5.1	Analyse des contacts et liaisons hydrogènes protéine-ligand.....	172
5.2	Evaluation de l'énergie d'interaction entre protéine et ligand cristallisé.....	172
5.3	Fraction de la surface enfouie du ligand.....	172
6	Ajustement du jeu de données protéine-nucléotide pour les calculs de docking.....	172
	<b>Bibliographie.....</b>	<b>176</b>

# Introduction

Les interactions protéine-ARN interviennent dans de nombreux processus cellulaires fondamentaux et sont essentielles dans tous les règnes du vivant. Elles font intervenir une variété de domaines de liaison à l'ARN (RBDs) dont la plupart reconnaît et lie spécifiquement une courte séquence nucléotidique accessible sous forme simple-brin. Une compréhension approfondie des modes de reconnaissance mis en jeu passe généralement par l'obtention de détails à résolution atomique de complexes structuraux. Ces derniers sont le plus souvent apportés par la cristallographie aux rayons X et la résonance magnétique nucléaire. Cependant, ces techniques sont malheureusement encore souvent longues et difficiles à mettre en place, particulièrement pour les complexes protéine-ARN. Le docking est une approche computationnelle qui, bien que complémentaire à ces méthodes expérimentales, peut être utilisée comme une alternative. Il permet en effet de fournir des modèles de la structure tri-dimensionnelle de complexes moléculaires à partir des coordonnées atomiques des partenaires d'intérêt. La qualité prédictive des approches traditionnelles de docking s'avère globalement très satisfaisante tant que les changements de conformations entre les formes liées et non liées des molécules restent limités. Modéliser le mode d'interaction entre protéines et ARNs simple-brin (ARNsb) représente en revanche une difficulté de taille en raison du caractère hautement flexible de ces derniers et donc du nombre important de conformations qu'ils peuvent potentiellement adoptées.

Dans les chapitres qui suivent, l'importance physiologique des protéines de liaison à l'ARN sera soulignée avant que leur arrangement structural ne soit présenté. Après une introduction aux généralités et principes du docking, un état de l'art de la modélisation des interaction protéine-ARN sera finalement décrit.

# I Importance fonctionnelle des protéines de liaison à l'ARN

Dès leur transcription, les molécules d'ARN se retrouvent rapidement associées à un ensemble de protéines et forment ainsi des complexes appelés ribonucléoprotéines (RNPs). Ces protéines qui lient l'ARN (RBPs) jouent un rôle essentiel dans le métabolisme cellulaire en assistant les ARNs dans leur biogenèse, leur fonction, leur adressage cellulaire, leur stockage, ou encore en participant à leur stabilité (Fig. 1). Les ARN ciblés par les RBPs sont les ARN messagers mais aussi tous les ARN non-codants: tRNA, rRNA, snRNA, miRNA, piRNA, etc . Pour illustrer l'importance des interactions protéine-ARN, nous nous focaliserons ici sur le rôle des RBPs dans la biogenèse des ARN messagers (ARNm) et leur régulation post-transcriptionnelle chez les eucaryotes sachant que certaines protéines peuvent s'associer aussi bien aux ARNm qu'aux ARN non-codants (Glisovic et al. 2008 ; Gerstberger, Hafner, and Tuschl 2014 ; Sutherland, Siddall, Hime, & McLaughlin, 2015).

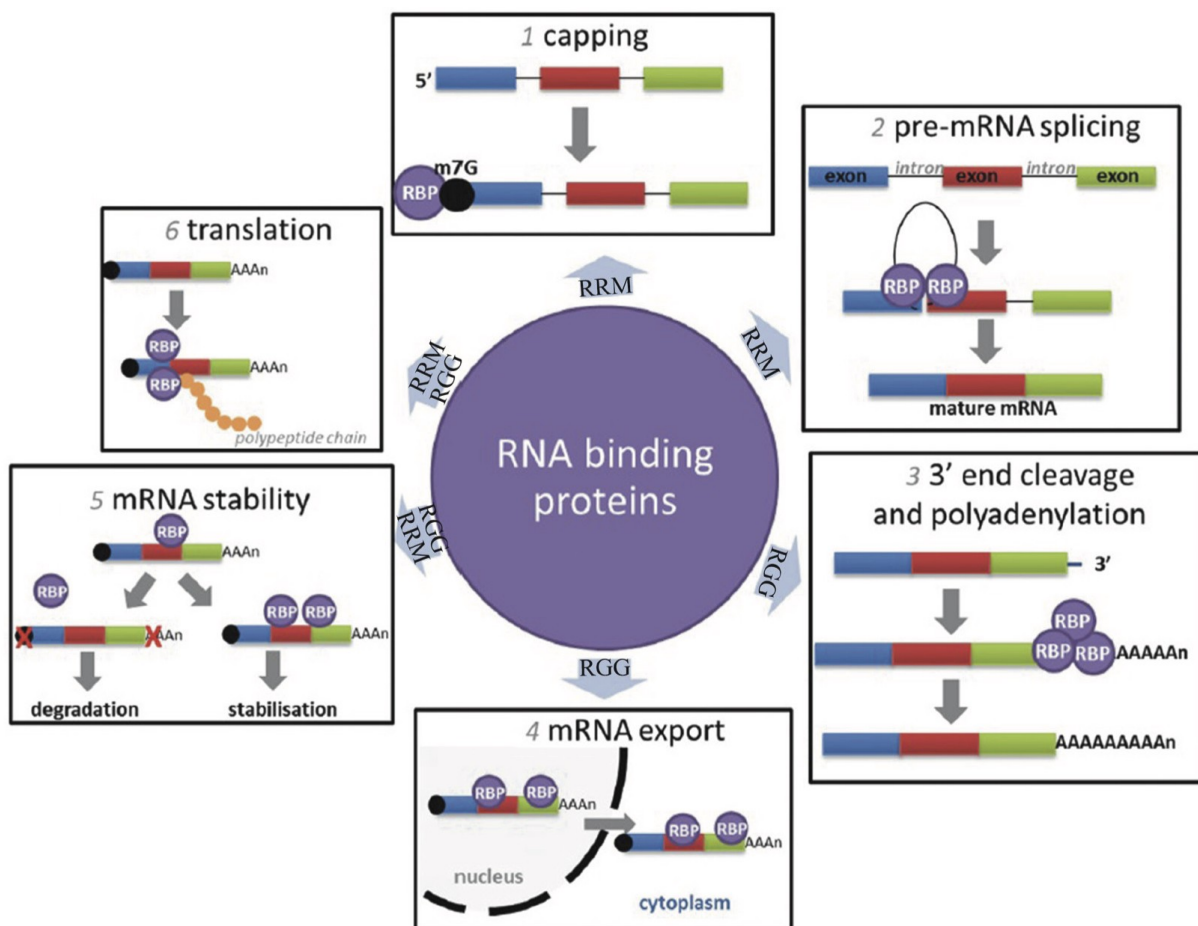


Figure 1: Régulations post-transcriptionnelles par les protéines de liaison à l'ARN. Les domaines impliqués dans la liaison à l'ARN sont indiqués dans les six cas de figures : RRM ("RNA recognition motif") ou RGG (arginine-glycine-rich). Tirée de Sutherland et al. (2015).

# 1 Maturation de l'ARN pré-messager précurseur

Chez les eucaryotes, avant qu'un ARNm ne devienne apte à être traduit en protéines, son ARN précurseur (pré-ARNm) est amené à subir un nombre important de modifications : l'ajout d'une coiffe methyl-guanosine en 5', un épissage, une polyadénylation en 3', et même parfois une édition. Toutes ces étapes de maturation du pré-ARNm font intervenir un nombre important et varié de RBPs.

## 1.1 *La famille des protéines nucléaires hétérogènes*

Dès le début de sa transcription, l'ARN primaire est pris en charge par un ensemble de RBPs qui participent à orchestrer toutes les étapes de maturation : les protéines nucléaires hétérogènes (hnP). Ces protéines, en association avec le pré-ARNm, forment un large assemblage amené à être remodelé et génériquement appelé complexe de ribonucléoprotéine hétérogène nucléaire (hnRNP). Sa composition évolue en effet au cours des différentes étapes de maturation selon la perte ou l'acquisition de RBPs, et d'autres protéines auxiliaires, impliquées dans des réactions spécifiques comme l'épissage ou la polyadénylation. Les hnPs se déclinent en 20 RBPs majeures et quelques autres dites mineures (Beyer, Christensen, Walker, & LeSturgeon, 1977; Chaudhury, Chander, & Howe, 2010; Singh & Valcárcel, 2005) nommées ainsi en raison de leur plus faible expression et de leur rôle régulateur (Dreyfuss, Matunis, Pinol-Roma, & Burd, 1993). Parmi toutes ces hnPs, certaines, pourtant précocement liées, se retrouvent encore associées à l'ARNm et participent à son export, son adressage subcellulaire, sa traduction ou encore sa stabilité (Dreyfuss, Kim, & Kataoka, 2002). Les protéines hnPs établissent un couplage physique et fonctionnelle entre toutes les étapes de maturation et influencent même la destinée ultérieure de l'ARNm mature (Maniatis & Reed, 2002). Leur diversité fonctionnelle ainsi que les pathologies auxquelles ces protéines ont été associées (Geuens, Bouhy, & Timmerman, 2016) reflètent leur caractère essentiel. La figure 2 ci-dessous illustre la structure d'une hnRNP en interaction avec une chaîne d'ARN.

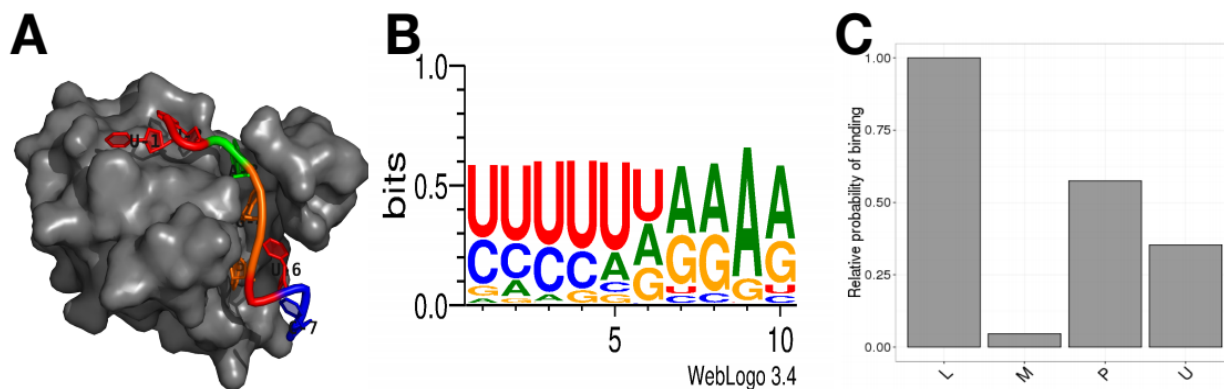


Figure 2: RBP impliquée dans l'épissage (RRM). A. Structure 3D du domaine RRM1 de la hnRNP A1 lié à l'ARN (PDB ID: 5MPG); B. Exemple de motifs reconnus (weblogo extrait POSTAR2: Zhu et al., 2018). C. Probabilité de structures secondaires pouvant être adoptées par les motifs reconnus (graphe RNAContext extrait de POSTAR2). L: internal loop et multiloop; M: multiloop; P: paired; U: unpaired.

## 1.2 L'épissage

L'épissage est un processus qui consiste à retirer les introns de l'ARN précurseur et à joindre ses exons. Cette étape est majoritairement catalysée par le spliceosome, un large assemblage moléculaire composé de diverses protéines associées à cinq complexes protéine-ARN essentiels appelés petites ribonucléoprotéines nucléaires ou snRNPs. D'autres RBPs sont par ailleurs impliquées dans la régulation de ce processus et sont fondamentales à l'existence de transcrits alternatifs, source importante de diversité protéomique chez les eucaryotes (Nilsen & Graveley, 2010). Chez l'Homme par exemple, près de 95 % des transcrits primaires contenant plusieurs exons sont sujets à un épissage alternatif (Q. Pan, Shai, Lee, Frey, & Blencowe, 2008; E. T. Wang et al., 2008). L'importance de l'épissage et sa régulation dans l'homéostasie cellulaire est reflétée par les pathologies humaines liées à un épissage défectueux (Licatalosi & Darnell, 2006).

## 1.3 La polyadénylation

L'étape de polyadénylation est étroitement liée à la terminaison de la transcription (Richard & Manley, 2009) et joue un rôle essentiel en influençant l'export de l'ARNm vers le cytoplasme, l'efficacité de sa traduction, et aussi sa stabilité. A l'exception de l'ARNm encodant l'histone réplication-dépendant, tous les ARNm eucaryotiques sont polyadénylés. Le processus se fait en deux étapes couplées consistant au clivage de l'ARN naissant suivi de la synthèse d'une queue poly-A à son extrémité 3' terminale. La machinerie moléculaire responsable de la polyadénylation inclut un noyau dur d'une vingtaine de facteurs protéiques (Shi et al., 2009), dont notamment le

facteur spécifique de polyadénylation/clivage (CPSF, Fig. 3) et la poly(A) polymérase (PAP). Les sites de polyadénylation (PAS) sont entourés et définis par des motifs plus ou moins conservés reconnus par cette machinerie. Plusieurs PAS peuvent être présents sur un même ARN et ainsi conduire à différentes isoformes d'ARNm (Barabino & Keller, 1999). Ce phénomène de polyadénylation alternative (APA) est en réalité très commun chez de nombreux eucaryotes et contribue, tout comme l'épissage alternatif, à accentuer la diversité protéomique. Il concernerait par exemple près de 70 % des gènes codant un ARNm chez les mammifères (Hoque et al., 2013). Globalement, le processus de polyadénylation fait intervenir de manière coordonnée de nombreuses RBPs dont des dysfonctionnements peuvent une fois encore conduire à des maladies graves (Curinha, Oliveira Braz, Pereira-Castro, Cruz, & Moreira, 2014).

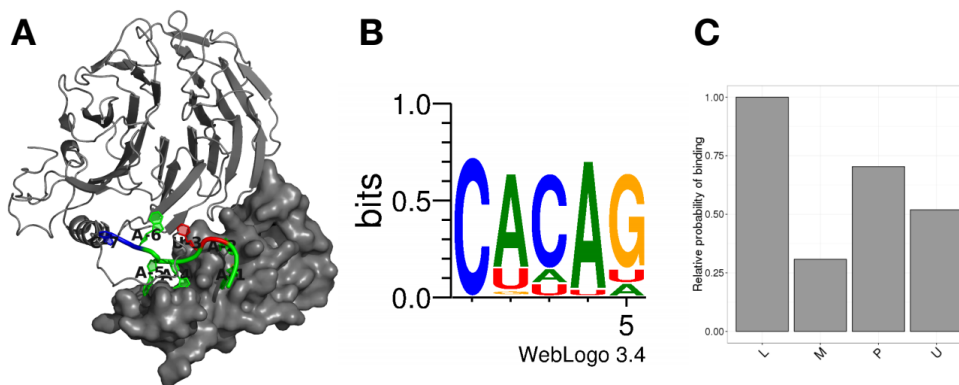


Figure 3: RBP impliquée dans la polyadénylation. A. Structure 3D partielle du complexe CPSF-160-WDR33-CPSF-30-PAS avec un ligand ARN (PDB ID: 6DNH); B. Exemple de motifs reconnus par CPSF4; C. Probabilité de structures secondaires pouvant être adoptées par les motifs reconnus.

## 1.4 L'édition

L'édition est une étape de maturation beaucoup moins fréquente chez les eucaryotes. Plusieurs types d'édition existent, mais le plus couramment observé chez les Primates consiste en la conversion d'adénosines (A) en inosines (I) (Bass, 2002) par des RBPs de la famille ADAR (Zipeto, Jiang, Melese, & Jamieson, 2015). Les inosines étant reconnues comme des guanosines, l'édition peu avoir d'importantes conséquences, notamment en changeant le potentiel codant de l'ARNm (Keegan, Gallo, & O'Connell, 2001). Une mutation du gène ADAR1 chez l'Homme conduit à une maladie de la peau, la dermatose pigmentaire (Miyamura et al., 2003).

## 2 Export de l'ARNm

Après qu'un système de surveillance se soit assuré que les étapes de maturation aient été correctement réalisées (Cole & Scarcelli, 2006; Moore, 2005; Rodriguez, Dargemont, & Stutz, 2004), le complexe ARNm-protéines (l'ARNm est toujours associé à un grand nombre de protéines), ou mRNP, est alors prêt à être transporté du noyau vers le cytoplasme. Ce transport peut se décliner en trois grandes étapes : la formation d'un complexe transporteur-mRNP (Fig. 4), sa translocation au travers d'un complexe formant le pore nucléaire (Pemberton and Paschal 2005) et la dissociation entre le transporteur et le mRNP dans le cytoplasme. La directionnalité du transport est assuré par un remodelage du mRNP en sortie du pore nucléaire. La protéine TAP/NFX1, qui interagit directement avec l'ARNm, est un exemple de RBP essentielle au processus de transport. Sa surexpression dans des ovocytes de *Xenopus laevis* a montré l'augmentation de l'export de transcrits matures qui sont, dans des conditions normales, transportés en faible quantité dans le noyau (Braun, Herold, Rode, & Izaurralde, 2002).

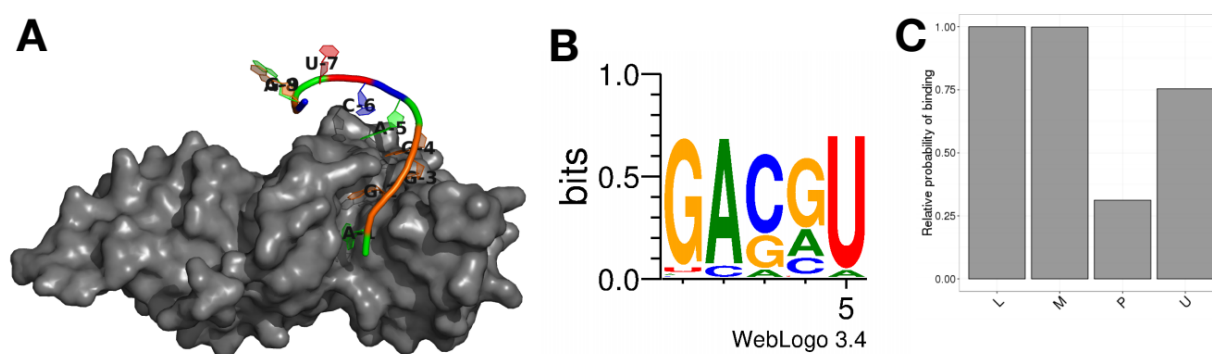


Figure 4: RBP impliquée dans le transport des ARNm. A. Structure 3D du complexe hnRNPA2/B1 avec un ligand ARN (PDB ID: 5WWE); B. Exemple de motifs reconnus par la hnRNPA2/B1; C. Probabilité de structures secondaires pouvant être adoptées par les motifs reconnus.

## 3 Adressage subcellulaire

Le transport des ARNm au sein d'une région précise du cytoplasme permet aux protéines d'y exercer spécifiquement leur(s) fonction(s). Cet adressage subcellulaire est particulièrement important dans les cellules polarisées. Un exemple souvent cité est celui concernant la protéine de base de la myéline (MBP). Son expression est restreinte à la région distale d'oligodendrocytes où elle participe au processus de myélinisation en s'insérant dans la membrane en cours de formation (Colman, Kreibich, Frey, & Sabatini, 1982). L'adressage de l'ARNm codant la MBP est dépendant de courts segments nucléotidiques présents dans les région 5' et 3' UTR. L'un de ces éléments



(A2RE) est spécifiquement reconnu par une RBP appartenant à la vaste famille des protéines hétérogènes nucléaires : hnP A2 (Ainger et al., 1997). La mutation d'une seule base dans la séquence A2RE est suffisante pour entraver sa reconnaissance et inhiber l'adressage de la MBP (Munro et al., 1999).

## **4 La traduction**

La traduction de l'ARNm en protéines est un processus complexe et finement régulé (Richter and Sonenberg 2005) faisant intervenir un large ensemble de facteurs protéiques incluant de nombreuses RBPs. La coiffe 5'-méthylguanosine est par exemple spécifiquement reconnue par le complexe eIF4F composé des facteurs d'initiation eucaryotiques (eIFs) eIF4A, eIF4E, et eIF4G. Le facteur eIF4E reconnaît et lie directement la coiffe, tandis que eIF4A possède une activité hélicase déroulant les éventuelles structures secondaires. L'action de ces deux RBPs combinée à celle de eIF4G participent à la fixation du complexe de pré-initiation 43S en 5' de l'ARNm. Ce dernier se compose, entre autres, de eIF3 qui interagit directement avec eIF4G. EIF4G s'associe également avec une RBP liée à la queue poly-A (PABP). Cette interaction conduit à une jonction des extrémités 5' et 3' de l'ARNm. La pseudo-circularisation de l'ARNm qui en résulte permettrait de stimuler la traduction, notamment en optimisant le processus de recyclage du ribosome (Kahvejian, Roy, & Sonenberg, 2001).

## **5 Dégradation et stabilité des ARNm**

Tous les ARNm sont amenés à être dégradés et ont donc une durée de vie limitée. Les protéines encodées par un ARNm peuvent en effet ne plus répondre au besoin actuel de la cellule. De plus, les étapes de la biogenèse d'un ARNm, bien que finement régulées, sont nombreuses et se déroulent dans un environnement moléculaire dynamique et complexe augmentant le risque de générer des erreurs. Les ARNm aberrants, et donc potentiellement délétères à la cellule, sont normalement reconnus par un système de surveillance qui les conduira à être dégradés (van Hoof & Wagner, 2011). La coiffe 5'-méthylguanosine et la queue poly-A en 3' protègent les ARNm d'une attaque exonucléolytique et contribuent donc à sa stabilité. La dégradation est souvent initiée par un processus de déadénylation qui raccourcit la queue poly-A. Deux voies distinctes peuvent ensuite être empruntées : soit une digestion est réalisée par une exonucléase 3' → 5', soit la coiffe en 5' est retirée pour permettre alors une digestion exonucléique dans le sens 5' → 3'. D'autres voies moins courantes sont possibles, comme la dégradation indépendante de la déadénylation ou encore la dégradation déclenchée par une attaque endonucléolytique (Garneau, Wilusz, & Wilusz, 2007).

Différentes RBPs interviennent dans la régulation de la stabilité des ARNm en reconnaissant certaines séquences spécifiques. Les éléments riches en AU (ARE) font partie de ces séquences régulatrices et représentent la classe d'éléments stabilisateurs la plus étudiée. Les ARE correspondent à des séquences variées qui se caractérisent néanmoins génériquement par la présence d'un triplet UUU entouré de A ou G à l'extrémité 3' des ARNm (Barreau, Paillard, & Osborne, 2005). Chez l'Homme, leur présence a été prédite chez près de 8 % des gènes (Bakheet, Frevel, Williams, Greer, & Khabar, 2001). La plupart des RBPs qui reconnaissent et lient les ARE exercent une action déstabilisante sur l'ARNm en autorisant le recrutement de protéines de la machinerie de dégradation (C. Y. Chen et al., 2001; Hollingworth et al., 2012; Lykke-Andersen & Wagner, 2005). D'autres contribuent au contraire à sa stabilité, soit en inhibant par compétition l'accès de facteurs déstabilisant à l'ARE, soit en stimulant la traduction (Lal et al., 2004).

## II Structures des protéines de liaison à l'ARN

La section précédente illustre le caractère essentiel des interactions protéine-ARN et montre la diversité fonctionnelle des protéines de liaison à l'ARN. La plupart des RBPs sont composées de domaines de liaison à l'ARN (RBDs) souvent répétées en tandem. Si certains RBDs interagissent avec l'hélice de type A caractéristique des ARN double-brin, beaucoup reconnaissent spécifiquement une courte séquence nucléotidique (entre deux et huit nucléotides selon les domaines) sous forme simple-brin. Quelques structures des domaines les plus communs sont décrits ci-dessous.

### 1 Les principaux domaines de liaison à l'ARN

#### 1.1 Le motif de reconnaissance de l'ARN

Le motif de reconnaissance de l'ARN (RRM) est retrouvé dans tous les domaines du vivant et est de loin le domaine de liaison à l'ARN le plus abondant chez l'Homme (Gerstberger, Hafner, Ascano, & Tuschl, 2014). Le domaine RRM se compose d'environ 90 acides aminés et présente une topologie  $\beta_1\alpha_1\beta_2\beta_3\alpha_2\beta_4$  qui adopte un repliement formé par un feuillet de quatre brins  $\beta$  entourés des deux hélices  $\alpha$  (Fig. 5). Les brins  $\beta_3$  et  $\beta_1$  sont constitués de deux motifs conservés appelés RNP1 et RNP2 (Fig. 6) composés respectivement de huit et six acides aminés (Adam, Nakagawa, Swanson, Woodruff, & Dreyfuss, 1986; Bandziulis, Swanson, & Dreyfuss, 1989 ; Afroz, Cienikova, Cléry, & Allain, 2015). Ces derniers exposent à la surface du feuillet trois résidus aromatiques conservés dont l'absence définit différentes sous-classes de domaines (Cléry, Blatter, & Allain, 2008) comme le quasi-RRM (qRRM) ou le pseudo-RRM ( $\Psi$ RRM).

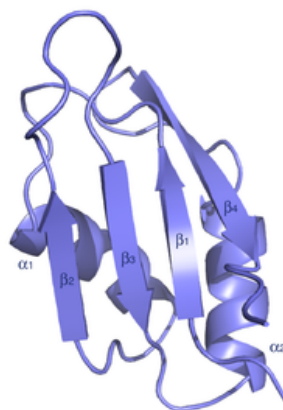


Figure 5: Topologie  $\beta_1\alpha_1\beta_2\beta_3\alpha_2\beta_4$  des domaines RRM.

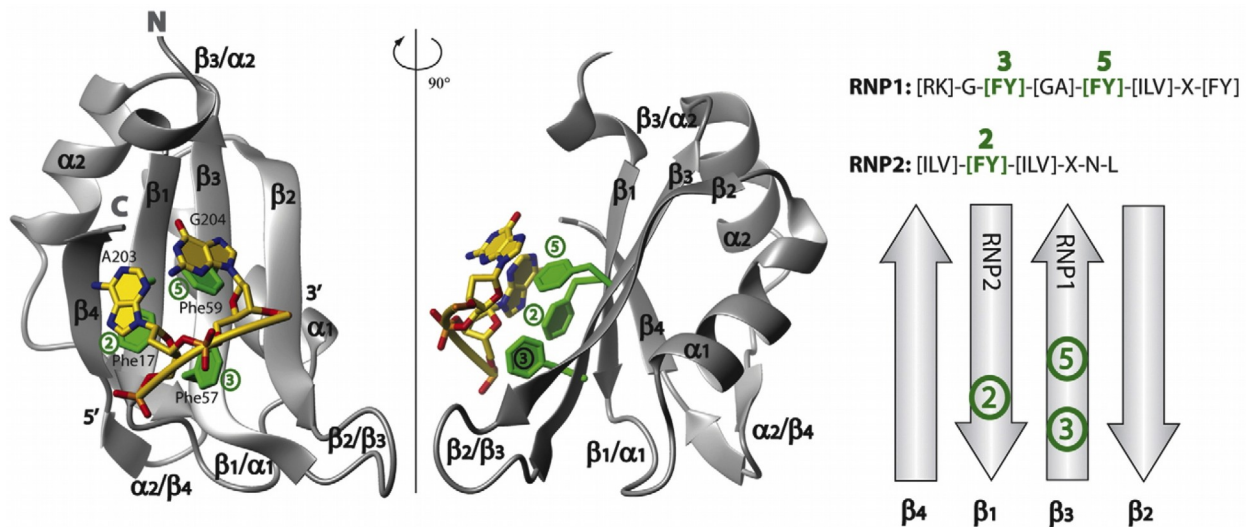


Figure 6: Liaison de l'ARN aux feuillet  $\beta$  du domaine RRM. La structure représentée est celle de la hnRNPA1 interagissant avec un simple brin d'ADN. Le schéma à droite indique la topologie des feuillet  $\beta$  et les motifs RNP1 et RNP2 conservés avec leur séquence consensus composée de résidus aromatiques. Tirée de Afroz et al., 2015.

Les domaines RRM sont très versatiles dans leur mode de reconnaissance des ARNs. Dans les interactions canoniques, l'ARN interagit au sein du feuillet  $\beta$  central où les résidus aromatiques des motifs RNP1 et RNP2 fournissent une parfaite plate-forme d'ancrage pour deux nucléotides consécutifs sous forme simple-brin (Fig. 7). Jusqu'à trois ou quatre nucléotides peuvent néanmoins y être accommodés grâce aux résidus des autres brins  $\beta$ . Certains domaines RRM présentent une extension de la surface du feuillet  $\beta$  permettant la reconnaissance de un ou deux nucléotides supplémentaires (Maris, Dominguez, & Allain, 2005; Oberstrass et al., 2005).

Des interactions non canoniques entre RRM et ARN peuvent également faire intervenir, en plus des résidus du feuillet  $\beta$ , des boucles et/ou les résidus des extrémités N- et/ou C-terminales souvent peu structurées (Oberstrass et al., 2005; Tsuda et al., 2009). Les boucles peuvent aussi être impliquées dans la reconnaissance d'éléments structurels de l'ARN. La structure du domaine RRM RBMY (Fig. 8) montre par exemple que les résidus de la boucle sont insérés dans le sillon majeur de l'ARN (Skrisovska et al., 2007) alors que les résidus du feuillet  $\beta$  interagissent spécifiquement avec les nucléotides non-appariés de la boucle de l'ARN.

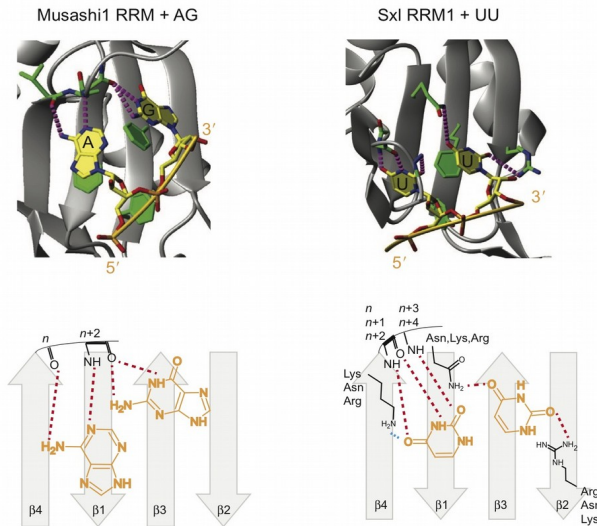


Figure 7: Reconnaissance par les domaines RRM de dinucléotides 5'-AG-3' et 5'-UU-3' (gauche: Musashi1 RRM; droite: Sxl RRM1). Tirée de Afroz et al., 2015.

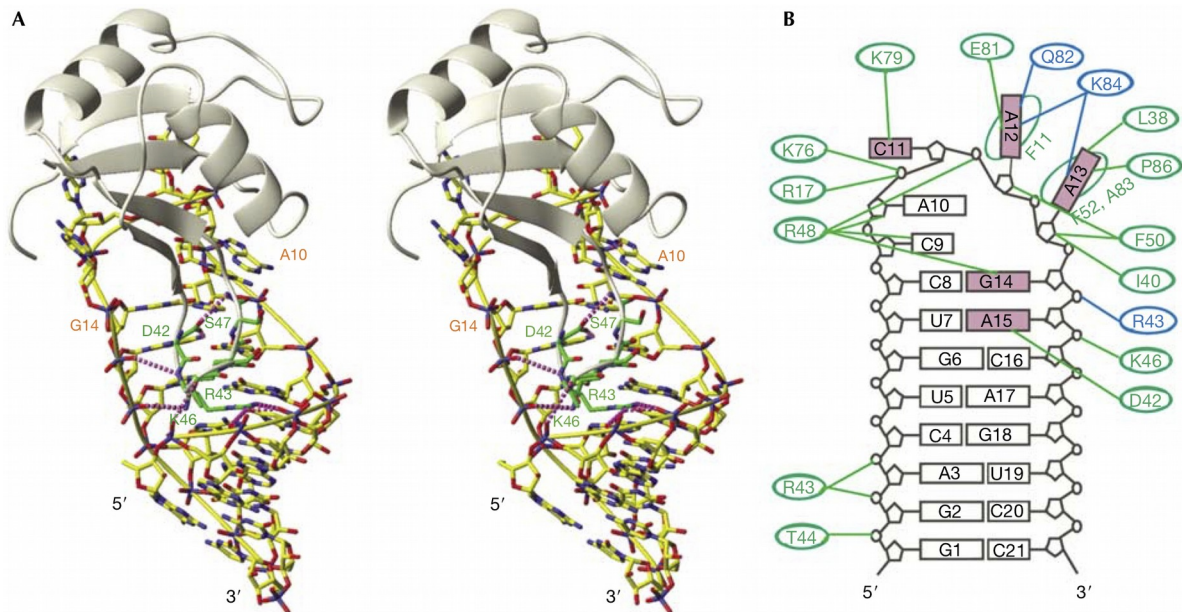


Figure 8: Interactions du domaine RRM de RBMY avec une région simple-brin en boucle de l'ARN. A. Vue stéréoscopique de l'interaction entre les feuillets  $\beta$ 2-  $\beta$ 3 et la boucle qui les connecte avec l'ARN. B. Détails des contacts entre résidus de la protéine et de l'ARN. Les résidus en vert correspondent à des contacts avec le squelette protéique, les résidus en bleu les contacts avec les chaînes latérales. Tirée de Maris et al., 2005.

## 1.2 Le domaine à doigts de zinc

Le domaine à doigts de zinc (Zn) est un petit domaine ubiquitaire composé d'environ 30 acides aminés. Sa topologie  $\beta\beta\alpha$  présente un repliement où l'épingle à cheveux  $\beta$  se retrouve associée à l'hélice  $\alpha$  par l'intermédiaire d'un ion  $Zn^{2+}$  coordonné à des résidus cystéines (C) et/ou histidines (H) comme le montre la figure 9. Différentes combinaisons de ces résidus définissent plusieurs classes de Zn comme par exemple les domaines Zn-CCHH, Zn-CCHC, Zn-CCCH ou Zn-CCCC. S'ils peuvent être retrouvés seuls, ces domaines sont le plus souvent répétés en tandem dans les RBPs, ou bien en association avec d'autres RBDs (Gerstberger, Hafner, & Tuschl, 2014). Le facteur de transcription TFIIA contient par exemple neuf domaines CCHH et peut lier aussi bien des molécules d'ADN que d'ARN (Pelham, 1980) en reconnaissant le squelette ribose-phosphate de régions double-brin (*Drosophila*, Lu, Searles, & Klug, 2003). Les protéines à domaines Zn-CCCH interagissent en revanche préférentiellement avec des séquences d'ARN simple-brin (Fig. 10). Des structures ont révélé que les domaines Zn-CCHC et Zn-CCCC peuvent aussi interagir avec des régions d'ARN simple-brin. Généralement, deux à trois nucléotides sont spécifiquement reconnus par l'établissement d'interactions de stacking et de liaisons hydrogènes spécifiques, particulièrement entre la base et des atomes de la chaîne principale des domaines. La structure 3D de ces domaines constitue donc une composante importante dans la reconnaissance spécifique de séquences d'ARN. Une caractéristique des Zn provient de leur plasticité structurale qui les autorise à adopter différents repliements en vue de reconnaître différentes séquences (Cléry & Allain, 2013).

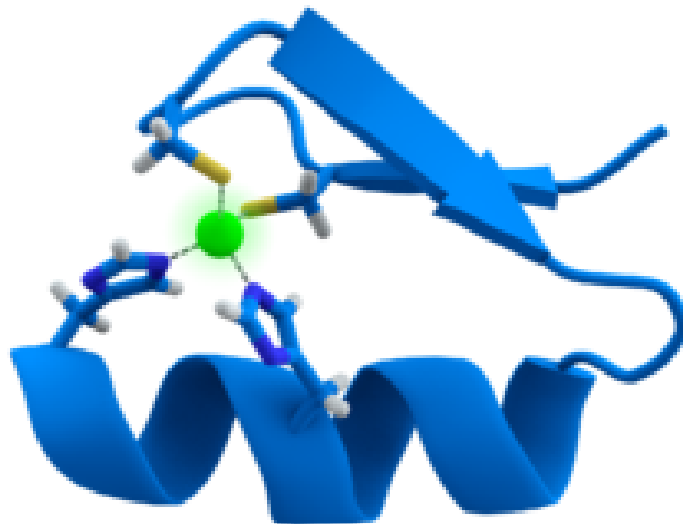


Figure 9: Domaine de doigt à zinc (CCHH) (Source: Wikipedia).



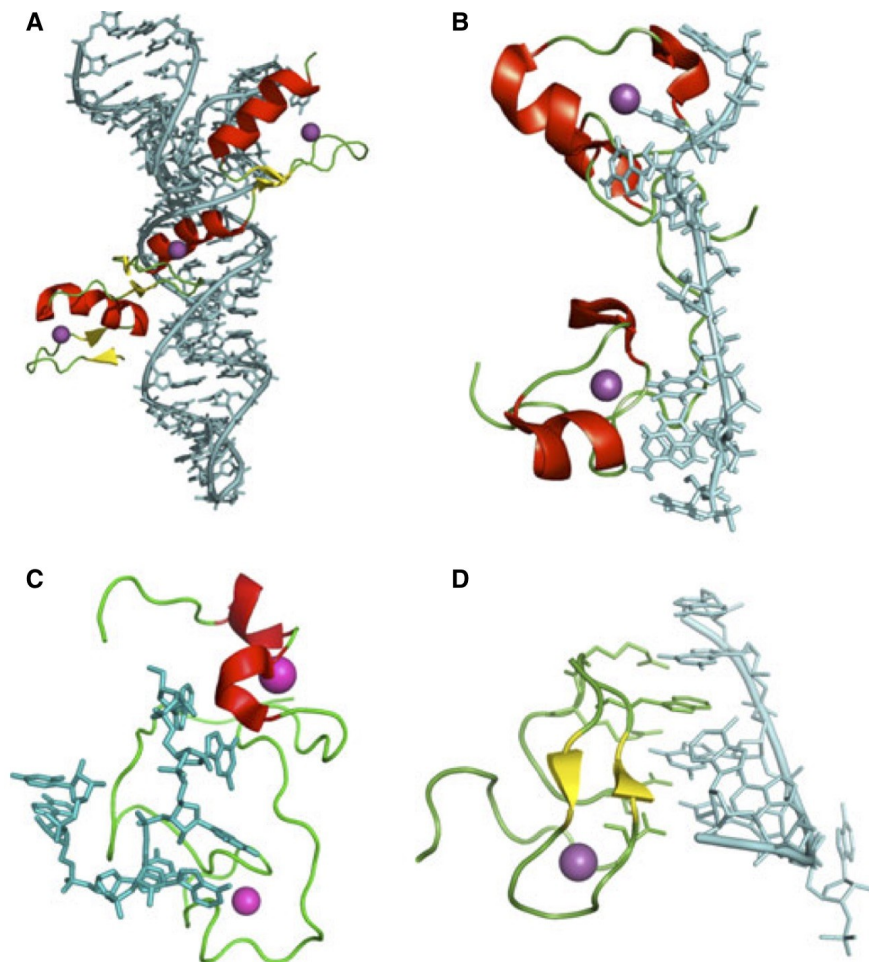


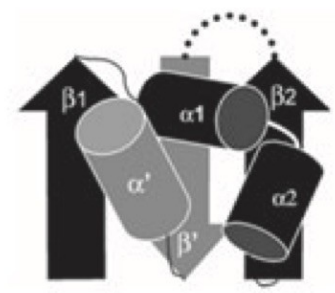
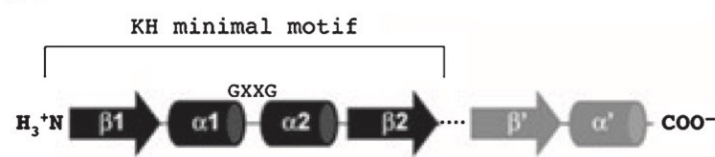
Figure 10: Modes d'interaction des protéines à doigts à zinc (Zn) avec l'ARN. A. Structure d'un domaine Zn de TFIIIA en complexe avec une région double-brin de l'ARNr 5S (PDB ID: 1UN6). B. Structure en solution de la protéine TIS11d (PDB ID: 1RGO). C. Structure en solution de la protéine Lin28-Zn en interaction avec un ARN simple brin de séquence 5'-AGGAGAU-3' (PDB ID: 2LI8). D. Structure cristallographique de la protéine ZRANB2 (PDB ID: 3G9Y) lié à un élément d'ARN simple-brin 5'-GGU-3'. Tirée de Chen et Varani, 2013.

### 1.3 Le domaine KH

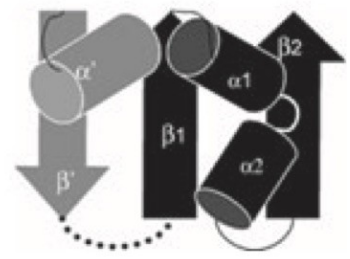
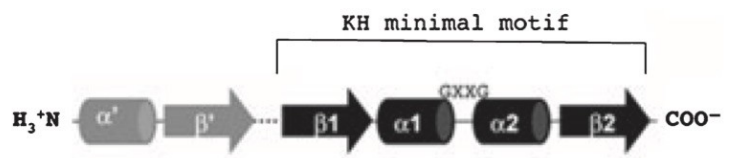
Le domaine KH (pour *K-homology*) tire son nom de la protéine nucléaire hétérogène K au sein de laquelle il a été identifié pour la première fois (Matunis, Matunis, & Dreyfuss, 1992). Il est lui aussi représenté dans tous les domaines du vivant et est capable de s'associer à des séquences simple-brin, aussi bien d'ADN que d'ARN (Valverde, Edwards, & Regan, 2008). Ce domaine est composé d'environ 70 acides aminés et se caractérise par la présence d'un motif conservé (I/L/V)-I-G-X-X-G-X-X-(I/L/V) occupant la position centrale du domaine (Fig. 11). Deux types de domaine KH se distinguent toutefois selon leur topologie (Grishin, 2001). Le type I est typiquement retrouvé

dans les protéines eucaryotiques et présente un agencement de ses structures secondaires de la forme  $\beta\alpha\beta\alpha$ . Il se caractérise en conséquence par un feuillet  $\beta$  composé de trois brins  $\beta$  anti-parallèles sur lesquels reposent les trois hélices  $\alpha$  (Fig. 11A). Le domaine KH de type II, trouvé chez les procaryotes, présente étonnement une structure tertiaire similaire malgré qu'il diffère dans l'arrangement de ses structures secondaires. Sa topologie  $\alpha\beta\beta\alpha\beta$  conduit cette fois à un feuillet  $\beta$  où le brin  $\beta'$  central se retrouve dans une position anti-parallèle à  $\beta_1$  et parallèle à  $\beta_2$  (Fig. 11B). Le mode de liaison typique des domaines KH implique la reconnaissance de quatre nucléotides. Ces derniers sont accommodés au sein d'une crevasse formée d'un côté par les hélices  $\alpha_1$ ,  $\alpha_2$ , et la boucle GXXG qui les relie, et d'un autre côté par le brin  $\beta_2$  et une boucle dont la longueur varie entre différents domaines KH (Fig. 12). Les deux nucléotides centraux du tétranucléotide reposent généralement sur un patch hydrophobe constitués des résidus I/L/V du motif conservé, où ils sont par ailleurs stabilisés par des interactions électrostatiques ou des liaisons hydrogènes. Ce mode d'interaction diffère de ceux observés chez les domaines RRM et Zn pour qui l'ancrage de nucléotides est principalement établi par des interactions de stacking avec des résidus aromatiques. Il explique certainement la faible affinité de liaison (de l'ordre du micromolaire) que possèdent les domaines KH seuls envers l'ARN. Certains domaines KH présentent cependant une extension de leur surface d'interaction par l'addition d'hélices  $\alpha$  les autorisant à contacter jusqu'à six nucléotides (Teplova et al., 2013). Ces interactions additionnelles leur permettent d'augmenter l'affinité de liaison et la spécificité de reconnaissance.

**A Type I KH domain:**



**B Type II KH domain:**



*Figure 11: Topologie des éléments de structure 2D des protéines à domaine KH. A. Domaine KH de type I. B. Domaine KH de type II. Les liens en pointillé indiquent la connexion entre les feuillets  $\beta_2$  et  $\beta'$  par une boucle variable. Tirée de Valverde et al., 2008.*



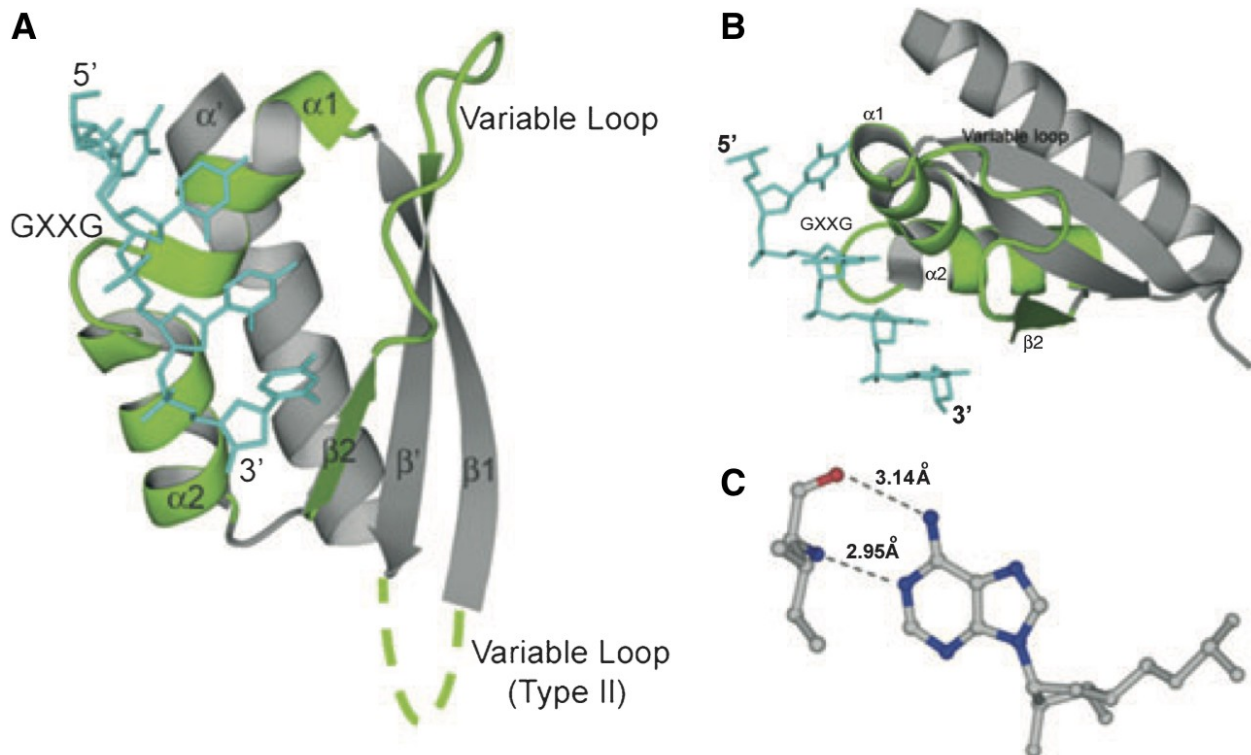


Figure 12: Traits caractéristiques des interactions protéine-ARN pour les domaines KH. A. Domaine KH de type I; la région de liaison à l'ARN inclut l'hélice  $\alpha_1$ , le motif GXXG, l'hélice  $\alpha_2$  et le feuillet  $\beta_2$  et une boucle variable (vert). B. Acide nucléique comportant quatre nucléotides empilés qui interagissent avec le domaine KH. C. Détails des contacts avec un résidu d'adénosine. (PDB ID: 1J5K, 2ASB). Tirée de Valverde et al., 2008.

#### 1.4 Le domaine de liaison à l'ARN double-brin

Le domaine de liaison à l'ARN double-brin (dsRBD) est un RBD composé d'environ 70 acides aminés agencés selon une topologie  $\alpha\beta\beta\alpha$ . Ces structures secondaires adoptent un repliement où les deux hélices  $\alpha$  reposent sur le feuillet formé des trois brins  $\beta$  anti-parallèles. Comme son nom l'indique, le dsRBD reconnaît l'hélice de type A caractéristique de l'ARN sous forme double-brin, en s'associant à deux sillons mineurs séparés par un sillon majeur. L'interaction fait intervenir d'une part, des résidus de l'hélice  $\alpha_1$  et de la boucle  $\beta_1$ - $\beta_2$  qui contactent chacun un sillon mineur, et d'autre part des résidus N-terminaux de l'hélice  $\alpha_2$  qui lient le sillon majeur (Fig. 13). Les contacts établis impliquent essentiellement les groupements 2'OH et phosphate et sont donc indépendants de la séquence nucléotidique (Stefl, Skrisovska, and Allain 2005). Quelques cas particuliers ont néanmoins été identifiés révélant que certains dsRBD peuvent établir des contacts spécifiques, comme cela a été montré pour les protéines Staufen (Ramos et al., 2000) et ADAR2 (Stefl et al., 2010).

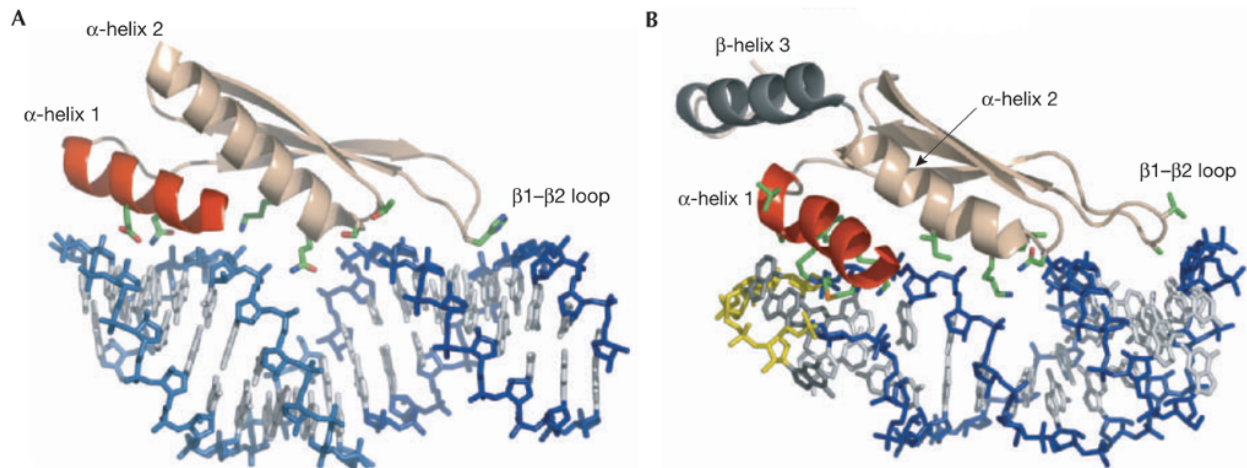


Figure 13: Reconnaissance de l'ARN double-brin par une protéine à domaine RBM. A. le domaine de liaison à l'ARN double-brin (dsRBM) de la protéine *Xlrp2*; les éléments de structure qui lient l'ARN sont: l'hélice  $\alpha 1$  (rouge), l'extrémité amino-terminale de l'hélice  $\alpha 2$  et les feuillets  $\beta 1$  et  $\beta 2$ . B. Domaine dsRBM de la protéine *Rnt1p* liée à un ARN en épingle à cheveux avec une tetraboucle terminale AGNN. L'hélice  $\alpha 1$  (rouge) et l'extrémité carboxy-terminale de l'hélice  $\alpha 3$  (noir) constituent les éléments de reconnaissance. Tirée de Stefl et al., 2005.

## 2 Arrangement modulaire des protéines de liaison à l'ARN

La propriété qu'ont les RBDs classiques à ne reconnaître qu'une courte séquence nucléotidique les rend intrinsèquement peu spécifiques envers les ARNs. Les protéines de liaison à l'ARN parviennent néanmoins à cibler spécifiquement leurs partenaires ARNs, et cela grâce à une architecture modulaire (Lunde, Moore, & Varani, 2007) où plusieurs RBDs se retrouvent répétés en tandem (Fig 14). Cet agencement leur permet de combiner les propriétés de liaison propres à chaque domaine et confère aux RBPs la capacité à s'associer spécifiquement à leurs ARNs cibles avec des affinités de liaison élevées. Par exemple, les protéines à domaines RRM peuvent comporter entre deux domaines (protéine U1A) et quatre domaines (protéine PABP) espacés les uns des autres par des résidus dont la longueur peut varier. Elles peuvent aussi inclure différents domaines RBDs agencés de façons spécifiques: la protéine U2AF35 comporte un domaine RRM et deux domaines à doigts à zinc (Zn-CCCH) qui l'encadrent, la protéine SF1 un domaine KH et un domaine un domaine à doigts à zinc (Zn-CCHC). Un certain nombre de RBPs appartiennent à l'une des grandes familles de protéines de liaison à l'ARN définie par leur type de RBDs ou combinaison de RBDs; elles peuvent aussi avoir une activité enzymatique qui est portée, dans ces cas de figure, par des domaines spécifiques: endonucléase, hélicase, kinase, etc.

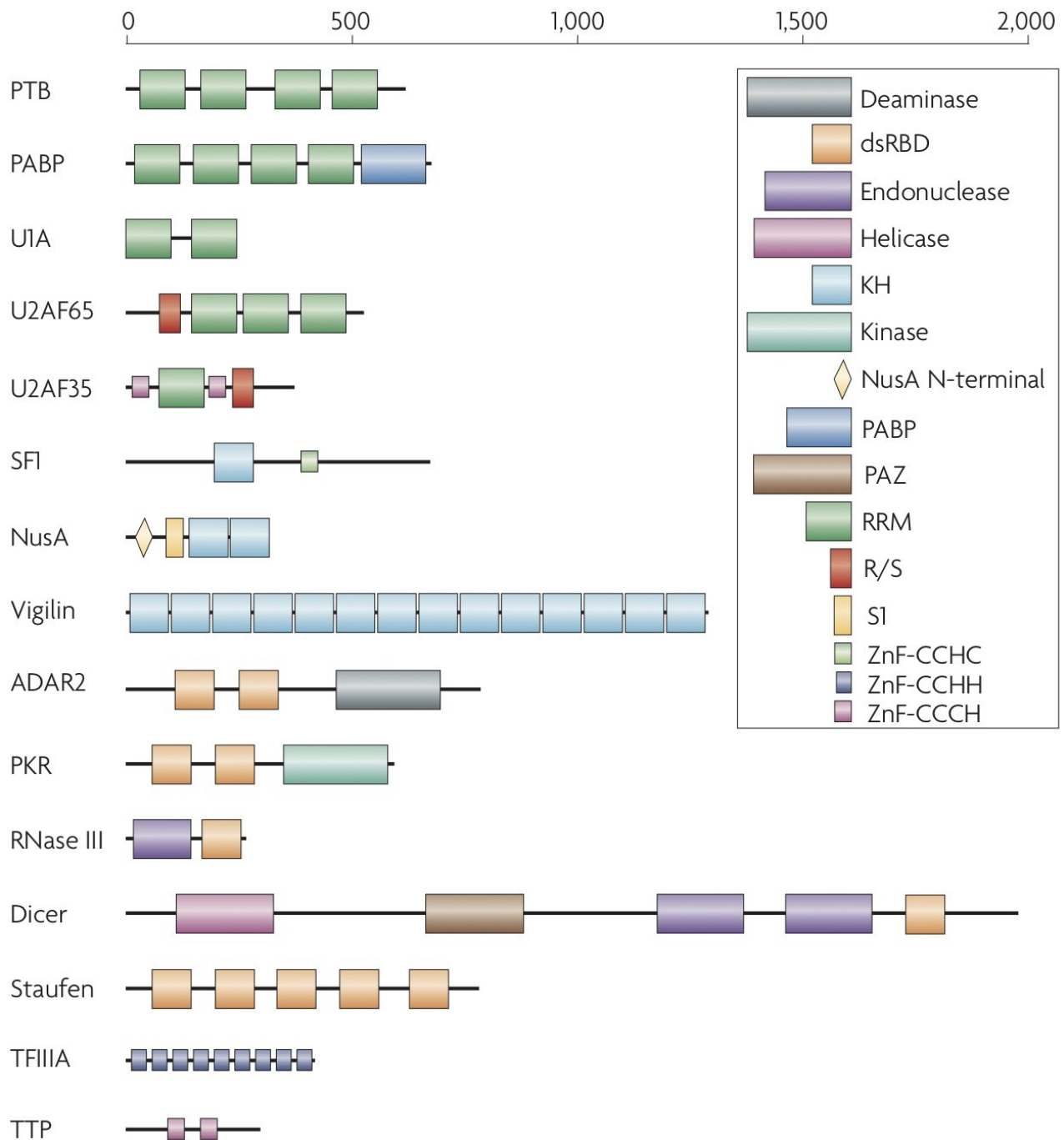


Figure 14: Modularité des protéines de liaison à l'ARN. Exemples représentatifs des grandes familles de protéine de liaison à l'ARN. Tirée de Lunde et al., 2007.

## 2.1 Résidus connectant différents domaines

Les résidus connectant différents domaines, ou plus simplement les résidus *linkers*, possèdent deux propriétés importantes qui jouent un rôle important dans la reconnaissance protéine-ARN: leur longueur et leur flexibilité qui vont déterminer la façon dont les RBDs vont être positionnés les uns par rapport aux autres et définir leur ou leurs interfaces potentielles avec l'ARN. Dans le cas très



formes libre et liée à l'ARN comme c'est le cas pour la protéine HuR à domaines RRM (H. Wang et al., 2013). Cette flexibilité du *linker* permet alors d'avoir une reconnaissance protéine-ARN de type "ajustement induit". Beaucoup de RBPs et en particulier celles à domaines RRM, possèdent une spécificité de reconnaissance étendue avec différents motifs ARN pouvant être liés. Il faut distinguer la spécificité de reconnaissance par la protéine entière et par les domaines RRM pris isolément. Pour HuR, la spécificité de liaison du domaine RRM3 n'est pas très forte avec des motifs de séquences différentes se liant avec des affinités similaires. La dimérisation induisant un changement de conformation ou modifiant l'interface protéine-ARN du monomère peut conduire à une altération de la spécificité de reconnaissance (Ripin et al., 2019).

### 3 Méthodes d'étude des interactions protéine-ARN

Les interactions protéine-ARN jouent un rôle important dans la cellule; on estime que les RBPs représentent entre 3 et 11% des protéines cellulaires bactériennes, d'archaea ou d'eucaryotes (Beckmann et al., 2016). Chez l'homme, elles représenteraient 7,5% du protéome (Gerstberger, Hafner, & Tuschl, 2014) mais leur proportion est sans doute sous-estimée pour deux raisons:

1. un certain nombre de protéines connues sont en fait des RBPs non identifiées et non annotées comme telles (y compris des enzymes non associés à la biologie des ARN) car elles ne présentent pas de domaine RBD ni de mode d'interaction "classiques" (Helder, Blythe, Bond, & Mackay, 2016; Hentze, Castello, Schwarzl, & Preiss, 2018) tels que ceux décrits précédemment (RRM, KH, Zn-CCCH, etc)
2. de nouvelles RBPs restent encore à identifier ("enigmRBPs") parmi les protéines inconnues ou non caractérisées (Zhao, Yang, Janga, Kao, & Zhou, 2014) mais souvent conservées de la levure à l'homme (Beckmann et al., 2015).

Les avancées technologiques, faisant appel à de nouvelles approches qui permettent d'étudier les réseaux d'interaction protéine-ARN ("RBPome"), offrent des opportunités pour identifier de nouvelles RBPs (Bao et al., 2018; R. Huang, Han, Meng, & Chen, 2018; Treiber et al., 2017). Parmi les approches utilisées qui font appel à des méthodes de la biologie moléculaire, de la biophysique et biologie structurale, de la génomique/transcriptomique/protéomique etc, on peut distinguer les méthodes focalisés sur un ARN (ou une famille d'ARN) d'intérêt ("RNA-centric") de celles focalisées sur une protéine d'intérêt ("protein-centric"). Ces méthodes décrites dans différents articles de revues ou de méthodes (Cook, Hughes, & Morris, 2015; Marchese, de Groot, Lorenzo Gotor, Livi, & Tartaglia, 2016a; McHugh, Russell, & Guttman, 2014; Ramanathan, Porter, &

Khavari, 2019; Ray et al., 2009; Sugimoto et al., 2012) ne seront pas détaillées; nous insisterons surtout sur les données qui en découlent. Celles-ci contribuent à apporter des informations précieuses sur les interactions protéine-ARN; la plupart a donné lieu à la création de banques de données et/ou de méthodes d'analyse et de prédiction de ces interactions.

### **3.1 Identification des RBPs**

Avec le développement des approches à haut débit, de nombreuses méthodes ont vu leur champ d'application étendu pour identifier les protéines de liaison à l'ARN par des approches "protein-centric" qui utilisent des conditions de purifications natives ou dénaturantes (Kramer et al., 2014; McHugh et al., 2014; Ramanathan et al., 2019). Dans le premier cas, les complexes protéine-ARN sont purifiés par immunoprécipitation de l'ARN via les techniques RIP ("RNA immunoprecipitation") dont la fiabilité est souvent limitée pour des raisons d'extraction, de spécificité des interactions des complexes isolés (faux-positifs), de localisation du site d'interaction, etc (Barra & Leucci, 2017; McHugh et al., 2014). En conditions dénaturantes, les techniques CLIP ("cross-linking immunoprecipitation") reposent sur les propriétés de formation de ponts chimiques induits par radiations UV entre les bases nucléiques ou des analogues des nucléotides et la plupart des acides aminés (Fig. 16).



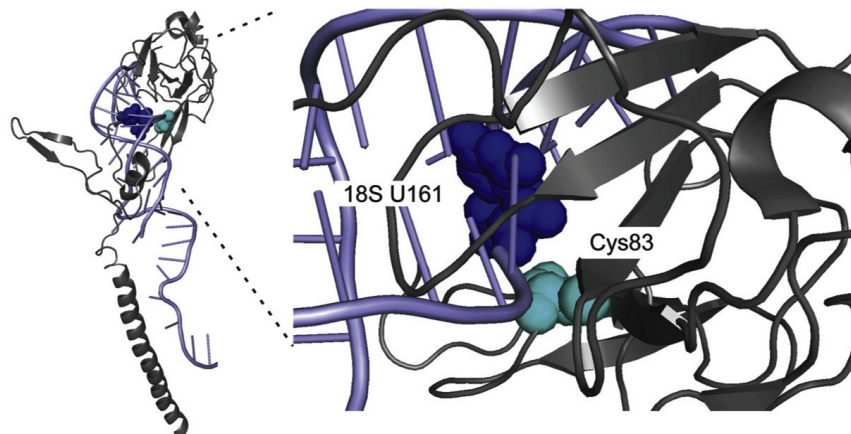
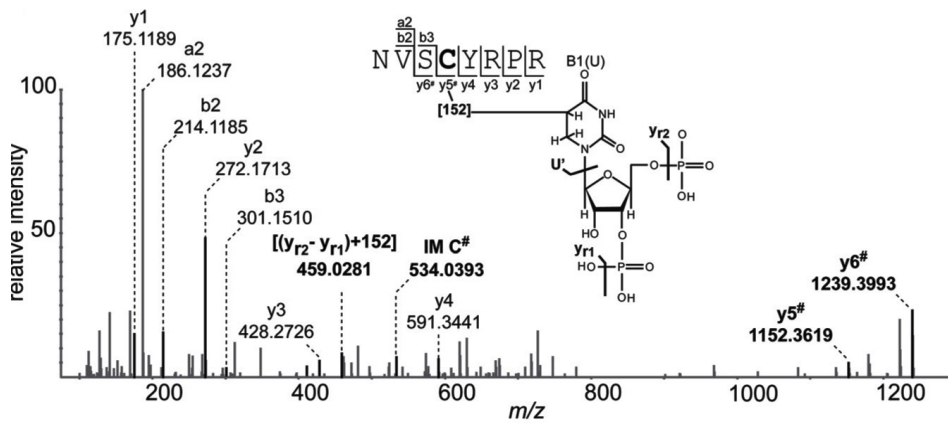


Figure 16: Exemple d'application du pontage chimique nucléotide-acide aminé (uridine et cystéine) pour la purification de complexes ARN-RBP. Le produit formé entre les résidus d'uridine et de cystéine est montré en haut. En bas, la structure du complexe entre une région de l'ARNr 18S et la protéine ribosomique S6 indique la proximité entre les deux résidus concernés par le pontage. Tirée de Zaman et al., 2015.

Ces techniques trouvent leur variante par l'utilisation d'approches à haut débit grâce aux nouvelles technologies (Cook et al., 2015; Marchese, de Groot, Lorenzo Gotor, Livi, & Tartaglia, 2016b; Wheeler, Van Nostrand, & Yeo, 2018) faisant appel aux puces ou "microarrays" (RIP-chip) à la PCR quantitative ou au séquençage massif "HITS" ou "high-throughput sequencing": RIP-seq (Fig. 17A), HITS-CLIP (Fig. 17B), PAR-CLIP (Fig. 17C), CLIP-seq, etc. L'avantage de l'ensemble de ces méthodes est qu'elles peuvent être appliquées *in vivo*.

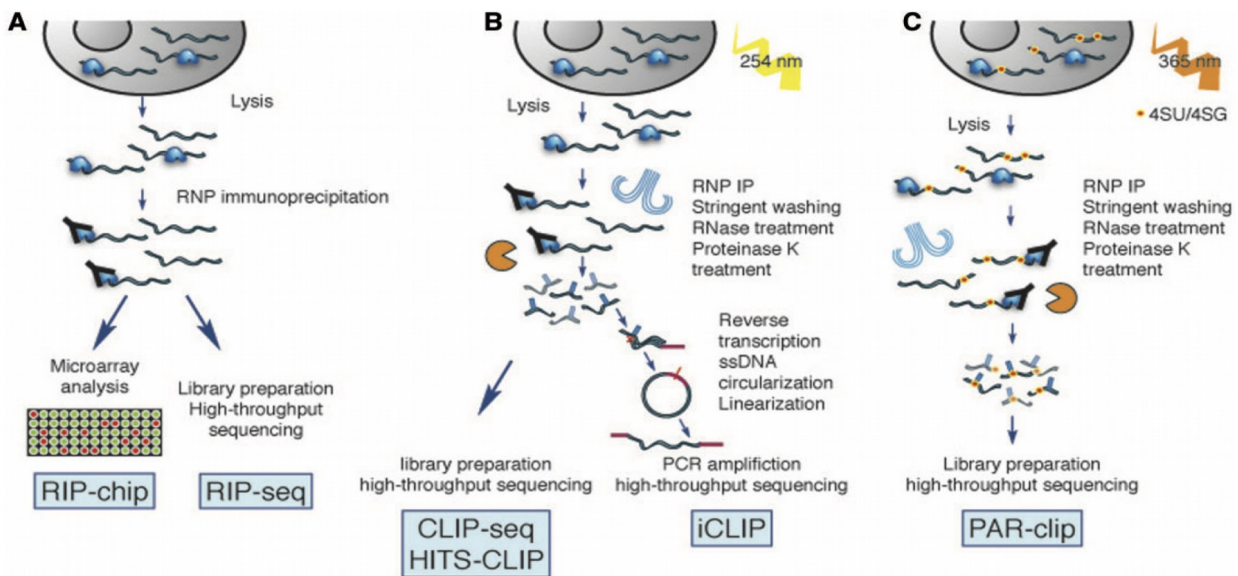


Figure 17: Approches "protein-centric" pour l'identification de ligands ARN *in vivo*. **A.** Méthodes RIP-chip et RIP-seq de détermination des ARN liés par immuno-précipitation et puces (microarrays) ou séquençage à haut débit. **B.** Méthodes de pontages chimiques et immuno-précipitation fournissant une meilleure résolution des sites de liaison. **C.** Méthodes dérivées de CLIP-seq utilisant des nucleotides modifiés pour des pontages chimiques induits par rayonnement UV. Tirée de Cook et al., 2015.

Deux plateformes compilent de nombreux jeux de données CLIP de différentes natures (CLIP-seq, PAR-CLIP, iCLIP, etc): POSTAR/POSTAR2 (Hu, Yang, Huang, Zhu, & Lu, 2017) et RNAct (Lang, Armaos, & Tartaglia, 2019). Elles allient aussi des méthodes *in silico* de prédiction des sites d'interaction protéine-ARN: RNApromo pour POSTAR (Rabani, Kertesz, & Segal, 2008) ou PARalyzer (Corcoran et al., 2011) et catRAPID pour RNAct (Agostini et al., 2013).

En parallèle, les méthodes "RNA-centric" permettent d'identifier un ensemble de RBPs qui se lient à un ARN spécifique (Ramanathan et al., 2019). Le fait que l'ARN soit connu permet de réaliser des systèmes d'expression ou d'hybridation de l'ARN d'intérêt qui vont faciliter la purification des RBPs que ce soit *in vitro* ou *in vivo*. Tout comme les méthodes "protein-centric", ces méthodes peuvent faire appel à la formation de pontages chimiques ou non via la biotinylation des RBPs fixées à l'ARN pour leur purification *in vivo* à partir de cellules (Fig. 18). *In vitro*, plusieurs options sont possibles avec des ARN modifiés par biotinylation en 5', fusion avec un aptamère permettant la fixation sur une résine ou fixation d'un composé fluorescent permettant l'utilisation de "protein microarray" (Fig. 19).



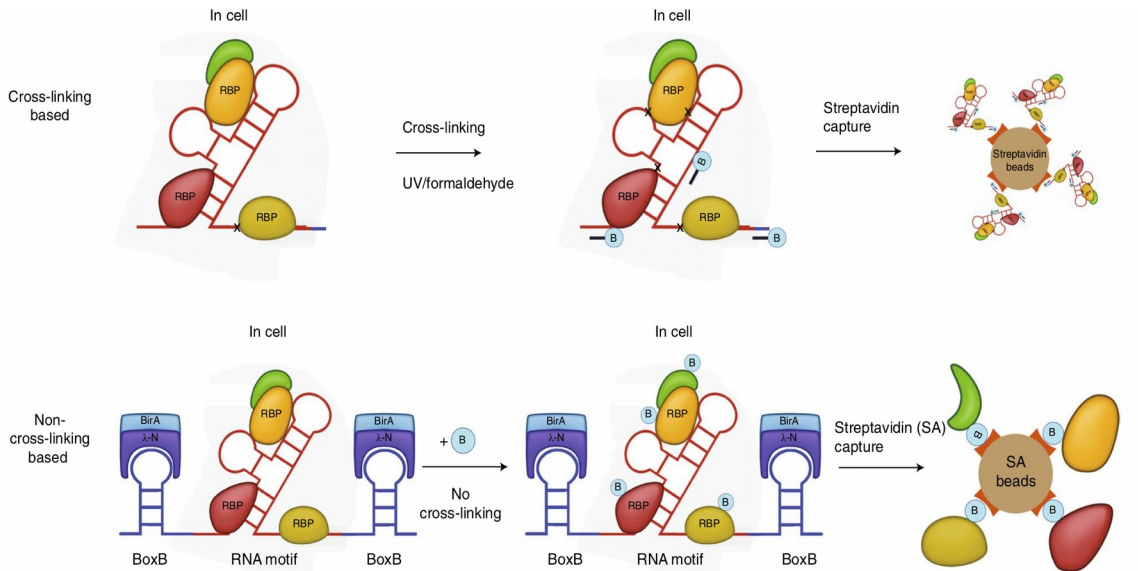


Figure 18: Approches "protein-centric" pour l'identification de RBPs in vivo. Méthodes faisant appel aux pontages chimiques (haut) et celles qui n'y font pas appel (bas). Tirée de Ramanathan et al., 2019.

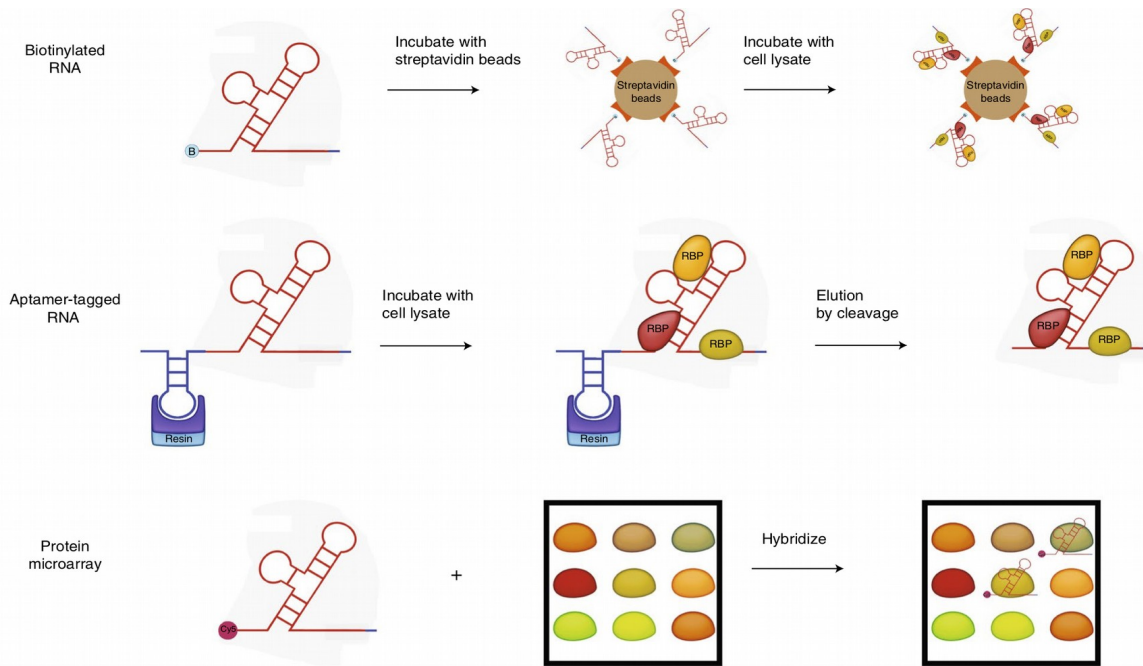


Figure 19: Approches "protein-centric pour l'identification de RBPs in vitro. Les différentes variantes de ces méthodes incluent l'utilisation : d'ARN biotinyllé (haut), d'apatamères « taggés » (milieu), ou de puces à protéine (bas). Tirée de Ramanathan et al., 2019.

### **3.2 Caractérisation des acides aminés interagissant avec l'ARN**

Les méthodes utilisées en conditions natives peuvent être utilisées pour identifier les sites d'interaction sur l'ARN (Corcoran et al., 2011). Mais elles sont difficiles à généraliser. Les méthodes qui utilisent la formation de ponts chimiques ("cross-links") entre ARN et protéine présentent l'avantage de pouvoir identifier les sites d'interaction aux positions de pontage par analyse de spectrométrie de masse (Kramer et al., 2014). La méthode catRAPID permet aussi de prédire les résidus de la protéine qui sont en contact avec la protéine mais assez peu de méthodes permettent ce type de prédiction. Bien que les méthodes à haut-débit permettent d'identifier les résidus de la protéine à l'interface (Kramer et al., 2014; Qamar, Kramer, & Urlaub, 2015), elles restent difficiles à mettre en œuvre. En parallèle, différentes méthodes *in silico* ont été développées avec ce même objectif (Cirillo, Agostini, & Tartaglia, 2013; Walia, EL-Manzalawy, Honavar, & Dobbs, 2017). catRAPID comme d'autres méthodes utilisent notamment les propriétés physico-chimiques des acides aminés et d'autres combinent ce type d'information avec des données phylogénétiques ou des données de similarité structurale. D'autres méthodes, comme PS-PRIP, utilisent directement les données extraites des motifs identifiés expérimentalement à l'interface protéine-ARN. Assez peu de méthodes utilisent en fait les informations de structures 3D.

Récemment, la banque de données InteracDome a été construite à partir de la compilation et de l'analyse des données structurales de ligands de protéines incluant notamment les ARN (Kobren & Singh, 2018). InteracDome répertorie les motifs de séquence des protéines qui interagissent avec tel ou tel ligand et prédit les résidus impliqués dans la liaison de ligands, notamment des ARN.

### **3.3 Caractérisation des séquences nucléotidiques reconnues par les RBPs**

Comme cela a été décrit dans le cas des méthodes "protein-centric" utilisées pour identifier des ligands ARN *in vivo*, de nombreuses données ont déjà été accumulées sur les motifs et leur contexte structural reconnus par les RBPs. Ces données sont disponibles dans les banques de données mentionnées comme POSTAR2 (Zhu et al., 2019), RNAct (Lang, Armaos, & Tartaglia, 2019) ou encore ATtRACT qui intègre des données expérimentales d'autres banques de données et des outils de recherche de motifs (Giudice, Sánchez-Cabo, Torroja, & Lara-Pezzi, 2016). Les approches dérivées de la technique CLIP avec ses variantes haut-débit peuvent comporter des biais dans les motifs les plus représentés qui ne sont pas nécessairement les motifs avec les meilleurs affinités de liaison (Friedersdorf & Keene, 2014). Des approches *in vitro* peuvent alors être utilisées pour confirmer les motifs identifiés (Fig. 20).

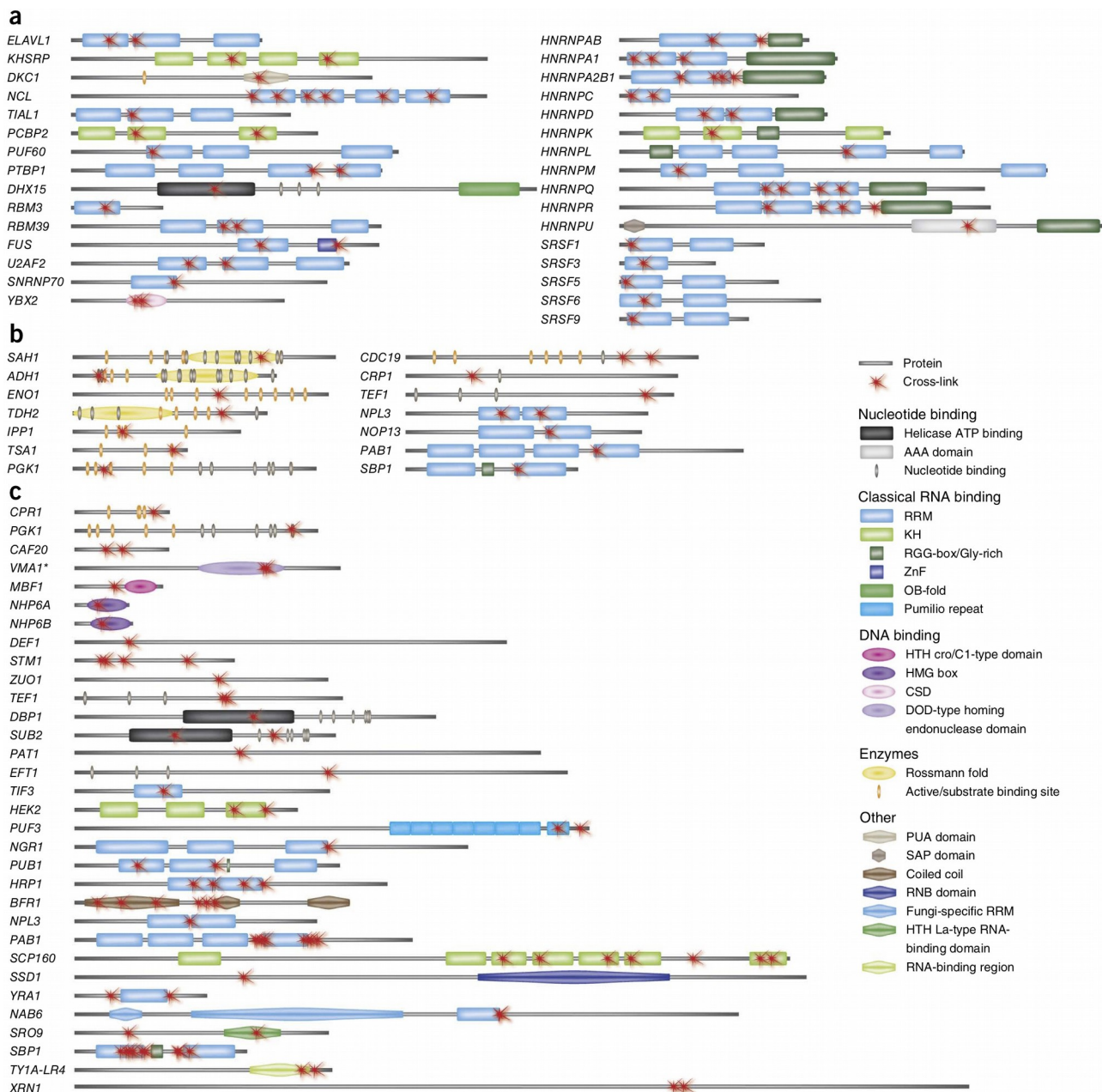


Figure 20: RBPs appartenant à différentes familles et leurs sites d'interaction avec l'ARN identifiés par pontage chimique. Tirée de Kramer et al. 2014.

Les approches *in vitro* comme SELEX (“Systematic evolution of ligands by exponential enrichment”, Fig. 21A), SEQRS (Campbell et al., 2012; Ellington & Szostak, 1990), RNAcompete (Fig. 21B) (Ray et al., 2009) et RNA Bind-n-Seq (Fig. 21C) (Lambert et al., 2014) ont également été largement utilisées pour l’identification de motifs spécifiques. La méthode SELEX, ou sa variante SEQRS à haut-débit, sont classiquement utilisées pour générer, par sélection *in vitro*, des ligands ARN (k-mers, k entre 20 et 40) à haute affinité ou aptamères. Ces aptamères correspondent à des ligands “optimaux” qui ne reflètent au mieux que partiellement les types de motifs reconnus

*in vivo*. L'approche RNAcompete a été conçue afin de surmonter ce biais en incubant la RBP d'intérêt avec une librairie exhaustive d'ARN (k-mers, k=9) et en n'exécutant qu'une seule étape de sélection. Elle met donc en compétition les ARN de la librairie et sélectionne les k-mers avec la meilleure affinité. Cette méthode a été appliquée à de nombreuses RBPs, en particulier celles des familles à domaines RRM et KH (Ray et al., 2013). Cette méthode est critiquée du fait de la courte taille des k-mers qui ne pourraient pas reproduire le contexte structural du motif de liaison dans l'ARN qui est lié par la RBP. RNA Bind-n-Seq repose sur le même principe en utilisant des librairies de plus longues séquences (k-mer, k=40).

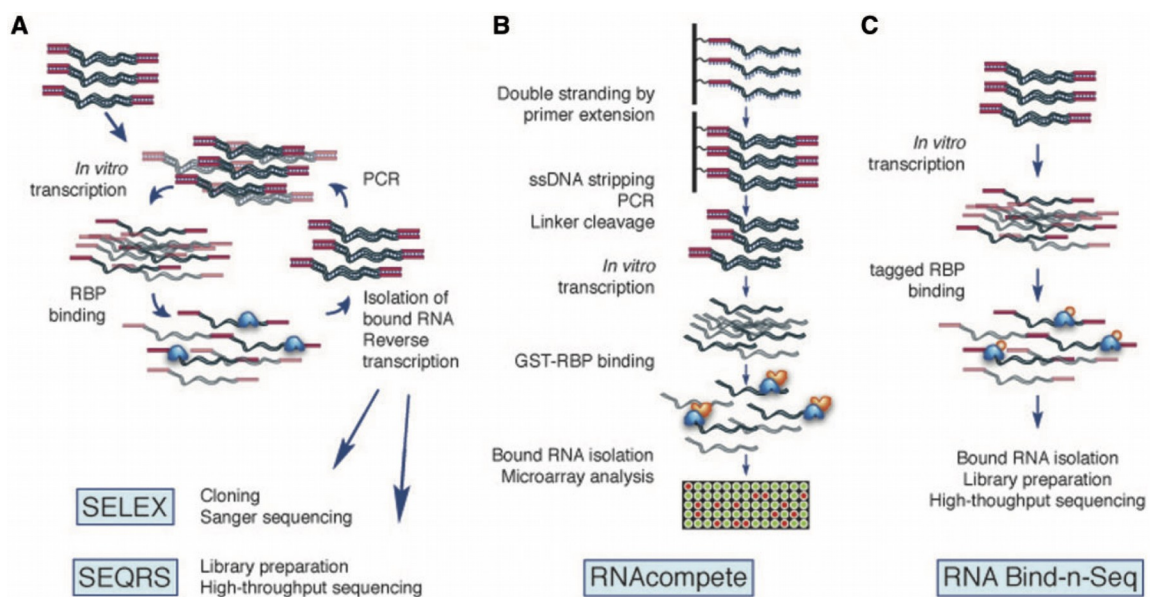


Figure 21: Approches "protein-centric" *in vitro* pour l'identification d'interactions protéine-ARN. **A.** Méthodes SELEX et SEQRS (SELEX avec séquençage à chaque cycle d'amplification). **B.** Méthode RNAcompete pour la liaison de séquences extraites d'une librairie prédéfinie en conditions compétitives et détection par puce. **C.** Méthode RNA Bind-n-Seq en présence de concentrations variables de protéine et séquençage à haut-débit. Tirée de Cook et al., 2015.

Plusieurs banques de données compilent des données de RNAcompete et SELEX avec d'autres sources de données : RBPDB (Cook, Kazan, Zuberi, Morris, & Hughes, 2010), CISBP-RNA (Ray et al., 2013), Deepbind (Alipanahi, Delong, Weirauch, & Frey, 2015); La banque de données généraliste AtTRACT (Giudice et al., 2016) reprend les données disponibles notamment dans CISBP-RNA et RBPDB et aussi dans PDB ("Protein Data Bank database"). L'ensemble de ces données permet d'avoir un panorama assez complet des motifs reconnus par telle ou telle RBP déjà identifiée et purifiée. Le thème de la spécificité de reconnaissance de séquences et motifs ARN est

largement débattu. Un point particulier nous intéresse, concernant les RBPs de liaison à l'ARN simple-brin : quelle est la nature des motifs/séquences reconnu(e)s ?

Dans la banque de données ATtRACT, les motifs de liaison aux RBPs correspondent à des k-mers de taille k compris entre quatre et huit dans 80% des cas chez l'Homme avec un pic correspondant à des motifs 7-mers (Fig. 22). Comme décrit ci-dessous, les techniques utilisées pour identifier ces motifs par les approches "protein-centric" *in vitro* peuvent induire des biais. Par exemple, les données RNAcompete utilisent habituellement des bibliothèques de 9-mers. Lorsque l'on examine les motifs reconnus par les RBPs à domaines RRM et KH déduits de données RNAcompete (Lang et al., 2019), ils correspondent généralement à des 7-mers (Fig. 23), ce qui est cohérent avec les données ATtRACT qui intègrent l'ensemble des informations issues de cette étude via CISBP-RNA. Une étude réalisée par SEQRS insiste sur le fait que les motifs 7-mers ne sont pas suffisamment longs pour définir une spécificité de liaison à de nombreuses RBPs (Campbell et al., 2012). Les cas étudiés (Puf4p, Puf5p, PUF-8, PUF-11, FBF-2, PUM2), incluent des motifs correspondant à des k-mers avec k entre huit et 10. Pour autant, les motifs en question ne sont pas des motifs consensus très contraints en terme de conservation de séquence. Ce sont des motifs avec au plus trois ou quatre positions conservées consécutives (Fig. 24). Les données issues des deux techniques, RNAcompete et de SEQRS, montrent que les motifs ARN sont généralement composites avec plusieurs sous-motifs de séquence conservées pouvant être séparées par plusieurs résidus. Cette caractéristique des sites de liaison des RBPs est confirmée par des travaux récents qui montrent que les motifs consensus sont des motifs généralement bi-partite voire tri-partite (Dominguez et al., 2018). Une autre étude sur les motifs de microARN met en évidence l'existence de nombreux sous-motifs correspondant à des 3-mers (Gao et al., 2018). Ces résultats sont cohérents avec la proposition de motifs bi- et tri-partites correspondant à des motifs dits: "3-mer cores" séparées par des espaceurs de longueur comprise entre 0 et 10 résidus (Dominguez et al., 2018).



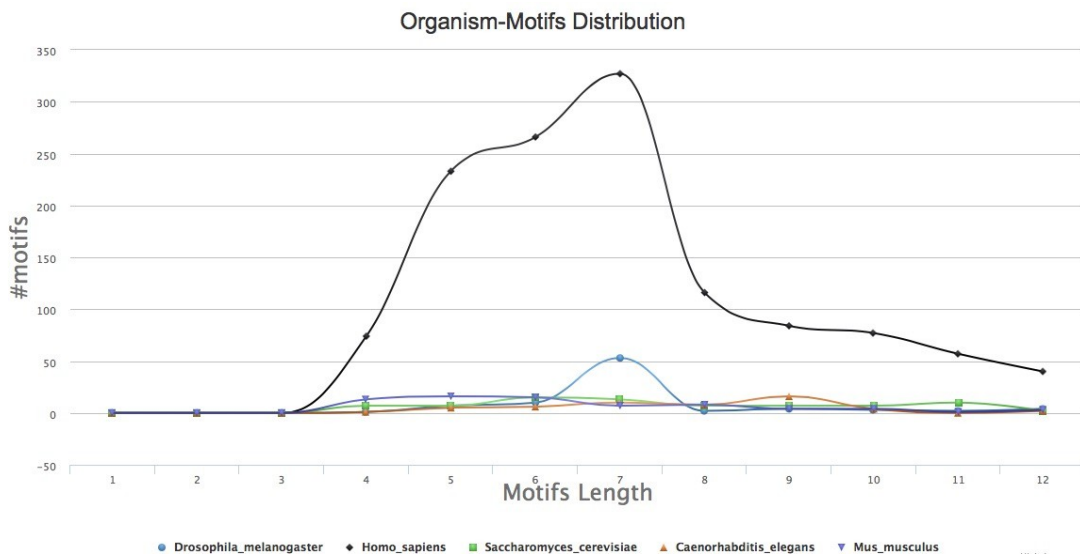


Figure 22: Distribution de tailles des motifs ARN reconnues par les RBP dans la banque de données ATtRACT. Générée à partir de la banque ATtRACT (<https://attract.cnrc.es/attract/default/stats>)

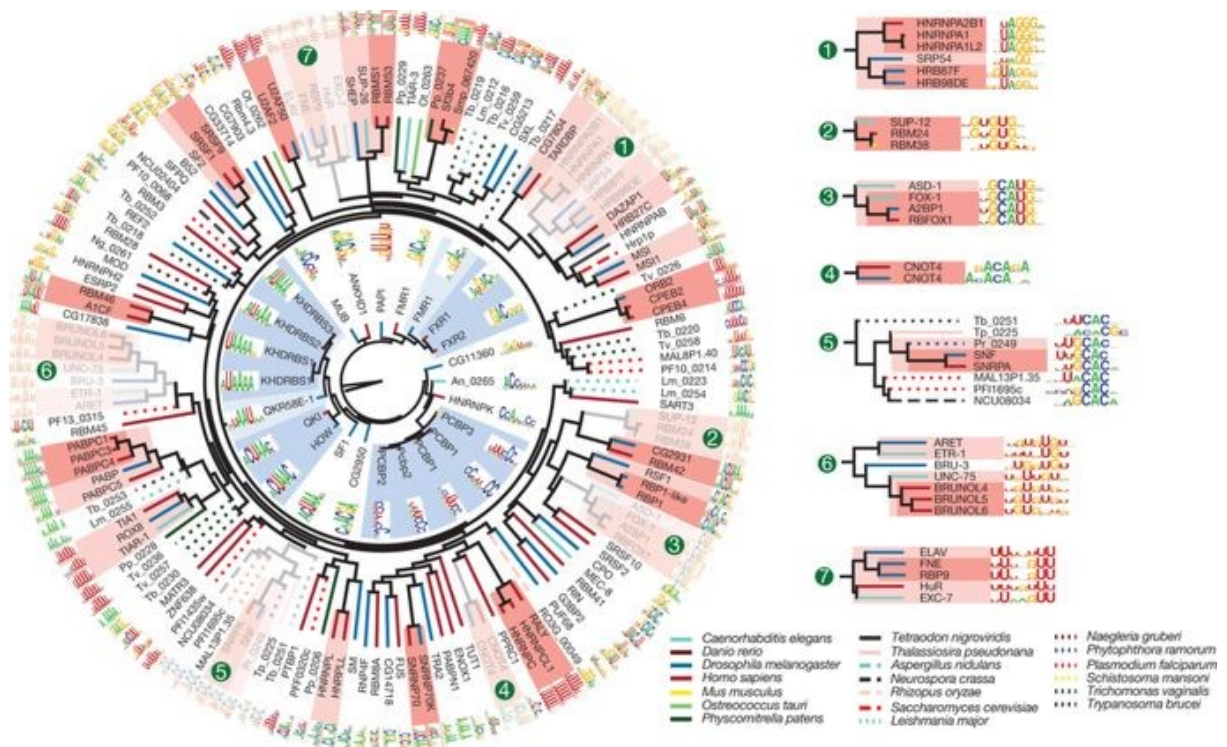


Figure 23: Motifs consensus identifiés par RNAcompete pour les RBPs à domaines RRM et KH. Extérieur du cercle : motifs reconnus par les RBPs à domaines RRM ; Intérieur: motifs reconnus par les RBPs à domaine KH. Tirée de Ray et al., 2013.

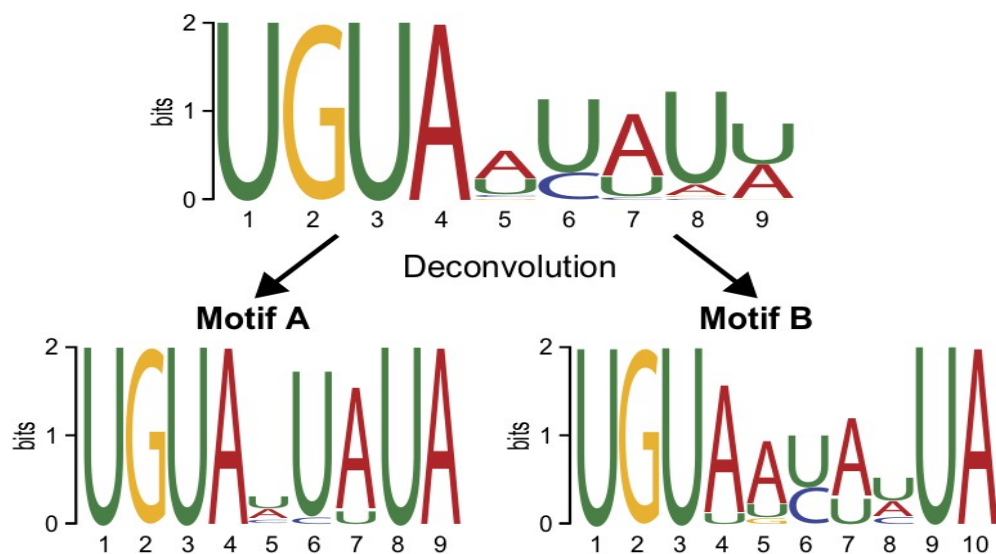


Figure 24: Motifs consensus identifiés par SEQRS pour la RBP *Puf5p* de levure. Motifs A et B de séquences consensus différents. Tirée de Campbell et al., 2012

Ces caractéristiques des sites de liaison des RBPs pour de l'ARN simple-brin peuvent être interprétées à la lumière de données structurales antérieures qui datent de 2006 (Auweter, Oberstrass, & Allain, 2006). Les contacts établis entre l'ARN et les RBDs de différentes familles (RRM, KH, doigts à zinc (CCCH, CCHC), Pumilio) sont concentrés sur des segments de trois à cinq résidus consécutifs. Des données structurales plus récentes abondent dans le sens de sites de liaison bi- et tri-partite (Afroz et al., 2015). Seules les données structurales permettent d'avoir une indication précise du site de liaison primaire de la RBP et de l'interface protéine-ARN.

### 3.4 Caractérisation tridimensionnelle des complexes protéine-ARN

La première structure 3D, déposée dans la "Protein Data Bank database" ou PDB (Berman et al., 2000), d'un complexe protéine-ARN date d'une trentaine d'années (Z. G. Chen et al., 1989). Depuis, beaucoup de structures de complexes protéine-ARN ont été déterminées par différentes méthodes grâce à des avancées technologiques qui ont permis de faciliter leur purification et la détermination de leur(s) structure(s) 3D. Il faut souligner aussi la capacité progressive de ces méthodes à traiter des ARN et complexes protéine-ARN de tailles grandissantes jusqu'aux complexes ribonucléoprotéiques tels que le ribosome ou le spliceosome. En dehors de la radiocristallographie et de la RMN dont les avancées techniques ont joué un rôle important, l'apparition de nouvelles approches basées sur la microscopie électronique ont aussi permis d'accumuler davantage de données structurales. Les plateformes de génomique structurale ont contribué de façon significative à l'augmentation des données structurales sur ce type de complexe



(S. Jones, 2016). En 2015, 1809 structures de complexes protéine-ARN étaient recensées dans PDB (Fig. 25). Par rapport à l'ensemble des protéines de liaison à l'ARN identifiées, annotées et disponibles dans la PDB (plus de 5600), les complexes protéine-ARN ne représentent que 31% de l'ensemble, soit 69% de RBPs pour lesquelles on dispose d'une structure 3D sans ARN, dont l'interface avec l'ARN et les sites d'interaction sont donc indéterminés. Dans ce contexte, les méthodes de modélisation 3D via le "docking" essaient de prédire la structure de complexes protéine-ARN à partir de données structurales partielles ou complètes des deux partenaires. Ces méthodes et leur application aux complexes protéine-ARN sont décrites plus loin.

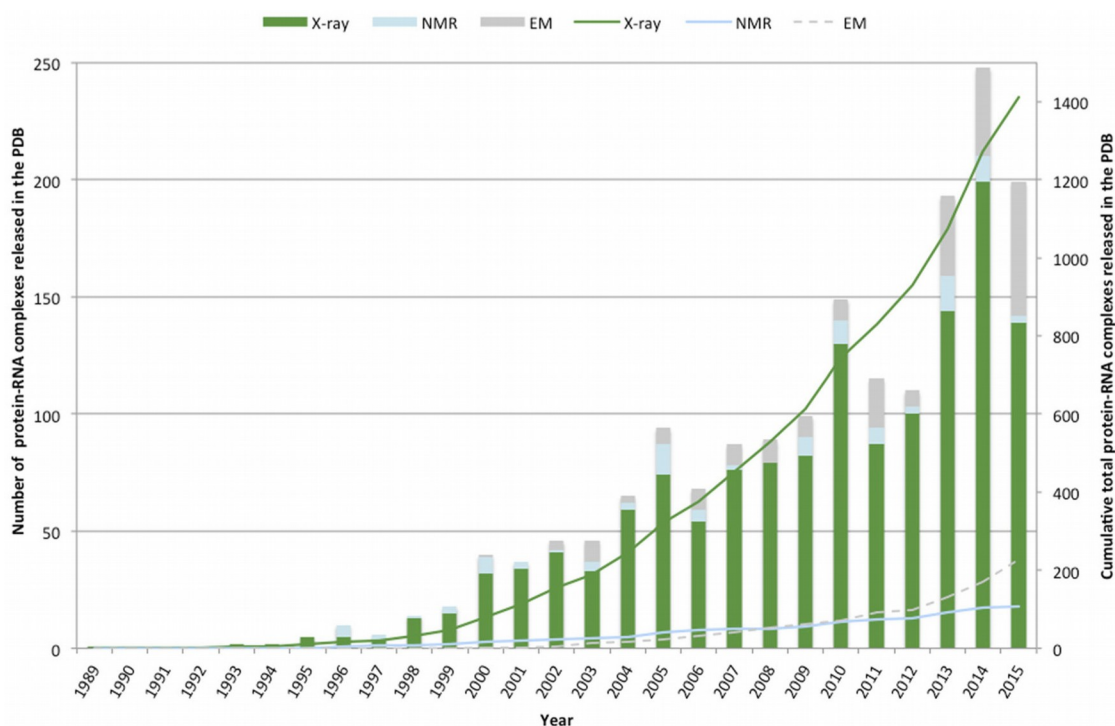


Figure 25: Données statistiques des complexes protéine-ARN disponible dans PDB jusque 2015. L'histogramme détaille la répartition des structures 3D en fonction des méthodes utilisées: diffraction des rayons-X (X-ray), RMN (NMR), microscopie électronique (EM). Tirée de Jones, 2016.

### 3.5 Code de reconnaissance protéine-ARN ?

Y-a-t'il un code qui définit les règles de reconnaissance entre un ARN et une protéine ? La question a été débattue dans le cas des interactions ADN-protéine en particulier sur la question de la reconnaissance de l'ADN par les facteurs de transcription où l'on exclut les structures non-canoniques de l'ADN (ADN simple-brin, G-quartet, jonctions ADN, etc). Dans les structures canoniques d'ADN double-brin, la diversité conformationnelle est surtout liée à la forme de l'ADN (A, B, Z) qui détermine notamment l'accessibilité des petits et grands sillons pour des interactions

avec des résidus d'acides aminés. Des contacts “types” dans l'un ou l'autre des sillons sont retrouvés dans beaucoup d'interactions ADN-protéine et on peut imaginer qu'il existe des patrons récurrents d'interactions à un niveau supérieur d'organisation. Pourtant, l'identification d'un code spécifique n'est pas triviale. En 2002, Benos et al. concluaient à la validité d'un code probabiliste de reconnaissance ADN-protéine applicable dans la majorité des cas (Benos, Lapedes, & Stormo, 2002).

La diversité conformationnelle de l'ARN ajoute un degré supérieur de complexité dans les régions simple-brin alors que les interactions entre les régions double-brin de l'ARN et les domaines RBD des protéines sont assez proches des interactions avec l'ADN en terme de codification. Comme cela a été décrit ci-dessus, les méthodes d'étude des interactions protéine-ARN “protein-centric” ou “RNA-centric” ont permis, notamment grâce aux approches à haut-débit, d'accumuler beaucoup de données et d'informations. Les apports combinés de ces approches avec ceux de la biologie structurale permettent de poser la question du code de reconnaissance (Hennig & Sattler, 2015). Des motifs consensus sont maintenant identifiés par les méthodes à haut-débit pour les domaines des grandes familles de RBPs (RRM, KH, Pumilio, Zn-CCCH, Sm-like, etc) même si certaines d'entre elles ont aussi la capacité de reconnaître des motifs différents en fonction du contexte. Les données structurales ont révélé la façon dont sont agencés les différents domaines de liaison à l'ARN et leur flexibilité. Pris individuellement, les domaines KH homologues (KH1, KH2, KH3) reconnaissent des motifs courts (trois à quatre résidus) assez conservés (Auweter et al., 2006).

Le contexte de ces interactions joue un rôle important. Pour la protéine, il s'agit de l'agencement des différents domaines de liaison à l'ARN et de la flexibilité inter-domaine qui permet aux protéines d'adopter un mode de reconnaissance par ajustement induit où la conformation liée diffère de façon significative de la conformation non-liée (Afroz et al., 2015). Pour l'ARN, il s'agit du contexte du motif et de son accessibilité. Le contexte de motifs ARN interagissant avec des protéines a été formalisé dans la banque de données “RNA Bricks” (Chojnowski, Walen, & Bujnicki, 2014). La flexibilité de l'ARN joue bien sûr aussi un rôle dans l'accessibilité des motifs présents dans des régions partiellement structurées mais dynamiques ; précisons également les RBPs pourraient aussi induire un changement de conformation de l'ARN. Des données récentes obtenues par des approches à haut débit suggèrent que les ARN seraient beaucoup moins structurés *in vivo* que ne le laissent penser les données *in vitro* et les prédictions *in silico* (Rouskin, Zubradt, Washietl, Kellis, & Weissman, 2014). Les ARN messagers synthétisés dans des conditions de stress seraient aussi significativement plus riches en régions simple-brin (Ding et al., 2014). Ceci suggère

un rôle potentiel des protéines à grande échelle dans la (re-)structuration et le remodelage des ARN dans la cellule. Tous ces éléments contribuent à la complexité des interactions protéine-ARN. L'évaluation de leur spécificité en font un défi majeur pour la compréhension des réseaux d'interaction.

### III Les principes du Docking

La cristallographie aux rayons X (X-ray) et la résonance magnétique nucléaire (RMN) sont deux techniques majeures permettant de décrire à un niveau de résolution atomique comment deux molécules interagissent. Ces méthodes expérimentales ont toutefois chacune leurs limitations et restent difficiles et coûteuses à mettre en place. L'amarrage moléculaire, ou docking, est une approche informatique représentant une alternative complémentaire à ces méthodes expérimentales. Le docking est globalement utilisé pour générer des modèles permettant de prédire le mode d'interaction entre deux molécules à partir de leurs coordonnées atomiques. Les structures tridimensionnelles des molécules d'intérêt sont donc nécessaires; elles peuvent provenir des approches expérimentales (X-ray ou RMN) ou bien, en leur absence, de modèles construits par homologie. Le docking peut se décliner en deux étapes successives (bien qu'elles soient souvent corrélées) : premièrement, l'échantillonnage des positions, orientations et conformations (ou plus simplement poses) d'un ligand à la surface d'un récepteur ; ensuite, le tri de ces poses grâce à une fonction de score qui leur associe une énergie d'interaction estimée. Idéalement, l'algorithme d'échantillonnage (ou sampling) doit pouvoir générer au moins une pose reproduisant le mode d'interaction expérimental ; la fonction de score doit être capable de lui associer une énergie d'interaction (scoring) permettant de la discriminer parmi l'ensemble des poses générées. Les différentes stratégies généralement utilisées pour les étapes de sampling et de scoring sont brièvement présentées ci-dessous. Les présentations traitent le cas d'un récepteur protéique et d'une petite molécule (par exemple un nucléotide) comme ligand.

#### 1 L'étape d'échantillonnage

L'étape d'échantillonnage a pour fonction de générer un ensemble de poses d'un ligand à la surface (circonscrite ou non) d'une protéine. Trois catégories de docking peuvent être établies selon la manière dont la flexibilité des molécules est traitée au cours de l'échantillonnage :

1. Le docking rigide, au cours duquel la protéine et le ligand sont tous deux traités comme entièrement rigides. Ainsi, seuls les degrés de liberté translationnels et rotationnels du ligand relativement au récepteur sont explorés. Cette simplification s'apparente à considérer un modèle de liaison de type "serrure-clé" où ni le ligand ni le récepteur ne subissent de réarrangements conformationnels suite à leur interaction. Le docking rigide est généralement employé pour l'amarrage entre deux macromolécules (docking protéine-protéine ou protéine/acide nucléique structuré) pour lesquelles les degrés de

liberté sont trop importants pour réaliser un échantillonnage conformationnel efficace dans des temps de calculs raisonnables.

2. Le docking semi-flexible, où seule la flexibilité du ligand est traitée, le récepteur restant rigide. Ainsi, l'échantillonnage des degrés de liberté du ligand s'ajoutent aux explorations translationnelles et rotationnelles. Ce type de docking repose sur l'hypothèse sommaire que la conformation du récepteur utilisée est apte à reconnaître le ligand.
3. Le docking flexible, qui considère à la fois la flexibilité du ligand et du récepteur. Les degrés de liberté conformationnels de ce dernier peuvent être limités à certaines chaînes latérales ou bien considérés également des mouvements plus larges impliquant par exemple les différents arrangements possibles entre domaines d'une protéine. Ce type de docking représente une approche plus réaliste en considérant des modèles de liaison du ligand au récepteur de type "ajustement induit" et/ou "sélection conformationnelle". En revanche, l'exploration de l'ensemble des degrés de liberté du ligand et du récepteur implique des difficultés importantes liées au temps de calculs nécessaires pour exploiter le docking flexible efficacement.

Les approches de docking semi-flexible ou flexible sont plus généralement appliquées à la modélisation d'interaction entre protéines et petites molécules. Les paragraphes qui suivent décrivent d'une manière non-exhaustive des stratégies généralement utilisées pour traiter la flexibilité du ligand et de la protéine.

## **1.1 Traitement de la flexibilité du ligand**

Les algorithmes d'échantillonnage peuvent être classés en différents groupes selon la méthodologie utilisée pour explorer la flexibilité du ligand : les algorithmes de recherche systématique, stochastique et les approches de simulation.

### **1.1.1 Recherche systématique**

Les algorithmes de recherche systématique explorent l'ensemble des degrés de liberté d'un ligand. Certaines approches systématiques sont dites exhaustives dans la mesure où elles visent à explorer l'ensemble des conformations possibles d'un ligand. Pour cela, chaque angle dièdre du ligand est incrémenté de façon discrète, et toutes les combinaisons d'angles autorisées sont échantillonnées. Si cette approche a le mérite d'évaluer l'ensemble des conformères possibles, leur quantité croît rapidement lorsque le nombre de degrés de liberté d'un ligand augmente. Par exemple, un échantillonnage des angles par intervalles de 30° conduit à 12 possibilités pour chaque

angle dièdre. Ainsi, le nombre de conformères à parcourir est de  $12^n$  pour une molécule à  $n$  degrés de liberté. Pour rendre la recherche exhaustive plus pratique, certains programmes de docking comme Glide (Friesner et al., 2004) génèrent en amont une bibliothèque de conformères d'un ligand donné permettant de traiter implicitement sa flexibilité. Ces conformères sont ensuite filtrés selon des contraintes géométriques/chimiques avant d'être positionnés et orientés aléatoirement dans un site pré-défini de la protéine.

Un autre type de recherche systématique concerne les approches par construction incrémentale du ligand. Plusieurs stratégies existent, toutes reposant initialement sur la décomposition d'un ligand d'intérêt en un ou plusieurs fragments. Le premier algorithme de construction incrémentale, décrit et intégré dans le programme DOCK en 1986 (DesJarlais, Kuntz, Sheridan, Venkataraghavan, & Dixon, 1986), peut se définir comme une approche de type "*anchor and link*". Dans cette implémentation, un ligand est d'abord décomposé en ses fragments les plus rigides, ou ancrés, les autres parties plus flexibles du ligand étant ignorées. Après un docking indépendant de chacune des ancrés, les paires de fragments pouvant être connectés sont identifiées par des contraintes de distance. Dans une approche légèrement différente correspondant à un procédé de type "*anchor and grow*", Leach et Kuntz (Leach & Kuntz, 1992) ne docke qu'un seul fragment. La partie plus flexible du ligand est greffée *a posteriori* à l'ancre en échantillonnant l'ensemble de ses angles dièdres par un nombre néanmoins réduits d'intervalles de manière à prévenir une explosion combinatoire. Enfin, Hammerhead (Welch, Ruppert, & Jain, 1996) utilise une approche de type "*anchor and merge*". Un premier fragment rigide correspondant généralement à une extrémité du ligand et appelé *head* est d'abord défini et docké à l'intérieur d'un site de liaison. Les autres parties du ligand sont également fragmentées de manière à ce que toutes possèdent des portions communes. Un premier fragment est alors joint à l'ancre au niveau de leur chevauchement puis la molécule résultante est optimisée par minimisation. Le processus est répété séquentiellement pour les autres fragments.

### **1.1.2 Recherche stochastique**

Les algorithmes de recherche stochastique visent à réduire la complexité combinatoire en explorant l'espace conformationnel de manière aléatoire. Cela implique que, contrairement aux algorithmes de recherche systématiques, deux simulations indépendantes peuvent générer des résultats différents. De plus, la recherche n'étant pas exhaustive, le risque de ne pas générer de solutions optimales est d'autant plus important que la complexité du système d'intérêt augmente. Les algorithmes Monte-Carlo et évolutionnaires sont deux exemples d'approches stochastiques.

Les algorithmes de Monte-Carlo génèrent aléatoirement des mouvements du ligand (translation/rotation et/ou torsions de ses angles). Ces algorithmes sont généralement associés à un critère de Métropolis permettant d'accepter ou rejeter une pose en fonction d'une probabilité. La méthode est itérative et repose sur le principe suivant. Des mouvements sont appliqués à une pose initiale d'énergie  $E_0$  de manière à obtenir une nouvelle pose d'énergie  $E_1$ . Si  $E_1 < E_0$  alors la nouvelle pose est acceptée et une nouvelle itération peut commencer. En revanche, si  $E_1 \geq E_0$ , alors une probabilité d'acceptation  $P$  est calculée de la manière suivante :

$$P = e^{-\frac{E_1 - E_0}{kT}} \quad (1)$$

où  $k$  représente la constante de Boltzmann et  $T$  la température. Si  $P$  est supérieure à une valeur tirée aléatoirement et comprise entre 0 et 1, alors la nouvelle pose est acceptée, sinon elle est rejetée. Cette stratégie permet d'accepter des poses d'énergie moins favorable et offre donc l'avantage de visiter un large espace du paysage énergétique de liaison en autorisant le franchissement de barrières d'énergie. La hauteur des barrières pouvant être franchies est dépendante de la température définie : plus elle augmente, plus la probabilité d'accepter des poses moins favorable est importante. AutoDock (Goodsell & Olson, 1990) est un exemple de programme de docking qui tire parti de cette propriété pour simuler une approche de recuit simulé. La température initiale est définie à une certaine valeur, suffisamment élevée pour autoriser la génération de poses franchissant d'éventuelles barrières énergétiques. Elle est ensuite progressivement abaissée au cours des cycles suivants. Cette stratégie augmente les probabilités d'obtenir une pose correspondant à un minimum énergétique global.

Un autre exemple d'approche stochastique concerne les algorithmes évolutionnaires. Ces derniers cherchent des solutions optimales à un problème donné en s'inspirant de la théorie de l'évolution : les individus les mieux adaptés à leur environnement ont une plus grande probabilité de survie et donc de transmettre leurs caractères aux générations suivantes. Les algorithmes génétiques sont certainement les plus populaires parmi les approches évolutionnaires. La première étape est de générer une population d'individus. Dans le contexte d'un échantillonnage conformationnel intégré à une approche de docking, chaque individu correspond à un ligand dont la position, l'orientation et la conformation sont encodées dans un vecteur définissant son chromosome. Une fonction de score est utilisée pour mesurer la qualité d'adaptation de chaque individu par rapport à son environnement (la surface de la protéine). Une nouvelle génération est ensuite générée à partir de la population initiale. Un accouplement est opéré entre les chromosomes



en fonction de leur qualité d'adaptation : les individus les mieux adaptés (i.e. les conformations de meilleur score) présentent une plus grande probabilité d'accouplement. Des *crossing-over* sont également opérés à cette étape. La variabilité des individus est aussi enrichi par des mutations intrachromosomiques aléatoires. Leur probabilité d'occurrence est toutefois réduite pour les individus les mieux adaptés de manière à s'assurer qu'ils se maintiennent au cours de l'évolution. Globalement, un algorithme génétique autorise donc une exploration variée de l'espace possible des poses et augmente ainsi les chances d'obtenir une solution optimale. La taille de la population, les taux de *crossing-over* et de mutations, ainsi que le nombre de cycles de reproduction sont les paramètres importants influençant les résultats. La troisième version d'AutoDock (Morris et al., 1998) et Gold (G. Jones, Willett, Glen, Leach, & Taylor, 1997) sont deux exemples de programme de docking utilisant un algorithme génétique.

### 1.1.3 Approches de simulation

La minimisation énergétique ainsi que la dynamique moléculaire sont deux approches de simulation.

La minimisation est principalement utilisée pour optimiser la géométrie d'un système en l'amenant dans un minimum énergétique local. Cette stratégie ne permettant pas de franchir de grandes barrières énergétiques, elle est souvent utilisée dans des approches de docking en combinaison avec d'autres méthodes de recherche qui autorisent une exploration plus large de l'espace possible des poses.

La dynamique moléculaire permet de simuler l'évolution d'un système au cours du temps dans des conditions thermodynamiques (température, pression et volume) pré-définies. Cette approche autorise donc une modélisation plus réaliste, d'autant plus qu'elle offre la possibilité de traiter explicitement les molécules d'eau ainsi que l'ensemble des degrés de liberté du système, aussi bien ceux du ligand, de la protéine et du solvant. Les temps de calculs sont en contrepartie évidemment plus coûteux que les autres approches présentées plus haut, rendant son utilisation rapidement prohibitive pour une étude efficace de systèmes complexes comme peuvent l'être les molécules biologiques. Ainsi, les simulations de dynamiques ne peuvent être, par défaut, réalisées sur une durée suffisamment longue pour franchir des barrières énergétiques très élevées, réduisant l'exploration de l'espace possible des poses. La dynamique moléculaire est en conséquence, comme la minimisation, le plus souvent utilisée en conjonction avec d'autres approches de docking, notamment pour évaluer la stabilité de poses candidats (Yu et al., 2018) et/ou les affiner (Rastelli, Degliesposti, Del Rio, & Sgobba, 2009).

## **1.2 Traitement de la flexibilité de la protéine**

De nombreuses données structurales (X-ray et RMN) mettent en évidence des différences conformationnelles entre la forme d'une protéine liée (holo-) à un ligand et sa forme libre, non-liée (apo-). Beaucoup d'approches de docking à grande échelle sont toutefois menées à partir d'une unique structure rigide de la protéine en raison de l'augmentation des temps de calculs associés à la prise en compte de ses degrés de liberté. Plusieurs stratégies sont néanmoins développées pour considérer leur flexibilité et ainsi rendre compte de modèles plus réalistes. Elles peuvent se diviser en deux classes : celles reposant sur une seule structure protéique, et celles exploitant plusieurs conformations d'une même protéine.

### **1.2.1 Approches à conformation protéique unique**

De subtils changements de positions des résidus de la protéine peuvent avoir une grande influence sur les résultats de docking. Le réarrangement d'une seule chaîne latérale peut par exemple modifier la topologie du site de liaison et décroître la précision dans la prédiction de modes d'interaction. Différentes approches de docking ont donc été développées afin de considérer ces réarrangements locaux, s'approchant ainsi d'un modèle de liaison de type ajustement induit. Elles reposent sur l'utilisation d'une unique conformation protéique comme point de départ, son squelette étant considéré rigide.

Le *soft* docking est une approche qui a été décrite pour la première fois en 1991 (Jiang & Kim, 1991). Le qualificatif *soft* fait référence à l'atténuation du terme répulsif du potentiel de Lennard-Jones employé dans la fonction de score. La réduction de ce terme autorise de légers chevauchements stériques entre atomes reflétant ainsi une certaine incertitude relative à leurs coordonnées atomiques. Ce type d'approche traite donc la flexibilité d'une manière implicite. Elle présente l'avantage d'être simple à mettre en place mais peut conduire à l'obtention de poses irréalistes (Vieth, Hirst, Kolinski, & Brooks, 1998).

Une autre stratégie consiste à explorer explicitement les conformations des chaînes latérales. Leur échantillonnage repose généralement sur des bibliothèques de rotamères (Dunbrack & Karplus, 1993; Tuffery, Etchebest, Hazout, & Lavery, 1991) dont l'utilisation peut être associée à une étape d'optimisation de la complémentarité stérique entre le ligand et la protéine (Källblad & Dean, 2003; Leach, 1994).

### **1.2.2 Approches à multiples conformations protéiques**

Les approches de docking n'utilisant qu'une seule conformation protéique permettent uniquement de traiter des mouvements locaux. L'exploitation de plusieurs conformations protéiques

offrent le potentiel de considérer de plus amples réarrangements conformationnels. Ces conformations peuvent provenir de structures résolues expérimentalement ou, en leur absence, de simulations de dynamique moléculaire, de Monte-Carlo ou encore par l'analyse de modes normaux. Plusieurs stratégies ont été envisagées pour intégrer dans le docking l'usage de plusieurs conformations protéiques. On peut distinguer parmi celles-ci les approches reposant sur l'utilisation de grilles pondérées, celles générant des protéines chimères moyennes, et celles utilisant séquentiellement un ensemble de conformations.

Certains programmes de docking calculent préalablement des grilles cartographiant l'ensemble des interactions possibles. Les énergies d'interaction entre chaque atome d'un ligand donné et les atomes d'un récepteur sont pré-calculées et stockées dans une grille. Cette méthode présente l'avantage de réduire les temps de calculs puisque les énergies d'interaction n'ont plus à être calculées à chaque étape du docking. Certaines approches exploitent l'utilisation de grilles pour considérer un ensemble de conformations d'un récepteur protéique. Une grille moyenne peut par exemple être calculée à partir de l'ensemble des structures disponibles. La pondération peut être basée sur l'énergie d'interaction ou sur la position des atomes du récepteur (Knegtel, Kuntz, & Oshiro, 1997).

Un ensemble de conformations peut également être utilisé pour construire une ou plusieurs protéines chimères. Par exemple, FlexE (Claußen, Buning, Rarey, & Lengauer, 2001) définit une structure moyenne rigide à partir des régions structurellement similaires entre les différentes conformations. Les portions montrant plus de variabilité sont ensuite fusionnées à cette structure moyenne d'une manière combinatoire. L'ensemble des protéines chimères qui en résultent est alors utilisé pour le docking.

Enfin, une stratégie qui a également été envisagée consiste plus simplement à réaliser un docking individuel sur chacune des conformations d'un ensemble. Pour optimiser les temps de calcul, une approche de ce type a été réalisée de manière à ce que pour chaque position et orientation échantillonnées pour un ligand, ce dernier soit optimisé pour l'ensemble des conformations, permettant de sélectionner la meilleure d'entre elle sur la base de l'énergie d'interaction associée (S. Y. Huang & Zou, 2007).

## 2 Les fonctions de score

### 2.1 Principes physico-chimiques de la reconnaissance protéine-ligand

De manière générale, les fonctions de score utilisées pour le docking donnent une estimation de l'affinité de liaison entre un ligand (L) et son récepteur (R). L'affinité peut être mesurée expérimentalement en déterminant la constante d'association à l'équilibre ( $K_{eq}$ ) qui représente le rapport des concentrations entre le complexe récepteur-ligand (RL) et les formes libres de R et L lorsque la réaction a atteint l'équilibre. Par ailleurs, la constante  $K_{eq}$  est directement reliée à la variation d'énergie libre de Gibbs ( $\Delta G$ ) qui peut aussi être décrite par les variations des contributions enthalpique ( $\Delta H$ ) et entropique ( $\Delta S$ ) :

$$\Delta G = -RT \ln K_{eq} = \Delta H - T \Delta S \quad (2)$$

où R est la constante universelle des gaz parfaits et T la température.

Dans des conditions de température et de pression constantes et lorsque le système a atteint l'équilibre, une variation négative de  $\Delta G$  correspond à un processus spontané d'association. L'équation 2 montre que la magnitude de  $\Delta G$  est déterminée par la constante  $K_{eq}$ . On peut donc considérer que la valeur de  $\Delta G$  reflète l'affinité de liaison entre un récepteur et un ligand, ou encore la stabilité du complexe RL. L'équation 2 montre également que  $\Delta G$  résulte des contributions enthalpique et entropique intervenant dans le processus de liaison. Dans des conditions où la pression est constante et la variation de volume du système ne varie pas, la variation d'enthalpie peut être définie comme la variation de l'énergie totale du système faisant suite à l'association d'un ligand à son récepteur. Elle découle d'une balance énergétique résultant de la formation et de la rupture d'un ensemble d'interactions. Une liaison enthalpiquement favorable peut ainsi mettre en jeu une perte de liaisons hydrogènes et d'interactions de van der Waals/électrostatiques entre chacun des solutés et le solvant, perte qui doit être compensée par la formation de nouvelles interactions non covalentes établies entre la protéine et le ligand. L'énergie interne des solutés intervient également dans la contribution enthalpique puisque les conformations de la protéine et du ligand peuvent être davantage contraintes dans l'état du complexe RL et donc énergétiquement moins favorables qu'elles ne peuvent l'être dans leurs formes libres respectives. L'équation 2 montre qu'un  $\Delta H$  négatif (*i.e.* favorable) conduira à un processus d'association spontané uniquement si cette variation n'est pas compensée par une variation d'entropie défavorable.

L'entropie représente une mesure du désordre qui, dans le contexte d'un système simple composé d'une protéine et d'un ligand baignant dans un milieu aqueux, peut être interprété par le

nombre d'états que peut adopter chacune des entités moléculaires du système. Les états possibles d'une molécule intègrent par exemple l'ensemble de ses conformations potentielles, ses possibilités de translations et rotations, ou encore les fréquences de vibrations de ses atomes. Lorsqu'un ligand se fixe à un récepteur, le nombre de ses états possibles réduit drastiquement, ce qui entraîne une forte pénalité entropique. Un phénomène important permet cependant de contrebalancer cette dernière : l'effet hydrophobe. Pour saisir son action, considérons une molécule hydrophobe (ou non polaire). Ses propriétés physico-chimiques ne lui permettent pas d'établir de liaisons hydrogènes (ou d'interactions électrostatiques "fortes") avec des molécules d'eau. Son introduction dans un milieu aqueux tend à rompre une partie du réseau dynamique de liaisons hydrogènes que forment les molécules d'eau entre elles. Les molécules d'eau au voisinage direct d'une surface hydrophobe ont toutefois tendance à s'orienter de manière à maximiser leur potentiel à établir des liaisons hydrogènes avec les molécules d'eau environnantes. Cet agencement tend à réduire leur mobilité (Ball, 2008; Noel T. Southall, Ken A. Dill, & Haymet\*, 2001) et entraîne donc une augmentation de l'ordre qui est entropiquement pénalisante. Par ailleurs, plusieurs études suggèrent que les molécules d'eau entourant un composé non polaire tendent à former moins de liaisons hydrogènes (Ji, Ostroverkhov, Tian, & Shen, 2008; Richmond, 2001), conférant également une pénalité enthalpique à son énergie de solvation. L'effet hydrophobe désigne le processus amenant des composés non polaires à spontanément s'associer en milieu aqueux. Le caractère spontané de cette réaction résulte, notamment, d'une diminution de la surface nette hydrophobe solvatée suite à l'association, cette dernière s'accompagnant d'une libération des molécules d'eau bénéfique sur le plan entropique. Ces dernières peuvent alors à nouveau participer au réseau de liaisons hydrogènes formé par les molécules d'eau environnantes apportant ainsi une contribution enthalpique favorable. La complémentarité stérique entre les régions non polaires contribue également à leur agrégation en favorisant des interactions de van der Waals.

L'effet hydrophobe représente un facteur majeur de stabilité des assemblages biologiques. Il est en effet essentiel au repliement des protéines et des acides nucléiques, à la formation des membranes lipidiques, et évidemment aux interactions intermoléculaires. Plusieurs travaux ont par exemple montré que l'affinité d'un ligand envers son récepteur peut être augmentée par l'ajout de groupements hydrophobes (Houk, Leach, Kim, & Zhang, 2003).

Les fonctions de score représentent des modèles mathématiques utilisés pour estimer une énergie d'interaction entre un récepteur et un ligand. Dans une approche de docking réalisée pour prédire leur mode d'interaction, les fonctions de score doivent pouvoir permettre de distinguer, parmi un ensemble varié de poses, celles qui sont pertinentes. Par ailleurs, dans les approches de criblage

virtuelle où une bibliothèque de ligands diverses est dockée sur une protéine, elles doivent aussi être capables d'identifier quels sont ceux qui peuvent réellement s'y lier et même pouvoir les classer selon leur affinité respective. Le nombre de poses à évaluer au cours de calculs de docking peut varier de plusieurs milliers à plusieurs millions selon la stratégie d'échantillonnage adoptée. Par conséquent, les fonctions de score doivent aussi être rapides. L'efficacité de leurs temps de calculs ne peut cependant être obtenue sans approximations ou simplifications du modèle décrivant le processus de liaison, et s'atteint donc au détriment de la précision dans l'évaluation de l'énergie d'interaction. De nombreuses fonctions de score ont été développées dans le but de répondre à ces contraintes d'efficacité et de précisions. Elles peuvent se classer en quatre grandes catégories : les fonctions de score basées sur champ de force, les fonctions empiriques, les fonctions reposant sur des potentiels statistiques et les fonctions basées sur des techniques d'apprentissage automatique.

## 2.2 Fonctions de score basées sur champ de force

En mécanique moléculaire (MM), un champ de force correspond à un ensemble de paramètres et de fonctions permettant de définir un système et de décrire son paysage d'énergie potentielle. Les paramètres inclus classiquement la masse des atomes, leur charge, leur rayon de van der Waals ainsi que différentes valeurs de référence correspondant à des longueurs de liaisons inter-atomiques et d'angles plans et dièdres. Ces paramètres sont généralement dérivés d'expériences ou de simulations quantiques habituellement réalisées sur de petites molécules organiques. Les fonctions d'un champ de force correspondent à des formalismes mathématiques qui intègrent ces paramètres et qui sont utilisés pour calculer plusieurs types d'énergie potentielle. AMBER (Weiner, Kollman, Nguyen, & Case, 1986) et CHARMM (Brooks, Brucoleri, Olafson, States, Swaminathan, et al., 1983) sont deux exemples de champs de force très utilisés pour la modélisation de biomolécules comme les protéines ou les acides nucléiques.

La fonction de score du programme de docking DOCK (Kuntz, Blaney, Oatley, Langridge, & Ferrin, 1982) est un exemple typique de fonction basée sur champ de force. Cette fonction de score, dont les paramètres sont dérivés d'AMBER, estime une énergie de liaison ( $E_{\text{liaison}}$ ) approximée par la somme d'énergies non-liées entre paires d'atomes du ligand et de la protéine : une énergie de van der Waals ( $E_{\text{vdw}}$ ) et une énergie électrostatique ( $E_{\text{elec}}$ ) :

$$\Delta G = E_{\text{vdw}} + E_{\text{elec}} \quad (3)$$

avec :

$$E_{\text{vdw}} = \sum_{i=1}^{\text{ligand}} \sum_{j=1}^{\text{protéine}} \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} \quad (4)$$

et :

$$E_{elec} = \sum_{i=1}^{ligand} \sum_{j=1}^{protéine} \frac{q_i q_j}{\epsilon(r_{ij}) r_{ij}} \quad (5)$$

où  $r_{ij}$  représente la distance séparant un atome  $i$  de la protéine d'un atome  $j$  du ligand,  $A_{ij}$  et  $B_{ij}$  sont les rayons de van der Waals des atomes,  $q_i$  et  $q_j$  leur charge respective, et  $\epsilon(r_{ij})$  une constante diélectrique dépendante de la distance.

Certaines fonctions de score, comme celle utilisée par MCSS pour les interactions protéine-nucléotide (décrite plus précisément dans la partie " Méthodes générales ", section 1.5), intègrent également des termes liés correspondant à une énergie interne. Les protocoles de docking classiques n'utilisant le plus souvent qu'une seule conformation protéique maintenue rigide, l'évaluation de l'énergie intramoléculaire est généralement restreinte au ligand. Cette énergie interne permet de considérer les contraintes induites sur la conformation du ligand faisant suite à sa liaison au récepteur.

Les liaisons hydrogène, importantes pour la reconnaissance spécifique entre molécules, ne sont qu'implicitement prises en compte par les termes coulombique et de van der Waals dans les fonctions de score suivant un formalisme similaire à celui de DOCK ou MCSS. Certaines fonctions de score comme celles des programme de docking Gold (Verdonk, Cole, Hartshorn, Murray, & Taylor, 2003) ou Autodock (Morris et al., 1998) incluent des termes permettant de considérer explicitement les liaisons hydrogène. Les différents termes incluent par la fonction de score de Gold dépendent de la nature des atomes mis en jeu et de la géométrie de l'interaction. La fonction de score d'AutoDock représente quant à elle les liaisons hydrogène par un potentiel de Lennard-Jones 12-10 couplé à une composante directionnelle.

Ces fonctions de score au formalisme très simple présentent l'avantage d'être relativement peu coûteuses en temps de calculs. Elles restent cependant très approximatives puisqu'elles n'estiment qu'une contribution enthalpique à l'énergie de liaison. Des facteurs importants intervenant dans l'affinité de liaison sont en effet négligés, comme l'effet du solvant ou les composantes entropiques.

L'effet du solvant joue un rôle essentiel, notamment en ayant un effet écran réduisant les forces d'interactions entre atomes chargés. Une mauvaise estimation de l'effet du solvant peut ainsi conduire à surestimer ce type d'interaction. Dans les approches classiques de docking, les molécules d'eau sont généralement retirées du système pour simplifier les calculs. L'effet du solvant est alors le plus souvent traité de manière implicite en intégrant dans le terme coulombique une constante diélectrique dépendante de la distance, comme le fait par exemple la fonction de score du



programme DOCK (Eq. 3). Ce traitement reste cependant une approximation très simpliste et peu réaliste. Des modèles plus rigoureux de solvant implicite ont été développés en considérant le solvant comme un milieu diélectrique continu. Les modèles de Poisson-Boltzmann (PB) et Generalized Born (GB) (Baker, 2005; Ghosh, Rapp, & Friesner, 2002) en sont deux exemples typiques, le second étant une approximation du premier et donc plus rapide. Pour une estimation se rapprochant d'une énergie libre de liaison (Eq. 2), les modèles PB/GB sont généralement combinés aux termes classiques de mécanique moléculaire (*e.g.* Eq. 3) ainsi qu'à un calcul de la variation d'accessibilité au solvant (SA) qui peut être perçu comme une estimation d'un facteur entropique lié à l'effet hydrophobe. Les modèles MM-PB/(GB)SA offrent un compromis intéressant entre temps de calculs et précision, mais sont néanmoins encore relativement coûteux pour être utilisés en routine dans les approches de docking. Ils sont le plus souvent employés dans une étape de post-traitement pour le ré-évaluation de l'énergie d'interaction des poses (Haider, Bertrand, & Hubbard, 2011).

Globalement, les fonctions de score basées sur champ de force reposent sur des modèles physiques théoriques. Cette caractéristique implique qu'elles sont indépendantes d'un apprentissage à partir de données d'affinités de liaison. Elles tendent en conséquence à bénéficier d'une transférabilité à divers systèmes plus importantes par rapport aux autres classes de fonctions de score. Les formalismes les plus simples de fonctions de score basées sur champ de force (*e.g.* Eq. 3) sont suffisamment rapides pour être appliquées à grande échelle. Cette rapidité découle de fortes simplifications pouvant entraîner des approximations pénalisantes dans l'estimation de l'énergie de liaison. Ces fonctions de score reposant sur des modèles physiques, elles peuvent bénéficier de représentations beaucoup plus précises (modèles de solvant, modèles quantiques) mais au prix de temps de calculs inappropriées pour des simulations de docking à grande échelle.

### **2.3 Fonctions de score empiriques**

Les fonctions de score empiriques estiment l'affinité de liaison  $\Delta G$  entre un couple récepteur-ligand sur la base d'un ensemble de descripteurs énergétiques  $\Delta G_i$ , chacun étant pondérés par un coefficient  $W_i$  :

$$\Delta G = \sum W_i \times \Delta G_i \quad (6)$$

Les descripteurs peuvent inclure différents termes énergétiques empruntés à la mécanique moléculaire (énergies de van de Waals, électrostatiques) mais aussi d'autres composantes idéalement non corrélées considérant par exemple l'hydrophobicité et/ou la polarité d'un site de liaison, l'accessibilité au solvant, l'entropie d'un ligand ou encore d'autres propriétés. Les

coefficients associés à chacun des termes énergétiques sont déterminés par des analyses de régressions linéaires de manière à optimiser la corrélation entre des énergies d'association estimées par la fonction de score et des affinités de liaison connues pour des complexes protéine-ligand dont la structure est résolue expérimentalement. Les fonctions de score empiriques estiment donc une affinité de liaison à partir d'un formalisme souvent bien plus simple que les fonctions de score basées sur champ de force incluant des modèles de solvant implicite (*e.g.* MM-PB/GBSA). Elles présentent ainsi l'avantage d'être plus rapides en raison de la faible complexité des termes à évaluer pouvant capturer implicitement des informations équivalentes. En revanche, l'apprentissage se faisant sur des jeux de données souvent réduits à une centaine de complexes tout au plus, la transférabilité de leur performance à des complexes différents du jeu d'entraînement reste problématique. L'augmentation croissante du nombre de structures de complexes et de leurs données d'affinité de liaison devraient néanmoins théoriquement permettre d'améliorer les fonctions de score empiriques en les rendant plus "universelles". LUDI (Böhm, 1992), ChemScore (Eldridge, Murray, Auton, Paolini, & Mee, 1997) ou GlideScore (Friesner et al., 2004, 2006) sont des exemples de fonctions de score empiriques. Elles diffèrent par leur nombre et types de descripteurs, le modèle de régression linéaire utilisé et les diversité et qualité des données d'affinité employées dans le processus de calibration.

## 2.4 Fonctions de score à potentiels statistiques

Les fonctions de score à potentiels statistiques sont basées sur une idée de la mécanique statistique permettant de dériver des potentiels de forces moyennes à partir de distributions de mesures observées entre paires d'atomes. Leur formulation générale est comme suit :

$$\Delta G = \sum_{i=1}^{ligand} \sum_{j=1}^{protéine} A_{ij}(r) \quad (7)$$

où  $\Delta G$  correspond à la somme des forces potentielles  $A$  entre les paires d'atomes  $i$  et  $j$  du ligand et de la protéine, respectivement. La force potentielle entre paires d'atomes  $i$  et  $j$  se calcule comme suit :

$$A_{ij} = -kT \ln \left( \frac{\rho_{ij}(r)}{\rho_{ref}} \right) \quad (8)$$

avec  $k$  la constante de Boltzmann,  $T$  la température,  $\rho_{ij}$  la densité de distribution des distances  $r$  séparant les atomes  $i$  et  $j$ , et  $\rho_{ref}$  une densité des distributions des distances  $r$  séparant la même paire d'atomes dans un état de référence où l'énergie d'interaction de la paire est supposée être nulle. Les distributions de distances entre différentes paires d'atomes sont obtenues à partir de structures

de complexes résolus expérimentalement. D'après l'équation 8, on peut voir que si des contacts entre une paire d'atomes donnée apparaît plus souvent que dans un état de référence, alors ces contacts auront une énergie d'interaction favorable. Les fonctions de score à potentiel statistique partagent avec les fonctions empiriques le fait d'essayer de capturer implicitement des facteurs intervenant dans le processus de liaison qui sont difficiles à modéliser explicitement. Elles sont également rapides en terme de temps de calculs au regard de la simplicité de leur formalisme. Contrairement aux fonctions empiriques, les fonctions à potentiels statistiques ne nécessitent pas d'informations de données d'affinité de liaison et peuvent donc profiter de données d'entraînement bien plus importantes les rendant potentiellement plus robustes face à des complexes diversifiés (Muegge & Martin, 1999). Leur précision est néanmoins dépendante de la définition d'une distribution de référence dont l'établissement constitue une difficulté importante puisque difficilement accessible pour des systèmes aussi complexes que des protéines (Thomas & Dill, 1996). DrugScore (Velec, Gohlke, & Klebe, 2005), PMF (Muegge & Martin, 1999) et ITScore (S. Y. Huang & Zou, 2008) sont des exemples de fonctions de score à potentiels statistiques différant, notamment, dans leur stratégie utilisée pour définir une distribution de référence.

## **2.5 Fonctions de score par méthodes d'apprentissage**

Les fonctions de score discutées jusqu'ici reposent toutes sur un formalisme linéaire et pré-définie qui se traduit par la somme de leurs différents termes. Ce principe d'additivité ne représente cependant qu'une approximation de la manière dont les différentes composantes énergétiques se manifestent au cours du phénomène d'association d'un ligand à son récepteur et ne peut rendre compte, par exemple, d'événements coopératifs (Dill, 1997; Sotriffer & Matter, 2011). Les fonctions de score basées sur un apprentissage automatique se distinguent de toutes ces fonctions conventionnelles puisqu'elles dérivent à partir d'un jeu d'entraînement une relation mathématique entre un large ensemble de descripteurs dont la forme n'est plus définie *a priori* et qui ne suit donc plus nécessairement un modèle additif et linéaire. Plusieurs études comparatives ont montré la supériorité des modèles non-linéaires dans leur capacité à prédire l'affinité de liaison (Ashtawy & Mahapatra, 2012; C. Wang & Zhang, 2017; Wójcikowski, Ballester, & Siedlecki, 2017). Cela a été particulièrement mis en évidence par deux études distinctes qui ont exploité les mêmes descripteurs que ceux utilisés par une fonction de score empirique pour reconstruire une nouvelle fonction cette fois-ci apprise à partir de forêts d'arbres décisionnels (H. Li, Leung, Wong, & Ballester, 2015; Zilian & Sotriffer, 2013). Ces fonctions ont toutes deux montré une amélioration significative dans la prédiction d'affinités de liaison par rapport au modèle linéaire des fonctions empiriques de référence. Cependant, une autre étude a montré qu'une bonne capacité à prédire l'affinité de liaison

ne garantit pas de bons résultats dans la discrimination d'une pose reproduisant un mode de liaison expérimental parmi un ensemble varié d'autres poses (Gabel, Desaphy, & Rognan, 2014). Cette même étude souligne par ailleurs que les fonctions de score basées sur un apprentissage automatique présentent le désavantage conséquent d'agir comme des boîtes noires rendant la contribution des différentes variables ainsi que leurs relations difficilement interprétables.

Comme pour les fonctions de score empiriques, le jeu d'entraînement consiste généralement en un ensemble de complexes protéine-ligand dont la structure est accessible et pour lesquelles des données d'affinité de liaison sont disponibles. Les descripteurs peuvent inclure des informations sur les interactions entre paires d'atomes, des analyses d'empreintes d'interactions (Zhan Deng, Claudio Chuaqui, & Singh\*, 2003) et prendre en compte des éléments de spécificité de reconnaissance (interactions électrostatiques, liaisons hydrogènes, stacking de résidus aromatiques), des propriétés physico-chimiques d'un site de liaison (polarité, hydrophobicité, accessibilité au solvant), sa géométrie ou encore diverses termes relatifs au ligand (poids moléculaire, nombre de degrés de liberté, etc ...). Les forêts d'arbres décisionnels, les machines à vecteurs de support, ou encore les réseaux de neurones artificiels sont des algorithmes couramment employés pour concevoir les fonctions basées sur apprentissage automatique.

## **2.6 Fonctions de score consensus**

De nombreuses fonctions de score ont été développées, mais aucune n'est universellement applicable. Certaines auront en effet de bonnes performances sur un ensemble de protéines apparentées à celles faisant partie du jeu d'entraînement utilisé pour calibrer leurs paramètres mais seront moins adaptées pour des protéines présentant des propriétés physico-chimiques différentes. De plus, chaque fonction de score possède ses propres avantages et inconvénients au regard du modèle formulé pour décrire le processus d'association d'un ligand à son récepteur. Ces caractéristiques suggèrent néanmoins que plusieurs fonctions de score doivent capturer des informations différentes. C'est sur la base de cette idée que l'application d'un scoring consensus a été introduite en combinant les prédictions de multiples fonctions de score (Charifson, Corkery, Murcko, & Walters, 1999). Plusieurs stratégies variant dans leur manière de combiner chaque score ont été entreprises et ont montré une amélioration dans la prédiction du mode de liaison, de l'affinité ou encore dans l'identification de ligands pouvant effectivement liés un récepteur lors de criblages virtuels (Bissantz, Folkers, & Rognan, 2000; Chaput & Mouawad, 2017; Ericksen et al., 2017; Kaserer et al., 2015). MultiScore (Terp, Johansen, Christensen, & Jørgensen, 2001) et X-Cscore (R. Wang, Lai, & Wang, 2002) sont deux exemples de fonctions de score consensus.

### **3 Approches par fragment**

#### **3.1 Principes des approches expérimentales et in silico**

Les prémisses des approches par fragment sont jetées avec la publication des travaux précurseurs des laboratoires Abbott sur la découverte de ligands à haute affinité construits à partir de données RMN sur des fragments localisés à des sites voisins sur une protéine cible et ensuite connectés entre eux (Shuker, Hajduk, Meadows, & Fesik, 1996). C'est le début de l'essor des approches par fragments (Murray & Rees, 2009) qui sont désormais largement utilisées dans l'industrie pharmaceutique (Hajduk & Greer, 2007). Les approches par fragment utilisées aujourd'hui pour la découverte et la conception de ligands ou drogues sont baptisées FBDD pour "Fragment-based Drug Discovery" et FBLD ou FBD pour "Fragment-based Ligand Design". Leur champ d'application a été largement étendue avec l'utilisation d'autres méthodes que la RMN et en tirant aussi profit des approches à haut-débit (Erlanson, 2006): la radiocristallographie (Hartshorn et al., 2007), la spectrométrie de masse (Pedro & Quinn, 2016), la radiocristallographie et la RMN combinées (Eaton & Wyss, 2011).

Développées pour permettre la conception de drogues à haute affinité, les approches FBDD/FBLD présentent un certain nombre d'avantages par rapport aux approches classiques de criblage de bibliothèques de plusieurs milliers de composés (Fig. 26A). Les fragments sont minimalistes avec un faible poids moléculaire et on s'attend à qu'il n'y ait pas de groupes chimiques superflus qui ne contribuent pas à l'interaction avec la cible comme cela peut être le cas avec des composés sélectionnés par criblage. Deuxièmement, un gain entropique lié aux mouvements de translation et rotation est obtenu lorsque les fragments sont connectés entre eux, sous réserve que les fragments conservent les interactions optimales établies par les fragments pris indépendamment (Fig. 26B). Enfin, il n'est pas nécessaire d'utiliser de larges bibliothèques de fragments pour couvrir un espace chimique important. Les inconvénients et les défis concernent la détection des fragments qui se lient à la cible et la stratégie de connexion des fragments.

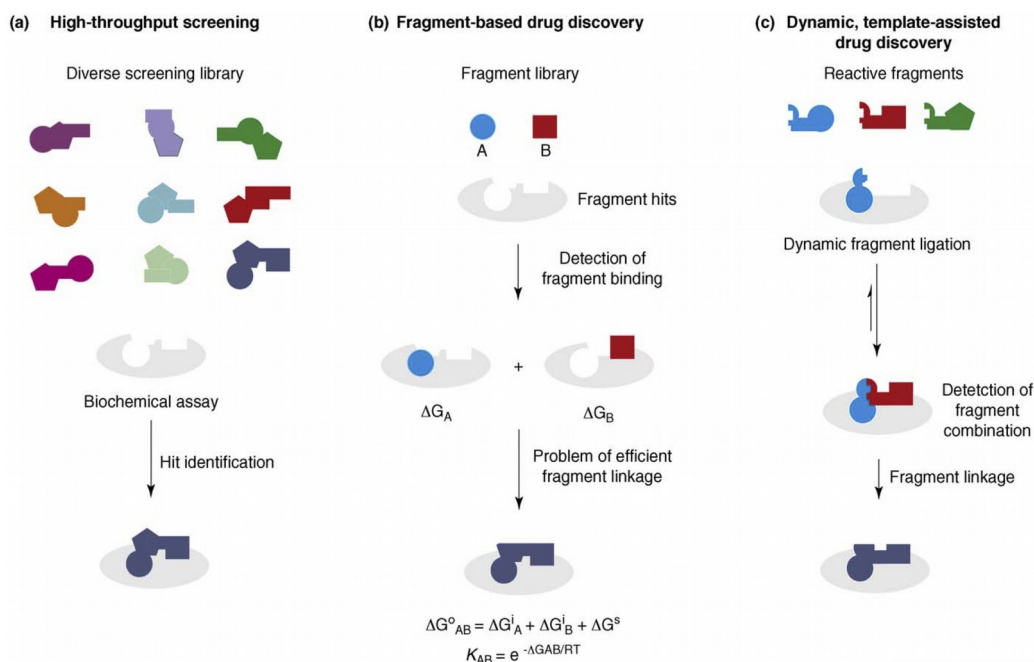


Figure 26: Approches pour la découverte et la conception de ligands. (a) criblage à haut-débit; (b) conception par fragment; (c) conception par fragments réactifs. Tirée de Schmidt et Rademann, 2009.

Une solution apportée par rapport à ces inconvénients est l'utilisation de fragments "réactifs" qui s'auto-connectent entre eux (Schmidt & Rademann, 2009), ce qui facilite à la fois leur détection et la formation du ligand (Fig. 26C). Selon la topologie des sites d'interaction potentiels à la surface de la protéine et la nature des fragments, la stratégie de connexion classique des fragments peut être basée sur le "fragment-linking" ou le "fragment-growing" (Fig. 27).

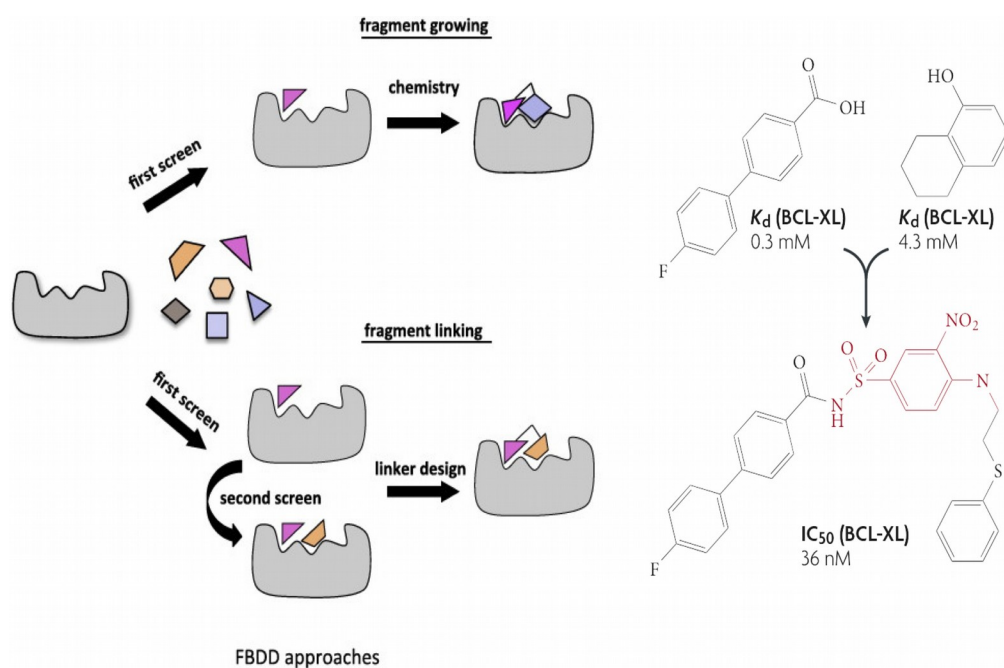


Figure 27: Stratégie de connexion dans les approches par fragments. Approches de type "fragment growing" ou "fragment linking" et exemple de gain entropique associé à la conception d'un ligand par connexion de fragments. Tirée de Hajduk & Greer, 2007.

Des approches par fragment *in silico* ont été développées à partir du début des années 1990 même si elles ne portaient pas les acronymes FBDD ou FBLD. Initialement, elles entrent dans la catégorie des approches *de novo* de conception de ligands "drug-like" (Schneider & Fechner, 2005). Selon que l'on utilise la structure de ligands connus ou la structure de la protéine ciblée, on parlera respectivement de méthodes "ligand-based" ou "receptor-based" (Zoete, Grosdidier, & Michielin, 2009). Avec les avancées de la biologie structurale, les méthodes utilisées actuellement sont "receptor-based", les ligands étant conçus à partir d'une ou plusieurs structures de la protéine cible ; elles font parti de la famille des méthodes dites "structure-based" pour la conception de ligands "drug-like" (Joseph-McCarthy, 1999). La première référence à une méthode computationnelle "fragment-based" apparaît en 1993 avec la méthode "GroupBuild" (Rotstein & Murcko, 1993) qui s'inspire notamment de l'implémentation de ce type d'approche dans le programme GRID (Goodford, 1985). GRID identifie des sites d'interaction favorables à la surface de la protéine cible en utilisant des "sondes" pseudo-atomiques (sphériques avec un atome lourd) correspondant à des groupes chimiques élémentaires (carbone aliphatique de méthyl, azote de groupe amine, oxygènes de groupes carboxyl ou hydroxyl, etc). On dénombre déjà près d'une vingtaine de méthodes "fragment-based" avant les années 2000 qui sont généralement associés à des approches de docking et utilisent des bibliothèques plus ou moins étendues de fragments (Schneider & Fechner, 2005; Zoete et al., 2009). Les méthodes proposées incluent l'ensemble des fonctionnalités de approches par fragment (LUDI, SPROUT, LigBuilder, BREED, LEA3D, etc) ou seulement une fonctionnalité



comme le docking de fragments (MCSS, HSITE, etc) et sont combinées avec d'autres méthodes pour la reconstruction de ligands par connexion des fragments (HSITE/2D Skeletons, MCSS/HOOK, MCSS/DLD, etc). Les différentes méthodes diffèrent notamment par les stratégies de connexion de fragments autorisées ("growing" ou "linking" ou les deux), et la fonction de score utilisée qui correspondent aux différentes classes de scoring utilisées dans les méthodes de docking (Zoete et al., 2009).

Les approches *in silico* présentent à peu près les mêmes avantages et inconvénients que les approches expérimentales. Les bibliothèques de fragments peuvent être davantage étendues en raison des coûts limités des approches *in silico* mais en contrepartie les ligands ainsi conçus peuvent être plus difficilement synthétisables en l'absence de filtres préalables pour sélectionner des fragments plus pertinents du point de vue de la synthèse chimique. Elles nécessitent aussi évidemment de disposer de la structure 3D de la protéine cible.

### **3.2 MCSS et méthodes complémentaires associées**

L'une des méthodes par fragments utilisée pour identifier, comme GRID, des sites d'interaction favorables à la surface d'une protéine cible est MCSS pour "Multiple Copy Simultaneous Search" (Miranker & Karplus, 1991). A la différence de GRID qui utilise des sondes pseudo-atomiques, MCSS est basée sur l'utilisation de fragments atomiques et une représentation tout-atome des fragments. Bien que la comparaison de la performance de MCSS et GRID donnent des résultats similaires, MCSS est plus adapté pour l'identification de fragments à des sites qui font intervenir plusieurs types d'interaction (hydrophobiques, polaires, électrostatiques...) et permet de mieux discriminer des sites d'intérêt (Bitetti-putzer, Joseph-mccarthy, Hogle, & Karplus, 2001). Bien que la fonction de score de MCSS soit un peu plus élaborée, elles sont assez équivalentes entre les deux méthodes (Fig. 28) et correspondent à des contributions enthalpiques de l'énergie de liaison. Dans le cas où la liaison du ligand comporte une contribution entropique dominante, la fonction de score ne fournira pas d'évaluation précise de l'interaction et pourra conduire à surestimer ou sous-estimer l'interaction à un site donné.

## CHARMM Energy Function

$$U = U_b + U_\theta + U_\phi + U_w + U_{vdw} + U_{elec}$$

### Internal Energy Terms

#### Bond Potential

$$U_b = \sum k_b (r - r_0)^2$$

#### Bond Angle Potential

$$U_\theta = \sum k_\theta (\theta - \theta_0)^2$$

#### Torsion Potential

$$U_\phi = \sum |k_\phi| - k_\phi \cos(n\phi), \quad n = 1, 2, 3, 4, 6$$

#### Improper Torsion Potential

$$U_\omega = \sum k_\omega (\omega - \omega_0)^2$$

### Nonbonded Energy Terms

#### Van der Waals Potential

$$U_{vdw} = \sum_{i>j=1} \left( \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} \right) sw(r_{ij}^2, r_{on}^2, r_{off}^2)$$

#### Electrostatic Potential

$$U_{elec} = \sum_{i>j=1} \frac{q_i q_j}{\epsilon r_{ij}}$$

Figure 28: Comparaison des fonctions de score de MCSS et GRID. Tirée de Bitetti-Putzer et al., 2001.

## GRID Energy Function

$$E_{xyz} = \sum E_{LJ} + E_{elec} + E_{hb}$$

$$E_{LJ} = \frac{A}{d^{12}} - \frac{B}{d^6}$$

$$E_{elec} = \frac{q_1 q_2}{K \zeta} \left[ \frac{1}{d} - \frac{(\zeta - \epsilon)/(\zeta + \epsilon)}{\sqrt{d^2 + 4s_p s_q}} \right]$$

$\zeta$  is the dielectric of the protein interior and  $\epsilon$  of the water;  $s_p$  and  $s_q$  are measures of the nominal depths (ref. 1)

$$E_{hb} = \left[ \frac{C}{d^8} - \frac{D}{d^6} \right] f(\theta, \theta', \dots)$$

La fonction de score de MCSS inclut également un facteur de correction par rapport à la déformation de la conformation optimale du ligand. Dans certaines implémentations, le scoring peut

aussi faire intervenir des contributions associées à la solvation/désolvation des fragments (Caflisch, 1996). Le score peut alors être approximée à une fonction d'énergie libre qui peut être exprimée par l'équation suivante:

$$\Delta G_{binding} = \Delta E_{intra} + \Delta E_{inter}^{vdw} + \Delta G_{inter}^{elec} + \Delta G_{desolv}^p + k \Delta G_{desolv}^m + \Delta G_{np} \quad (9)$$

où  $\Delta E_{intra}$  correspond à l'énergie intramoléculaire donnée par MCSS,  $\Delta E_{inter,vdw}$  l'énergie de van der Waals,  $\Delta G_{elec}$  l'énergie d'interaction électrostatique,  $\Delta G_{np}$  le terme non-polaire et  $\Delta G_{desolv}$  les termes liés à la désolvation.

Les termes électrostatiques peuvent être calculées en utilisant différents modèles comme par exemple ceux basés sur la résolution de l'équation de Poisson-Boltzmann (PB) (M. E. Davis & McCammon, 1989). Une alternative est l'utilisation de modèles plus approximatifs comme Generalized Born (GB) qui sont moins gourmands en temps de calcul et qui donnent des résultats intéressants lorsque couplé à MCSS (Haider, Bertrand & Hubbard, 2011). Dans les deux cas (PB ou GB), les contributions de solvation sont calculées à partir d'un modèle de solvant implicite qui peut aussi inclure un terme non polaire de l'énergie de solvation. Ce dernier est souvent calculée à partir de la variation de surface accessible au solvant (SA) entre le complexe protéine-ligand et les partenaires sous forme non-liée. Les scores les plus complets fournis par ces méthodes sont basés à la fois sur un calcul MCSS-CHARMM de mécanique moléculaire (MM) pour les interactions protéine-fragments (van der Waals et électrostatique) et sur un calcul de solvation (PBSA ou GBSA) où les minima obtenus par MCSS sont "re-scorés" partiellement en ajoutant les termes PBSA ou GBSA. Les modèles MM-GBSA et MM-PBSA sont parmi les plus populaires pour les scores estimés par mécanique moléculaire (Genheden & Ryde, 2015) et ont été tous les deux utilisés pour le rescoring de minima générés par MCSS (Haider, Bertrand & Hubbard, 2011 ;Caflisch, 1996).

La façon dont MCSS est implémenté permet aussi d'utiliser un modèle de solvant explicite où les fragments qui sont répliqués à la surface de la protéine peuvent être solvatés. Le fragment solvate et la/les molécule(s) d'eau (modèle TIP3) sont alors répliqués. Bien que ce modèle permette d'appréhender les interactions protéine-fragment qui pourraient faire intervenir des contacts médiés par des molécules d'eau, il présente aussi des inconvénients. *A priori*, il peut être difficile de déterminer combien de molécules d'eau il est pertinent d'inclure dans le fragment solvate. Par ailleurs, les molécules d'eau répliquées pendant les calculs MCSS ont tendance à « s'évaporer », c'est-à-dire à s'éloigner du fragment pour se lier à d'autres sites de la protéine. Ce phénomène peut

conduire à des variations importantes du score des fragments. Ce modèle demanderait des tests et une calibration pour évaluer sa pertinence (Leclerc, communication personnelle). Une alternative est d'inclure une solvation partielle de la protéine où les molécules d'eau peuvent être tirées de données cristallographiques. Ce modèle demande à être testé et validé ; des tests préliminaires sont en cours de réalisation (Gonzalez-Aleman & Leclerc, communication personnelle).

MCSS est exécuté en plusieurs étapes et de façon itérative afin de faire converger les calculs vers des modes de liaison correspondant à un positionnement optimal des fragments ou minima. Les étapes d'initialisation permettent de générer une distribution initiale de  $n$  fragments dans une région donnée de la protéine cible. A chacune des  $m$  itérations, les  $n$  fragments sont optimisés en utilisant la fonction REPLICIA implémentée dans CHARMM (Brooks, Bruccoleri, Olafson, States, SWAMINATHAN, et al., 1983) permettant d'échantillonner localement et de façon efficace l'espace des modes de liaison des fragments à la surface de la protéine. Une méthode de clustering permet, *in fine*, de ne conserver que les minima non redondants et d'aboutir à une distribution de fragments qui constitue une cartographie de fragments correspondant à des groupes chimiques fonctionnels (Fig. 29).

Comme MCSS est conçu seulement pour identifier des fragments se liant favorablement à des sites à la surface de la protéine, elle doit être combinée avec d'autres méthodes pour réaliser l'étape de connexion des fragments et l'optimisation des ligands. Elle peut être associée au programme HOOK (Eisen, Wiley, Karplus, & Hubbard, 1994) ou à DLD (Stultz & Karplus, 2000). Le principe de HOOK est de placer des "squelettes" moléculaires à proximité des minima MCSS afin de connecter les fragments entre eux. Une banque de "squelettes" est disponible dans HOOK ; elle inclut par exemple des groupes benzène. Dans le cas de DLD, les connexions entre fragments sont effectués avec des atomes de carbone de différentes hybridations ( $sp^3$ ,  $sp^2$ ) ou des groupes cycliques. Les applications de MCSS pour la conception de ligands par fragment sont nombreuses et variées puisqu'elles vont de l'optimisation de ligands à partir de pharmacophores (Zheng et al., 2007) à la conception d'oligopeptides et de composés peptide-like (Evensen, Joseph-McCarthy, Weiss, Schreiber, & Karplus, 2007). MCSS a aussi été utilisé pour reproduire les interactions d'autres biopolymères sur des protéines cibles comme des dérivés d'oligosaccharides qui sont des antibiotiques de la famille des aminoglycosides qui se lient au site A de l'ARNr 16S (Leclerc & Karplus, 1999). Les calculs MCSS de mécanique moléculaire sont effectués par CHARMM, ce qui permet de bénéficier de l'ensemble des fonctionnalités du programme tant pour l'utilisation des paramètres du champ de force CHARMM que pour l'utilisation de modèles de solvation. CHARMM inclut des paramètres pour des résidus modifiés d'acides aminés (Grauffel, Stote, &

Dejaegere, 2010) ou de nucléotides (Xu et al., 2016) et un nouveau champ de force général CGenFF (Vanommeslaeghe & MacKerell, 2012) permettant de dériver des paramètres génériques pour à peu près n'importe quel type de molécules. Différents modèles de solvant sont ainsi potentiellement utilisables avec MCSS.

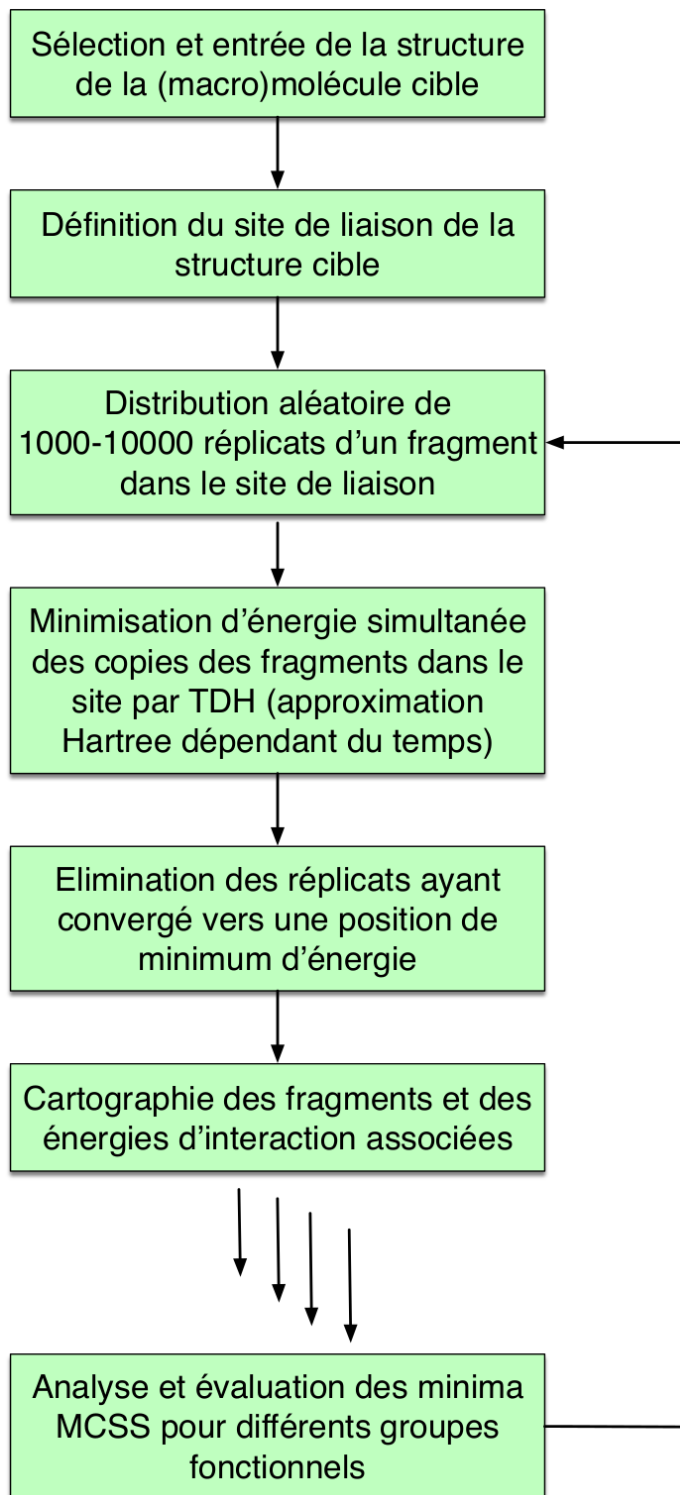


Figure 29: Schéma de fonctionnement de MCSS étape par étape. D'après Schubert & Stultz, 2009.

## **IV Modélisation des interactions protéine-ARNsb : état de l'art**

### **1 Du 2D à la 3D**

L'intérêt grandissant pour les interactions protéine-ARN a beaucoup stimulé la communauté durant les 10 dernières années pour développer des méthodes de prédiction de leurs modes d'interaction. On peut toutefois englober des approches différentes que l'on peut distinguer en fonction de la représentation implicite ou explicite du ligand ARN et du type d'information structurale utilisée (séquence et structure 2D ou structure 3D). Dans les approches où l'ARN est décrit de façon implicite, il s'agit de méthodes de type SAR ou "structure/activity relationships" qui ont pour objectif de prédire la région et les sites d'interaction de l'ARN d'après les connaissances disponibles sur telle ou telle famille de protéines de liaison à l'ARN. Ce sont essentiellement des approches par apprentissage qui exploitent des données de séquence ou de structure 3D dans l'objectif de prédire l'interface protéine-ARN (Cirillo et al., 2013; Walia et al., 2012). Par défaut, les approches basées sur la séquence utilisent la composition et similarité de séquence et les données évolutives. Le même type d'approche basée sur la structure utilise d'autres descripteurs associés à la structure 3D tels que les profils électrostatique et hydrophobe, la topologie (cavités), ou autres. Un article récent par Miao et Westhof dresse le bilan de la performance de ces approches (Miao & Westhof, 2015).

### **2 Le docking protéine-ARN**

Les approches basées sur une description explicite de l'ARN ont pour objectif quant à elles de prédire à la fois l'interface protéine-ARN et la conformation de l'ARN lié, ce sont des méthodes de docking classique. Les stratégies développées et appliquées lors du défi CAPRI 2008 consistaient essentiellement à effectuer un docking rigide puis flexible ou combinée (Fleishman et al., 2010). La conformation de l'ARN étant fournie comme donnée de départ, le traitement de la flexibilité de l'ARN ne se posait pas. A l'image de l'approche NPdock publiée récemment et disponible en ligne (Tuszynska et al., 2015), les méthodes développées pour le docking protéine-ARN sont pour la plupart dérivées de méthodes existantes pour le docking protéine-protéine telles que HADDOCK, HEX, FTDOCK (Lensink & Wodak, 2010; Pons, Grosdidier, Solernou, Pérez-Cano, & Fernández-Recio, 2010).

Plusieurs benchmarks ont été construits et utilisés pour évaluer la performance des méthodes de docking protéine-ARN (S. Y. Huang & Zou, 2013; Pérez-Cano & Fernández-Recio, 2010). L'ensemble des benchmarks existants traitent de façon indifférenciée les complexes avec des ARN

essentiellement structurée sous forme double-brin ou à la fois double-brin et simple-brin (sans éléments de structure secondaire) ou essentiellement simple-brin. Différentes méthodes de scoring ont aussi été développées récemment pour les interactions protéine-ARN avec des performances variables (Fornes, Garcia-Garcia, Bonet, & Oliva, 2014; S.-Y. Huang & Zou, 2014a; C. H. Li, Cao, Su, Yang, & Wang, 2012) mais aucune d'entre elles ne permet *a priori* de discriminer parmi des ligands ARN liés à l'interface protéine-ARN ceux qui sont spécifiques en terme de séquence (Lensink & Wodak, 2010).

Les ARN simple-brin présentent un intérêt biologique et structural particulier car ils sont une composante qualitative (flexibilité) et quantitative (degré de régions simple-brin/double-brin) importante des ARN. Outre le fait qu'ils soient présents dans tous les ARN sous la forme de bulges, de boucles internes ou terminales, ils correspondent aussi à des motifs reconnus par de nombreuses protéines de liaison à l'ARN spécifiques des régions simple-brin. Des données récentes obtenues par des approches à haut-débit suggèrent d'ailleurs que les ARN seraient beaucoup moins structurés *in vivo* que ne le laissent penser les données *in vitro* et les prédictions *in silico* (Rouskin et al., 2014). Les ARN messagers synthétisés dans des conditions de stress seraient aussi significativement plus riches en régions simple-brin (Ding et al., 2014). Les protéines jouent un rôle important dans le changement de conformations des ARN en structurant certaines régions de l'ARN, ou bien en déstructurant d'autres rendant certains motifs accessibles.

### **3 Approches par fragment pour modéliser les interactions entre protéines et ARN liés sous forme simple-brin non structurée**

Ce n'est que très récemment que des approches permettant de modéliser les interactions entre protéines et ARN simple-brin (ARNsb) ont vu le jour. Elles ne sont qu'au nombre de trois au moment de l'écriture de ce manuscrit : RNA-LIM parue en 2015 (Hall et al., 2015), une approche basée sur le programme de docking ATTRACT parue en 2016 (Chauvot de Beauchene, de Vries, & Zacharias, 2016a; De Beauchene, De Vries, & Zacharias, 2016) et *RNP-denovo* (Kappel & Das, 2019) parue en 2019. Ces méthodes s'inscrivant directement dans la thématique de cette thèse, une attention particulière leur est portée.

#### **3.1 RNA-LIM**

L'approche RNA-LIM a été développée pour prédire les positions les plus probables d'une chaîne d'ARNsb de séquence connue à la surface d'une protéine. Elle repose au préalable sur la connaissance de la propension de chacun des acides aminés de la protéine à interagir avec des nucléotides. A chaque nucléotide de la séquence ARN donnée en entrée est associée une liste

d'acides aminés, ceux présentant la plus grande propension. Ces propensions acide aminé-nucléotide spécifiques sont utilisées pour réduire l'espace de recherche. Un ensemble de chemins est défini à partir de ces acides aminés en connectant ceux dont la longueur d'arc superficiel (Hall et al., 2014) satisfait un critère de distance. Les 100 chemins les plus probables sont ensuite sélectionnés pour énumérer d'une manière systématique et discrète l'ensemble possible des chaînes d'ARN. Pour chaque chemin, environ 30 nucléotides représentés sous forme d'une sphère (centrée sur l'atome C3' du ribose) sont aléatoirement distribués localement autour de chaque acide aminé d'intérêt. Ces nucléotides sont séparés au minimum de 3 Å (distance centre-centre). Une énergie conformationnelle est assignée à chacune des chaînes possibles par un pseudo-potential statistique dérivé de NAST (Jonikas et al., 2009) correspondant à la somme d'énergies de liaison et d'angle. Les 100 chaînes de meilleure énergie sont conservées, conduisant ainsi à un ensemble de 10000 solutions potentielles (100 chemins x 100 chaînes). Testée sur la forme liée d'une protéine à domaine RRM liant un ARNsb de cinq nucléotides à partir du prédicteur aaRNA (S. Li, Yamashita, Amada, & Standley, 2014), l'approche a permis de générer parmi les 100 meilleures solutions des chaînes présentant un RMSD d'environ 5 Å par rapport au modèle expérimental.

La méthode RNA-LIM permet l'énumération d'un ensemble de positions de chaînes d'ARNsb grâce à la réduction de la complexité du système. Cette réduction provient d'une part de la représentation simplifiée en mode gros-grain de la protéine et des ribonucléotides, d'autre part d'une recherche de chaînes localisée autour d'acides aminés présentant une plus grande probabilité d'interagir avec l'ARN. La représentation ultra simplifiée des nucléotides sous forme d'une sphère présente l'avantage de réduire le nombre de degrés de liberté à explorer mais, en contrepartie, la résolution des modèles obtenus est limitée : les meilleurs modèles obtenus sur le système test présentent un RMSD autour de 5 Å et plus globalement, l'orientation des nucléotides au sein des chaînes ne peut être définie. La recherche systématique place autour des acides aminés de plus grande probabilité d'interaction des sphères séparées de 3 Å. Une discrétisation plus fine de leur placement permettrait d'augmenter la résolution des modèles mais au prix d'une expansion massive de l'espace des solutions. Les auteurs mentionnent par ailleurs une limitation similaire pour des chaînes composées de plus de six nucléotides. Remarquons également que la pertinence des modèles découle directement des capacités prédictives du programme utilisé pour définir la propension des acides aminés à interagir spécifiquement avec un ribonucléotide. Les programmes dédiés à la prédiction de résidus pouvant interagir avec l'ARN ont globalement montré d'excellentes capacités prédictives (AUC  $\sim$  0,85) sur plusieurs jeux de données protéine-ARN (Miao & Westhof, 2015). C'est le cas de aaRNA utilisé sur le système testé par RNA-LIM. Il reste



néanmoins difficile d'évaluer la robustesse de l'approche RNA-LIM puisque celle-ci n'a été testée que sur un système avec un seul prédicteur d'interaction.

### **3.2 Approche basée sur le programme de docking ATTRACT**

Une deuxième approche parue en 2016 traite le problème de complexité associée à l'échantillonnage conformationnelle d'un ARNsb en le restreignant à des fragments trinuécléotidiques (3-nt). L'approche repose sur une bibliothèque de conformères 3-nt prélevés sur un large ensemble de structures protéine-ARN. Plusieurs milliers de conformations sont ainsi associées à chacun des 64 triplets possibles. Considérons une séquence d'ARN d'intérêt. L'ensemble des positions possibles des triplets composant cette séquence est échantillonné en dockant à la surface d'une protéine les conformations de la bibliothèque correspondante. Le docking est effectué par le programme ATTRACT et les fragments 3-nt étant rigides, seuls des mouvements de translation et rotation sont explorés. Un ensemble de chaînes peut être reconstruit en connectant les poses trinuécléotidiques répondant à un critère de chevauchement spatial et compatibles au niveau de leur séquence. Sans autre contrainte que la séquence d'ARN, cette approche a permis d'obtenir sur la forme liée de deux protéines à domaine RRM liant une chaîne d'ARNsb de huit nucléotides un ensemble d'une dizaine de milliers de solutions dont la majorité était enrichie au niveau du site de liaison (dans un « rayon » RMSD de 10 Å par rapport aux chaînes natives). Ces dernières incluent par ailleurs des modèles de chaînes d'ARN reproduisant le mode d'interaction natif avec une excellente résolution ( $\sim 1,5$  Å) mais qui ne pouvaient être discriminés avec précision par la fonction de score gros-grain utilisée (top 2000).

L'approche fut par la suite adaptée en incluant des contraintes sur quelques résidus conservés de la protéine. Ces contraintes sont utilisées pour prédire des poses d'ancrage initiales à partir desquelles des chaînes d'ARN sont construites par un processus de docking itératif. Dans le premier cycle de docking, les poses d'ancrage de meilleure énergie d'interaction sont conservées. Elles sont ensuite utilisées dans un second cycle de docking pour contraindre le positionnement de poses associées au second triplet. Les poses d'énergie plus favorable sont une nouvelle fois conservées et la procédure est ainsi répétée jusqu'à ce que la totalité des triplets composant la séquence d'ARN d'intérêt a été échantillonnée. Cette approche a été testée sur la forme liée de sept protéines à domaines RRM et a permis de réduire le nombre de modèles à une petite centaine. Des chaînes reproduisant le mode d'interaction avec une haute résolution ( $\sim 2$  Å) se trouvaient dans le top 100 pour cinq de ces protéines, et dans le top 10 pour quatre d'entre elles. La méthode a par ailleurs montré des résultats similaires sur deux protéines à domaines PUF suggérant qu'elle pouvait être transférable à d'autres familles de domaines de liaison à l'ARN. La méthode a de plus été appliquée

à partir de la forme non-liée de deux protéines et de la connaissance de l'orientation de leurs deux domaines RRM. Fait notable, les changements conformationnels locaux (de 1,5 et 1,7 Å) entre les formes liées et non-liées n'ont eu aucun effet négatif sur la qualité des modèles obtenus ( $\sim 2$  Å).

Globalement, l'approche a permis de prédire avec une très bonne précision le mode d'interaction de chaînes d'ARNsb allant jusqu'à 11 nucléotides, dénotant ainsi son efficacité à traiter la flexibilité inhérente aux ARNsb. La précision de ces modèles est dépendante du caractère exhaustif de la bibliothèque de conformères 3-nt utilisée. Cette dernière est construite à partir de structures de complexes protéine-ARN déjà résolues qui ne couvrent pas tout l'espace possible des conformations 3-nt. Les conformations ne sont en effet pas représentées de manière équivalente pour chaque triplet. L'approche présente donc une certaine limitation à prédire des conformations locales de 3-nt déjà observées. Notons toutefois que cette limitation n'a eu un effet que modéré sur les sept systèmes testés et qu'elle est vouée à s'atténuer au fur et à mesure que de nouvelles structures protéine-ARN seront résolues.

L'utilisation de fragments 3-nt autorise une construction de chaînes aisée par un simple critère de chevauchement spatial. Notons cependant que les nucléotides constituant une chaîne d'ARNsb liée à une protéine contribuent généralement différemment à l'interaction : certains nucléotides présentent une énergie d'interaction plus favorable que d'autres. Dans ce contexte, la sélection de fragments d'intérêt présentant une faible énergie d'interaction peut constituer une difficulté pour l'approche. Leur sélection impose en effet de retenir un plus grand nombre de poses, ce qui peut conduire à augmenter l'espace des solutions. L'utilisation de fragments 3-nt présente néanmoins l'avantage que l'énergie d'interaction d'un nucléotide établissant peu de contacts avec la protéine peut être compensée par une énergie plus favorable des autres nucléotides du même fragment si ces derniers établissent suffisamment de contacts avec la protéine.

La sélection des poses et des chaînes modélisées repose sur une fonction de score à potentiel statistique où les molécules sont décrites par un modèle gros-grain : chaque pyrimidine/purine est représentée par six ou sept sphères et chaque acide aminé par trois ou quatre sphères. Cette représentation simplifiée permet des calculs plus rapides autorisant l'évaluation d'un large ensemble de conformations 3-nt (de l'ordre du millier au million). Elle conduit par ailleurs à un paysage énergétique moins rugueux qui explique certainement la relative insensibilité observée face à de faibles changements conformationnels de la protéine. En contrepartie, l'utilisation d'une fonction de score de type gros-grain peut conduire à des imprécisions dans l'estimation énergétique des poses. Ces imprécisions peuvent rendre difficile la discrimination de fragments d'intérêt et entraînent la nécessité de sélectionner un plus large ensemble de poses. Notons que les 1000 chaînes

de plus basse énergie sont reconstruites en tout-atome puis optimisées dans un champ de force AMBER. Cette dernière étape suivie d'un clustering a permis globalement d'améliorer la précision des modèles pré-sélectionnés et de réduire leur nombre à une centaine tout au plus.

Soulignons finalement que les modèles à haute précision obtenus par l'approche repose sur la connaissance de résidus conservés utilisés pour contraindre le positionnement de poses d'ancrage. Les auteurs mentionnent qu'une imprécision dans leur positionnement d'environ 3 Å conduirait probablement à réduire la qualité des modèles.

### **3.3 RNP-denovo**

Une troisième approche est apparue encore plus récemment : Rosetta *RNP-denovo*. L'approche repose sur la méthode FARNA (Das and Baker, 2007) qui a été développée pour prédire la structure tertiaire d'un ARN. FARNA replie une séquence d'ARN à partir d'un ensemble d'angles de torsion associés à des fragments 3-nt et stockés dans une bibliothèque. Les angles de torsion des fragments 3-nt sont tirés de la structure d'un ARN ribosomique. La procédure de repliement utilise un échantillonnage de type Monte-Carlo guidé par une fonction de score gros-grain à potentiel statistique. La séquence ARN d'intérêt est initialisée dans une conformation étendue à partir d'angles standards. A chaque étape, une position aléatoire est choisie dans cette chaîne pour y remplacer ses angles de torsions par ceux d'un fragment sélectionné au hasard dans la bibliothèque. Le mouvement est accepté ou rejeté selon un critère de Métropolis. La fonction de score utilisée dans FARNA prend en considération la conformation du squelette ribose-phosphate et différentes interactions entre paires de bases. La méthode Rosetta *RNP-denovo* a été développée en modifiant FARNA de manière à y intégrer des mouvements protéine-ARN durant l'échantillonnage ainsi que divers types d'interactions protéine-ARN dans la fonction de score. Elle requiert en entrée une séquence d'ARN ainsi que la connaissance de la position exacte de quelques nucléotides en contact avec la protéine, ces derniers étant utilisés comme contraintes pour le repliement de l'ARN. Cette approche de docking et repliement simultanés a été testée sur la forme liée de 10 RBPs différentes : sur les 100 modèles retenus pour chaque système, le RMSD moyen des solutions reproduisant au mieux le mode d'interaction natif est de 4,3 Å. Parmi les 10 RBPs, quatre correspondent à des protéines liant spécifiquement un ARN sous forme simple-brin sans élément de structures secondaires dans sa forme liée ; chacune de ces quatre protéines est composée de domaines de liaison à l'ARN de famille différente (RRM, KH, PUF et Hut). Pour ces dernières, le RMSD du meilleur modèle varie entre 3,1 et 4,2 Å.

Comme l'approche basée sur ATTRACT, *RNP-denovo* repose sur une bibliothèque de fragments 3-nt pour explorer l'espace conformationnel d'une séquence ARN. Cette bibliothèque permet de restreindre l'échantillonnage à des conformations locales déjà observées ; elle est cependant moins exhaustive que celle de l'approche basée sur ATTRACT puisqu'elle découle d'une seule structure d'un ARN ribosomique d'environ 2700 nucléotides. La méthode *RNP-denovo* semble donc limiter par un manque de couverture de l'espace conformationnel local pouvant affecter la précision de ses modèles.

L'approche *RNP-denovo* requiert en entrée, en plus de la séquence de l'ARN, la connaissance exacte de quelques contacts protéine-nucléotide pour réduire l'espace conformationnel à échantillonner. Ces informations peuvent être obtenues par des approches expérimentales comme la RMN ou des données FRET (*Förster Resonance Energy Transfer*). Remarquons que pour les 10 systèmes testés, la précision des meilleurs modèles obtenus avoisine les 4 Å malgré l'utilisation de contacts protéine-nucléotide précisément déterminés. Cela met en évidence la limitation de l'approche à prédire des modèles à plus haute résolution.

Comparée aux deux approches présentées précédemment, la méthode *RNP-denovo* a la particularité de ne pas être spécifiquement dédiée à la modélisation d'ARN interagissant uniquement sous forme simple-brin avec une protéine. En s'attaquant simultanément à la prédiction du repliement des ARN et du mode d'interaction protéine-ARN, la méthode doit faire face à une complexité accrue : l'espace conformationnel à explorer est plus important et il est de plus nécessaire d'estimer l'énergie d'interaction incluant des termes aussi bien intermoléculaires (interactions protéine-ARN) qu'intramoléculaires (interactions ARN-ARN). La fonction de score utilisée est de type potentiel statistique et repose sur un modèle gros-grain (la description des atomes des acides aminés et des nucléotides varient selon les termes énergétiques considérés) pour bénéficier de temps de calculs plus rapides. Cette fonction de score a permis un enrichissement de modèles à basse résolution ( $\sim 4$  Å) parmi les 100 meilleures solutions, mais n'autorise pas une discrimination plus précise (présence dans le top 1 ou top 10 par exemple) des meilleurs modèles. Notons que la comparaison à d'autres fonctions de score ne considérant que des termes énergétiques intermoléculaires a mis en évidence que l'inclusion de termes intramoléculaires conduit à des modèles plus précis pour les systèmes présentant un ARN structuré sous sa forme liée, mais n'entraîne pas de changements pour ceux où l'ARN interagit sous forme simple-brin, sous-entendu sans élément de structures secondaires dans sa forme liée.

Le tableau 1 ci-dessous résume les principales caractéristiques de la méthode RNA-LIM, de l'approche basée sur ATTRACT et de l'approche Rosetta *RNP-denovo*.

Tableau 1: Principales caractéristiques des approches de docking existantes permettant la modélisation des interactions entre protéines et ARNs non-structurés dans leur forme liée

	<b>RNA-LIM</b>	<b>Approche basée sur ATTRACT</b>	<b>Rosetta <i>RNP-denovo</i></b>
<b>Entrées</b>	- Structure de la protéine - Séquence de l'ARN - Liste d'acides aminés ayant une probabilité importante d'interagir avec un nucléotide donné	- Structure de la protéine - Séquence de l'ARN - Liste d'acides aminés utilisés comme contraintes	- Structure de la protéine - Séquence de l'ARN - Liste de contacts protéine-nucléotide utilisés comme contraintes
<b>Pré-supposé sur le mode d'interaction</b>	- ARN interagit sous forme simple-brin	- ARN interagit sous forme simple-brin	- Aucun
<b>Type de fragments</b>	- 1 mono-nucléotide	- Ensemble de conformères tri-nucléotidiques	- Ensemble de conformères tri-nucléotidiques
<b>Flexibilité des fragments</b>	- Aucune	- Aucune	- Aucune
<b>Représentation du fragment</b>	- très gros-grain : 1 nucléotide = 1 bille	- semi gros-grain : 1 nucléotide = 6/7 billes	- semi gros-grain : jusqu'à 9 atomes par nucléotide, selon les termes énergétiques évalués
<b>Représentation de la protéine</b>	- très gros-grain : 1 acide aminé = 1 sphère	- gros-grain : de 3 à 4 sphères par acide aminé	- gros-grain : 6 sphères par acide aminé
<b>Construction des chaînes d'ARN</b>	- Connexion de nucléotides par satisfaction de contraintes (séquence, angle, distance)	- Assemblage de conformères 3-nt par chevauchement spatial	- Enfilement de conformères 3-nt guidé par l'évaluation énergétique de contacts ARN-ARN et protéine-ARN
<b>Type de fonction de score</b>	- Potentiel statistique dérivée de complexes protéine-ARN	- Potentiel statistique dérivée de complexes protéine-ARN pour les poses et une pré-sélection des chaînes - Basée sur un modèle physique en tout-atome pour les modèles finaux	- Potentiel statistique dérivée de complexes protéine-ARN
<b>Résolution globale des meilleurs modèles</b>	- ~ 5 Å (centre des nucléotides) sur 1 système à domaine RRM	- ~ 2 Å (atomes lourds) sur 7 systèmes à domaines RRM et 1 système à domaines PUF	- ~ 4 Å (atomes lourds) sur 10 systèmes incluant notamment les domaines RRM, KH, PUF et Hut
<b>Longueur des chaînes modélisées</b>	- 5 nucléotides	- jusqu'à 11 nucléotides	- jusqu'à 45 nucléotides

# Travaux de thèse

## I Objectifs

Au commencement de cette thèse, en octobre 2014, aucune méthode n'était encore parue pour la modélisation spécifique aux interactions protéine-ARN simple-brin. Tous les travaux réalisés au cours de cette thèse ont eu pour objectif le développement d'une approche permettant de prédire leur mode d'interaction. De manière comparable aux approches qui ont depuis été publiées, la méthode développée, qui sera référencée par l'acronyme FBDRNA pour *Fragment-Based design of RNA*, repose sur une approche de docking par fragments. Par rapport à ces dernières, le développement de FBDRNA a été entrepris en vue de pouvoir répondre à deux objectifs distincts : à partir de la structure d'un domaine de liaison à l'ARN (RBD) que l'on sait reconnaître spécifiquement un ARN simple-brin non structurée dans sa forme liée et de la connaissance de son site d'interaction :

- prédire le mode d'interaction d'une séquence ARN donnée
- prédire la séquence préférentiellement reconnue par le RBD simultanément à son mode d'interaction

L'approche FBDRNA se distingue des autres approches existantes en explorant l'espace des conformations d'une séquence ARN à l'échelle d'un fragment mono-nucléotidique représenté en tout-atome. Elle repose sur deux programmes pré-existants. Le premier est le programme de docking MCSS qui a été brièvement introduit et qui sera plus précisément décrit dans la partie suivante. MCSS est utilisé pour générer un ensemble de poses nucléotidiques à la surface d'intérêt d'une protéine. Les deux molécules sont représentées en tout-atome et sont décrites d'après les paramètres du champ de force CHARMM27. Le deuxième programme, Molpy (<https://github.com/mianuel/Molpy>), a été développé par Manuel Simoes et est présenté dans la section "Matériels et méthodes" de la partie II. Molpy est utilisé pour "reconstruire" des chaînes ARN à partir de poses préalablement sélectionnées satisfaisant différentes contraintes. Mon travail s'est essentiellement focalisé sur l'étape de sélection des poses à donner en entrée au programme Molpy. C'est une étape essentielle à la faisabilité de l'approche puisqu'elle doit permettre de retenir l'ensemble des poses nucléotidiques d'intérêt autorisant la construction de chaînes reproduisant le mode d'interaction natif tout en évitant un phénomène d'explosion de l'espace des solutions.

Après une description des procédures générales (partie II) utilisées durant les travaux, l'attention sera portée sur la mise au point d'un protocole de sélection pour la prédiction du mode d'interaction entre des domaines de liaison à l'ARN et des chaînes d'ARNsb de séquence connue (partie III). La

partie suivante (IV) présente une adaptation de ce protocole au problème de prédiction dans le cas où aucun *a priori* n'est établi sur la séquence ARN. Les deux dernières parties (parties V et VI) s'inscrivent dans des perspectives d'amélioration de l'approche FBDRNA. Une dernière partie conclura ce manuscrit en résumant l'ensemble des travaux présentés et les projets identifiés pour améliorer les résultats.

## II Méthodes générales

Cette partie présente les méthodes communes aux différents travaux présentés (parties III, IV , V et VI).

### 1 Simulations de docking avec MCSS

#### 1.1 Procédure d'échantillonnage et paramètres utilisés

Dans un protocole typique de MCSS, plusieurs copies d'un ligand (ou fragment) sont aléatoirement distribuées à la surface d'une protéine circonscrite à un espace pré-défini. Chacune des poses résultantes est alors optimisée simultanément par plusieurs cycles de minimisation. Au cours de cette étape, la protéine est maintenue rigide tandis que le ligand, rendu flexible, est libre de tout mouvement. Si deux poses convergent à une même position, seule celle de meilleure énergie d'interaction (énergie la plus basse) est conservée. Par ailleurs, si l'énergie d'interaction d'une pose est supérieure a une valeur définie, alors cette dernière est rejetée. L'ensemble de ce processus peut être répété durant un nombre d'itérations défini par l'utilisateur. A la fin du docking, toutes les poses retenues sont associées à une énergie d'interaction et sont rangées de la pose d'énergie la plus favorable à la moins favorable.

Tous les calculs de docking ont été réalisés avec les paramètres suivants :

- 2500 copies d'un fragment ont été dockées durant 20 itérations
- les cycles de minimisation utilisés comprennent 500 étapes initiales de SD suivies de 300 autres étapes de SD, puis 20 répétitions de 500 étapes de gradient conjugué et enfin 200 étapes de ABNR.
- un seuil RMSD (voir plus bas pour le calcul) de 0,5 Å a été utilisé pour définir deux poses convergeant à une même position. Le RMSD est calculé sur tous les atomes lourds à l'exception des atomes d'oxygène du groupement phosphate.
- un seuil énergétique de +40 kcal/mol a été défini comme valeur à ne pas dépasser pour conserver une pose. Ce seuil est volontairement fixé à une valeur élevée de manière à maximiser les chances de retrouver des poses natives n'établissant qu'un nombre limité de contacts avec la protéine. Les autres paramètres utilisés correspondent aux valeurs par défaut de MCSS.



## 1.2 Préparation des protéines

Une étape préalable à toute simulation de docking consiste en la préparation des structures protéiques utilisées pour le docking. Toutes les structures utilisées durant les travaux de thèse proviennent de leur forme liée. Elles ont été préparées à partir du serveur web <http://www.charmm-gui.org/> (Jo, Kim, Iyer, & Im, 2008). Le ligand (ARN ou nucléotide), ainsi que tous les hétéroatomes (molécules d'eau, ions, métaux) ont été retirés des protéines. Les atomes d'hydrogène ont été reconstruits par la commande HBUILD de CHARMM. Les histidines ont été traitées comme neutres. Une fois ces étapes effectuées, les chaînes latérales des protéines ont été relaxées par 1000 étapes de minimisation selon la méthode de plus grande pente (SD) suivies de 1000 autres étapes par la méthode de Newton-Raphson (ABNR). Enfin, la totalité des atomes des protéines a été minimisée par 500 étapes de la méthode SD suivies de 500 étapes de la méthode ABNR.

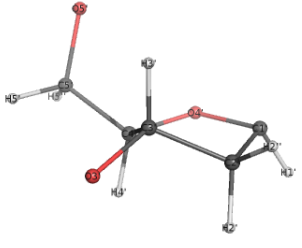
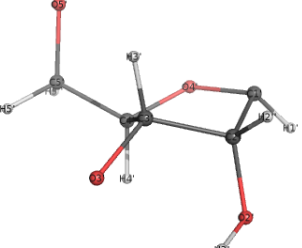
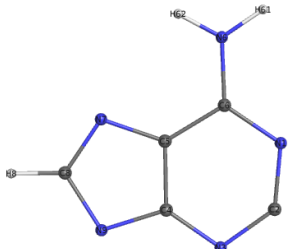
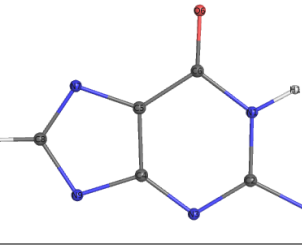
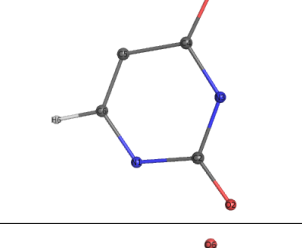
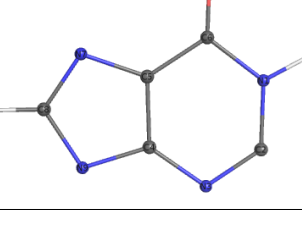
NB : Ces étapes de minimisation conduisent à de légers réarrangements dans le positionnement des chaînes latérales. La comparaison visuelle des structures avant et après minimisation montre cependant que ces réarrangements sont infimes, y compris pour les résidus directement impliqués dans l'interaction avec le ligand. En conséquence, par commodité, la structure minimisée utilisée pour le docking pourra être référencée dans la suite du manuscrit comme étant une structure dans sa forme liée, bien que cela ne soit pas strictement exacte.

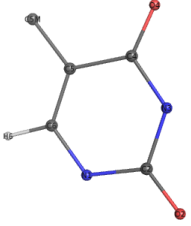
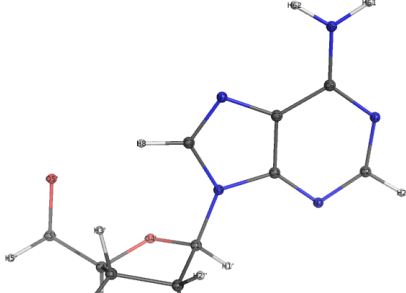
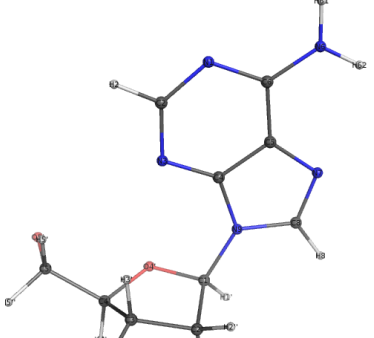
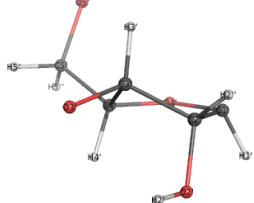
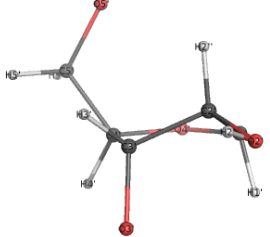

## 1.3 Fragment nucléotidique utilisé comme ligand

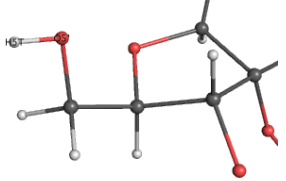
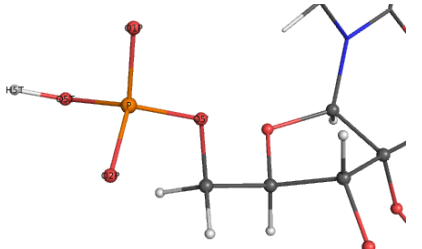
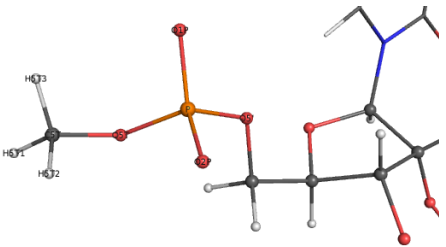
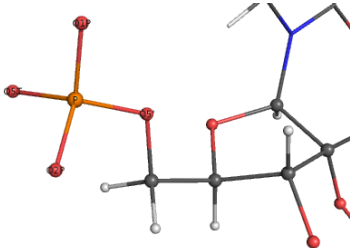
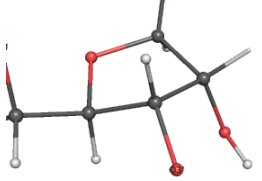
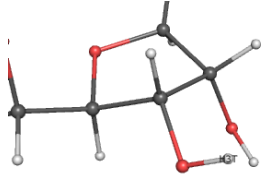
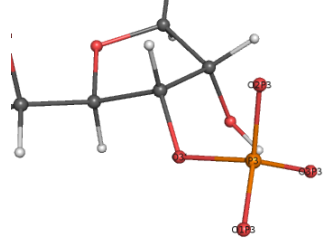
Toute simulation de docking requiert également la structure d'un ligand à partir de laquelle pourront être explorées différentes positions, orientations et conformations à la surface d'intérêt d'une protéine. Tous les calculs de docking réalisés durant cette thèse ont été effectués à partir d'une structure nucléotidique intégrée dans le programme MCSS. MCSS contient une variété de structures de nucléotides présentant une conformation pré-définie et prêtes à l'emploi ; elles peuvent être choisies à partir d'une chaîne à sept caractères répondant à une nomenclature présentée dans le tableau 2 ci-dessous. Chaque position de caractère correspond à une partie modulable du nucléotide.

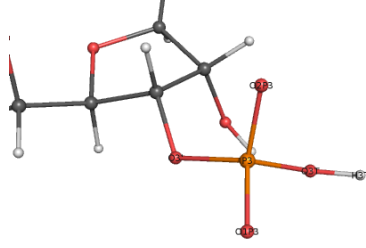
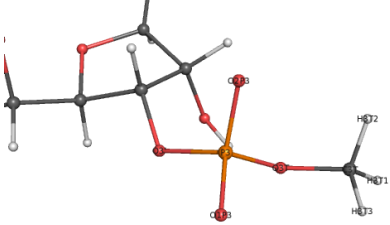
Tableau 2: Description de la nomenclature utilisée dans MCSS pour désigner la structure du ligand utilisé pour le docking.

Position du caractère	Partie modulée	Caractère	Module	Représentation du module
1	Type de ribose	0	Aucun	/

Position du caractère	Partie modulée	Caractère	Module	Représentation du module
		D	Désoxyribose	
		R	Ribose	
2	Type de base	A	Adénine	
		G	Guanine	
		U	Uracile	
		I	Inosine	

Position du caractère	Partie modulée	Caractère	Module	Représentation du module
		T	Thymine	
3	Orientation de la base	X	Anti	
		Y	Syn	
4	Plissement du ribose	N	Nord (C3'-endo)	
		S	Sud (C2'-endo)	
5	Extrémité 5'-ter du sucre	0	O5'-PO <sub>2</sub> <sup>-</sup>	

Position du caractère	Partie modulée	Caractère	Module	Représentation du module
		1	O5'-H	
		2	O5'-HPO <sub>3</sub> <sup>-</sup>	
		3	O5'-CH <sub>3</sub> PO <sub>3</sub> <sup>-</sup>	
		4	O5'-PO <sub>3</sub> <sup>2-</sup>	
6	Extrémité 3'-ter du sucre	0	O3'	
		1	O3'-H	
		2	O3'-PO <sub>3</sub> <sup>2-</sup>	

Position du caractère	Partie modulée	Caractère	Module	Représentation du module
		3	O3'-HPO <sub>3</sub> <sup>-</sup>	
		4	O3'-CH <sub>3</sub> PO <sub>3</sub> <sup>-</sup>	
7	Connexion inter-nucléotidique (ne concerne pas les mono-nucléotides)	0	Aucune	/
		1	5' -> 3'	/
		2	5' -> 2'	/

Les structures nucléotidiques RAXN010, RCXN010, RGXN010 et RUXN010 ont été utilisées par défaut pour les calculs de docking dont les résultats sont présentés dans les parties III et IV des travaux de thèse. Chaque structure correspond à un ribonucléotide présentant un plissement du ribose C3'-endo, une base orientée en anti, un groupement phosphate PO<sub>2</sub><sup>-</sup> en 5'-ter et une extrémité 3'-ter définie par un O3' protoné.

Différentes structures nucléotidiques ont ensuite été utilisées pour les résultats présentés dans la partie V : RNXN010, RNXN110, RNXN210, RNXN310 et RNXN410, où N correspond aux bases A, C, G et U. Ces structures présentent un plissement du ribose et une orientation de la base identiques aux précédentes, mais elles diffèrent au niveau de l'extrémité 5'-terminale du ribose.

Notons que toutes les structures nucléotidiques accessibles sont optimisées, et se trouvent donc dans un minimum énergétique propre à leur type de conformation (*e.g.* ribose C3'-endo et base orientée anti). Leur énergie d'interaction est utilisée comme référence pour corriger l'énergie intramoléculaire des poses résultantes (voir section 1.5)

#### 1.4 Définition de l'espace d'échantillonnage

L'échantillonnage des poses a été restreint au site d'interaction des structures protéiques, site balisé par une boîte parallélépipédique. La définition de cette boîte pouvant différer en fonction des

jeux de données utilisés, elle est décrite plus spécifiquement dans la partie "Matériels et méthodes" des travaux correspondants.

### 1.5 Evaluation de l'énergie d'interaction

L'énergie d'interaction protéine-ligand calculée dans MCSS correspond à une énergie enthalpique d'interaction qui est estimée d'après le schéma suivant :

$$\Delta H_{interaction} = H_{PL} - H_P - H_L \quad (10)$$

avec  $\Delta H_{interaction}$ ,  $H_{PL}$ ,  $H_P$  et  $H_L$  les énergies enthalpiques d'interaction du complexe protéine-ligand, de la protéine, et du ligand, respectivement.

L'énergie enthalpique  $H_L$  du ligand correspond à une énergie intramoléculaire ( $H_{L-pose}$ ) corrigée par rapport à l'énergie d'un ligand de référence ( $H_{L-ref}$ ) dont la conformation est optimisée :

$$H_L = H_{L-pose} - H_{L-ref} \quad (11)$$

Toutes les composantes enthalpiques des équations 10 et 11 résultent de la somme de différentes énergies potentielles évaluées entre atomes liés et non-liés :

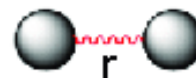
$$H = E_{lié} + E_{non-lié} \quad (12)$$

La somme des énergies associées aux atomes liés inclue les termes suivant :

$$E_{lié} = E_{liaison} + E_{angle} + E_{Urey-Bradley} + E_{dièdre} + E_{planarité} \quad (13)$$

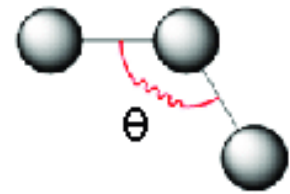
Le terme  $E_{liaison}$  évalue l'énergie associée à la longueur de liaison covalente ( $r$ ) entre deux atomes par rapport à une longueur de référence à l'équilibre ( $r_0$ ).

$$E_{liaison} = \sum_{liaisons} K_r (r - r_0)^2 \quad (14)$$



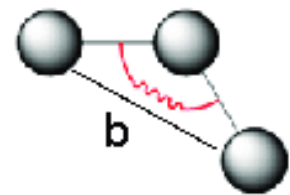
Le terme  $E_{angle}$  évalue l'énergie associée à l'angle de valence ( $\theta$ ) formé par trois atomes par rapport à un angle de valence de référence à l'équilibre ( $\theta_0$ ).

$$E_{\text{angle}} = \sum_{\text{angle}} K_{\theta} (\theta - \theta_0)^2 \quad (15)$$



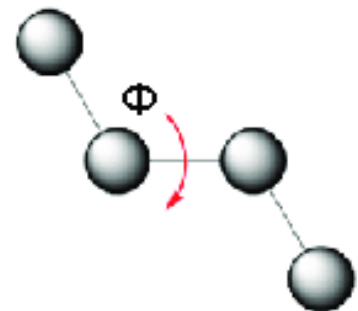
Le terme  $E_{\text{Urey-Bradley}}$  est complémentaire au terme  $E_{\text{angle}}$  ; il évalue l'énergie associée à la longueur ( $b$ ) séparant le premier atome du troisième atome dans un triplet d'atomes formant un angle de valence, toujours par rapport à une valeur de référence à l'équilibre ( $b_0$ ).

$$E_{\text{Urey-Bradley}} = \sum_{\text{Urey-Bradley}} K_{\text{UB}} (b - b_0)^2 \quad (16)$$



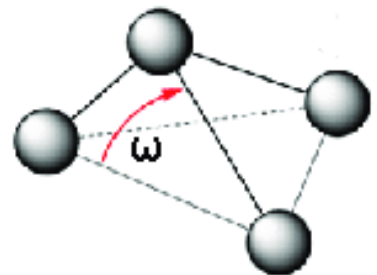
Le terme  $E_{\text{dièdre}}$  évalue l'énergie associée à un angle dièdre ( $\Phi$ ) formé par quatre atomes présentant une périodicité de rotation ( $n$ ) et un angle de phase ( $\delta$ ) donnés.

$$E_{\text{dièdre}} = \sum_{\text{dièdres}} K_{\phi} [1 + \cos(n\Phi - \delta)] \quad (17)$$



Le terme  $E_{\text{planarité}}$  évalue une énergie associée à la déformation du plan établi par certains groupes d'atomes tels les groupes aromatiques ou les atomes hybridés  $sp^2$ . L'énergie de l'angle du plan ( $\omega$ ) est corrigée par rapport à une valeur de référence à l'équilibre ( $\omega_0$ ).

$$E_{\text{planarité}} = \sum_{\text{planarités}} K_{\omega} (\omega - \omega_0)^2 \quad (18)$$



Remarquons que tous les calculs des énergies entre atomes liés sont corrigés par une constante de force K (K<sub>r</sub>, K<sub>θ</sub>, ...).

La somme des énergies associées aux atomes non-liés inclue les termes suivant :

$$E_{non-lié} = E_{LJ} + E_{elec} \quad (19)$$

Le terme E<sub>LJ</sub> correspond à un potentiel de Lennard-Jones :

$$E_{LJ} = \sum_{ij} \epsilon_{ij} \left( \frac{r_0^{ij}}{r_{ij}^{12}} \right) - 2 \epsilon_{ij} \left( \frac{r_0^{ij}}{r_{ij}} \right) \quad (20)$$


où ε<sub>ij</sub> est un paramètre du champ de force calculé pour chaque type d'atomes et qui définit la profondeur du puits du potentiel, r<sub>0ij</sub> est la combinaison des rayons de van der Waals des atomes i et j et définit la distance à laquelle l'énergie est minimale, et r<sub>ij</sub> est la distance séparant la paire d'atomes considérée.

Le terme E<sub>elec</sub> correspond à un potentiel coulombique :

$$E_{elec} = \sum_{ij} \frac{q_i q_j}{4 \epsilon_0 r_{ij}^2} \quad (21)$$


où q<sub>i</sub> et q<sub>j</sub> correspondent respectivement à la charge des atomes i et j, r<sub>ij</sub> la distance séparant la paire d'atomes considérée, et ε<sub>0</sub> est la constante diélectrique dont la valeur a été fixée à 3. La forme au carré de r<sub>ij</sub> rend compte de l'utilisation d'une constante diélectrique dépendante de la distance pour mimer l'effet d'écrantage du solvant sur les interactions électrostatiques.

Le calcul des énergies non-liées est effectué sur toutes les paires dont les atomes sont séparés de moins de 7,5 Å. Par ailleurs, pour l'évaluation de l'énergie intramoléculaire du ligand où les termes non-liés sont également inclus, seuls les atomes séparés par au moins quatre liaisons covalentes sont considérés comme non-liés.

Tous les paramètres utilisés pour le calcul énergétique (constantes de force, valeurs à l'équilibre, rayons de van der Waals, charges des atomes ...) sont tirés du champ de force CHARMM27. Les charges sont associées ponctuellement au centre de chaque atome ; celles portées par les atomes du groupement phosphate ont été réduites de manière à mimer implicitement la présence de contre-ions (Leclerc & Karplus, 1999).



## 2 Mesures de la déviation quadratique moyenne (RMSD)

Le RMSD a été calculé pour comparer les conformation et position des poses, soit entre elles deux à deux, soit par rapport à un ligand expérimental. Le RMSD est une mesure de distance exprimant la différence de conformation et de position entre une pose et une molécule de référence (ref) :

$$RMSD = \sqrt{\frac{1}{n} \sum_{i=1}^n [(pose_{ix} - ref_{ix})^2 + (pose_{iy} - ref_{iy})^2 + (pose_{iz} - ref_{iz})^2]} \quad (22)$$

où x, y et z correspondent aux coordonnées des atomes i de la pose et du nucléotide de référence (ref). A moins que cela ne soit explicitement précisé, les calculs RMSD sont effectués sur tous les atomes lourds à l'exception des atomes d'oxygène du groupement phosphate qui ont été écartés pour des raisons de symétrie. Globalement, une pose a été considérée comme reproduisant le mode d'interaction natif si son RMSD présentait une valeur inférieure ou égale à 2Å. Ce seuil correspond à une valeur standard couramment utilisée dans les applications de docking protéine-ligand.

## 3 Procédure de regroupement des poses (clustering)

Des étapes de clustering ont été appliquées aux poses issues du docking. Le clustering a été réalisé grâce au script *cluster\_struc* de la suite Haddock<sup>1</sup> (Jo et al., 2008). L'algorithme utilise la procédure de regroupement suivante : la pose ayant le plus de voisins (selon un seuil de distance RMSD défini) est regroupée, avec ses voisins, dans un premier cluster. Après un retrait des poses constituant ce premier groupe, la procédure est répétée sur l'ensemble des poses restantes, et ainsi de suite jusqu'à ce que l'ensemble des poses appartienne à un groupe. Lorsque le clustering est réalisé pour réduire le nombre de poses, seule la pose de plus basse énergie à l'intérieur de chaque groupe est choisie comme pose représentative. L'ensemble des poses représentatives est alors ordonné par ordre croissant de leur énergie d'interaction.

Le script *cluster\_struc* requiert en entrée une matrice de distance RMSD entre toutes les paires de poses (*matrice.rmsd*), un seuil de clustering (*seuil*) et la taille minimale des clusters (*taille*) à afficher en sortie :

```
cluster_struc matrice.rmsd seuil taille
```

Le RMSD a été utilisé comme mesure de distance entre les poses. Ainsi, une matrice RMSD a été d'abord calculée entre toutes les paires de poses possibles d'un système donné.

---

1 <http://www.bonvinlab.org/software/haddock2.2>

### **III Développement de l'approche FBDRNA lorsque la séquence ARN est connue**

#### **1 Introduction**

L'objectif abordé ici est de développer l'approche FBDRNA pour pouvoir répondre à la question suivante : étant données la structure d'un domaine de liaison à l'ARN (RBD), la connaissance de son site d'interaction, et une séquence ARN, quel est le mode d'interaction de la chaîne ARN avec le domaine ? La séquence ARN est supposée être sans élément de structures secondaires dans sa forme liée. L'échantillonnage de l'espace conformationnel de cette séquence est traitée en fragmentant l'ARN en ses unités élémentaires : les nucléotides. Chaque nucléotide de la séquence est représenté par une unique structure pré-définie et utilisée comme ligand de départ dans les simulations de docking réalisées par MCSS. Le principe général de l'approche peut se décliner en trois grandes étapes (Fig. 30) :

1. chaque ligand nucléotidique correspondant aux nucléotides composant la séquence ARN est distribué dans des positions et orientations aléatoires à la surface du domaine restreinte à son site d'interaction.
2. une sélection de poses est établie pour chaque distribution obtenue.
3. des chaînes ARN potentielles sont recherchées à partir de ces poses en identifiant avec le programme Molpy celles pouvant être connectées entre elles par satisfaction de contraintes (distances, séquence, angles).

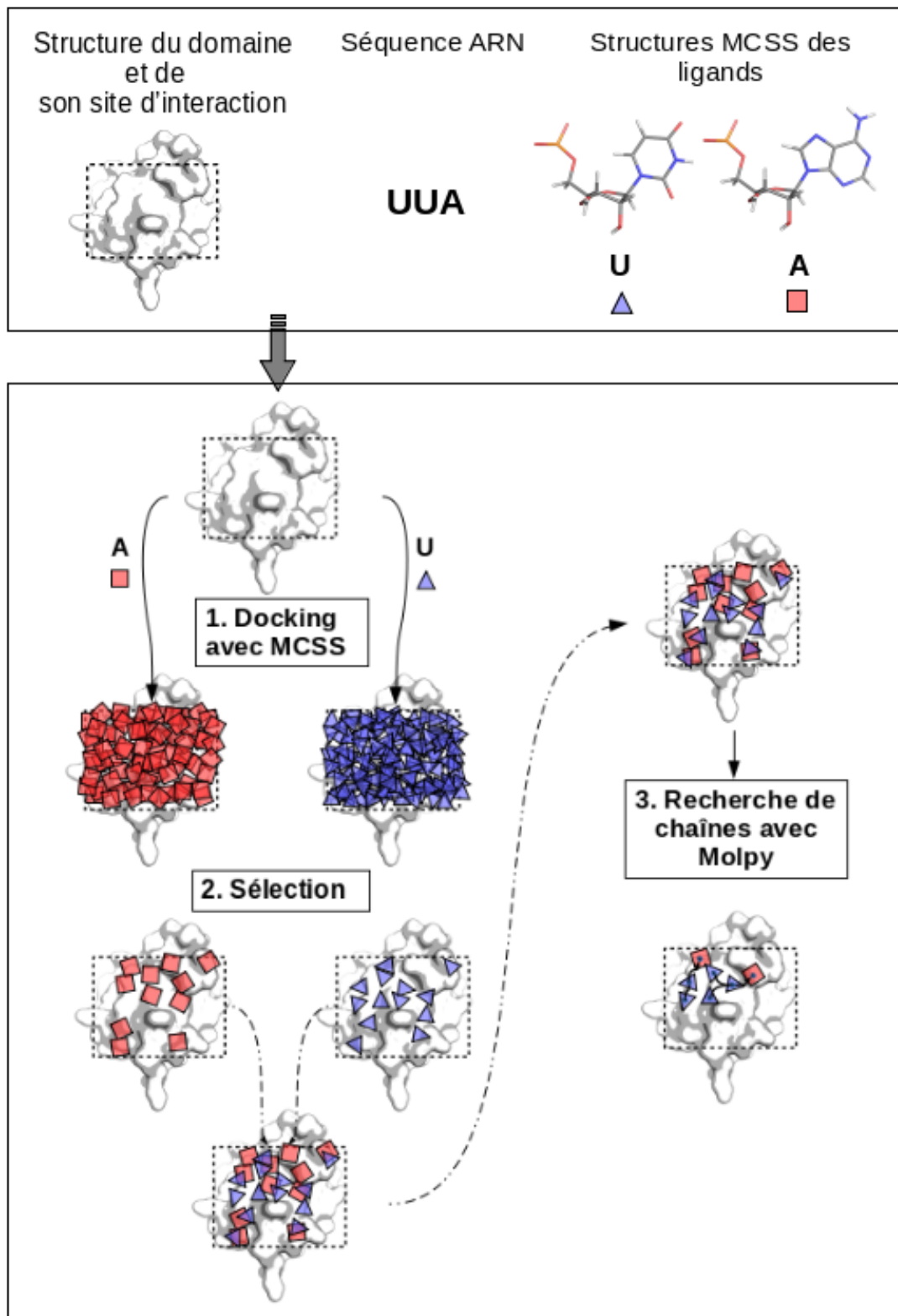


Figure 30: Illustration du principe de l'approche FBDRNA pour la prédiction du mode d'interaction entre un domaine de liaison à l'ARN (RBD) et une chaîne ARN à partir sa séquence UUA. Le cadre du haut contient les données requises et utilisées en entrée : la structure d'un RBD avec la connaissance de son site d'interaction, une séquence ARN (ici UUA) et les structures de ligands nucléotidiques utilisées pour le docking avec MCSS. Le protocole du cadre du bas est décrit dans le texte.

Trois complexes RBD-ARNsb ont été utilisés pour le développement de l'approche. Chaque complexe provient d'une structure cristallographique à haute résolution ( $< 2 \text{ \AA}$ ). Chaque complexe implique des RBDs appartenant aux trois familles les plus représentées chez l'Homme (Gerstberger, Hafner, & Tuschl, 2014) :

- le domaine RRM de la protéine Nab3 lié à une chaîne UCU (Lunde, Hörner, & Meinhart, 2011). Son code PDB est 2XNR.
- un groupe de trois domaines de doigts à zinc CCCH (Zn-CCCH) de la protéine Unkempt lié à une chaîne UUAUU (Murn, Teplova, Zarnack, Shi, & Patel, 2016). Son code PDB est 5ELH.
- le domaine KH2 de la protéine MEX-3C lié à une chaîne CAGAGCU (Yang et al., 2017). Son code PDB est 5WWX.

Les critères de sélection de ces complexes sont décrits dans la section 2.1. L'un des critères utilisés correspond à la disponibilité de données d'affinité relatives à différents variants de séquences (informations nécessaires pour la réalisation des travaux présentés dans le chapitre suivant). Pour les trois domaines, les variations ne concernent que trois nucléotides successifs dans la séquence ARN qui sont directement impliqués dans la reconnaissance spécifique. Pour cette raison, seuls ces triplets nucléotidiques ont été retenus pour définir les chaînes d'ARN utilisées comme référence. Précisons également que les nucléotides écartés soit ne présentent pas de densité électronique claire, soit n'établissent que peu voire aucun contact avec le domaine de l'unité asymétrique. Les informations relatives aux trois systèmes considérés sont résumées dans le tableau 3 et leur structure est représentée à la figure 31.

*Tableau 3: Caractéristiques des complexes RBD-ARN utilisés comme références.*

Type	code PDB	Résolution (Å)	Ligand (5' → 3')	Référence
RRM	2XNR	1,6	UCU	(Lunde, Hörner, & Meinhart, 2011)
Zn-CCCH	5ELH	1,8	UUA	(Murn, Teplova, Zarnack, Shi, & Patel, 2016)
KH	5WWX	2,0	AGA	(Yang et al., 2017)

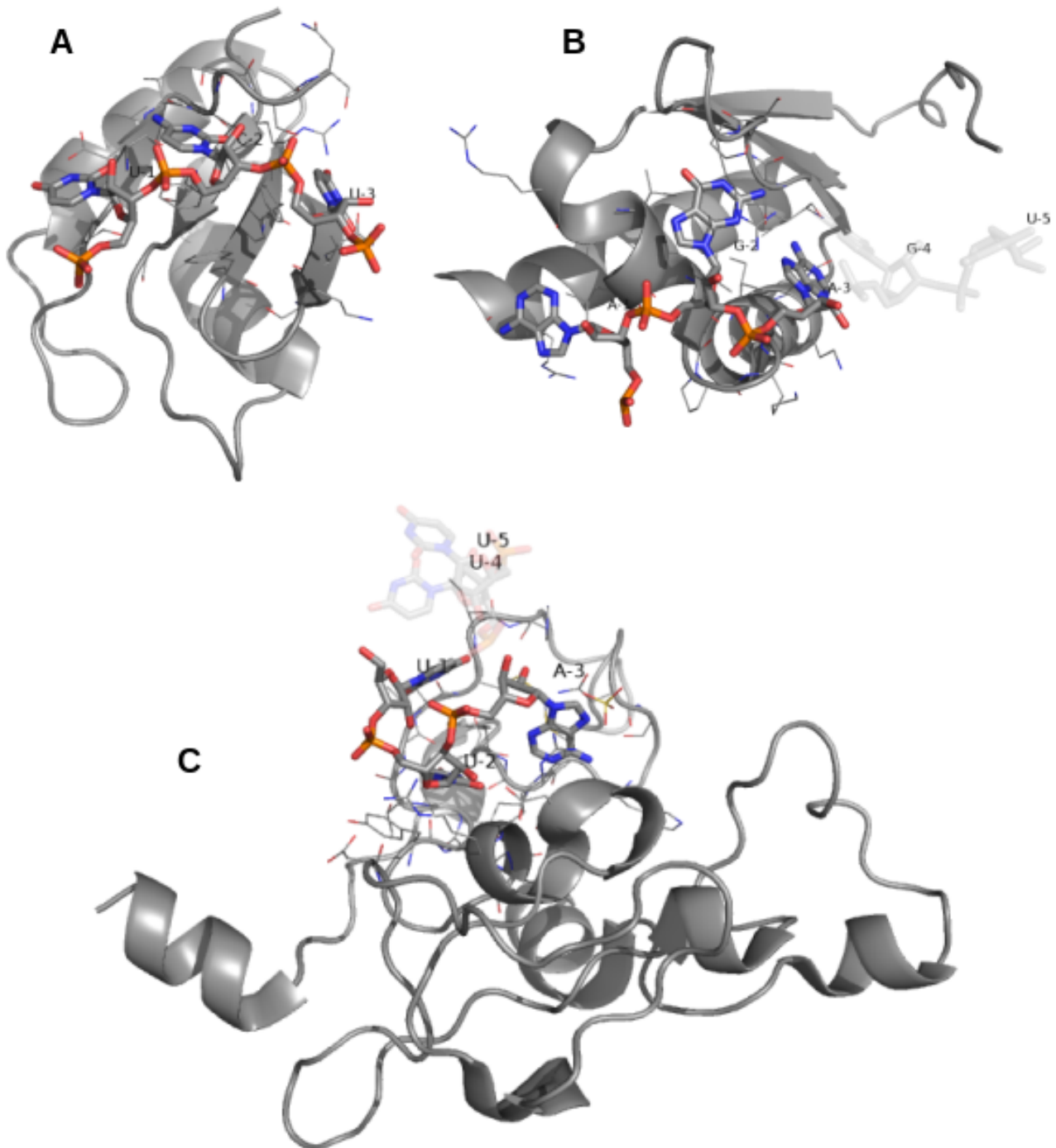


Figure 31 : Structures cristallographiques retenues pour le développement de l'approche FBDRNA. (A) Structure du domaine RRM de la protéine Nab3 lié à un ARNsb de séquence 5'-UCU ; code PDB 2XNR. (B) Structure du domaine KH2 de la protéine MEX-3C lié à un ARNsb de séquence 5'-AGAGU ; code PDB : 5WWX. (C) Structure des domaines Zn-CCCH de la protéine Unkempt liés à un ARNsb de séquence 5'-UUAUU ; code PDB : 5ELH. Les domaines sont représentés en mode cartoon et de couleur grise. Les nucléotides sont représentés en mode bâtonnet ; leur numérotation débute en 5' des chaînes ARN ; les atomes de carbone sont en gris, les oxygènes en rouge, les azotes en bleu et le phosphore en orange. Les nucléotides écartés de la sélection sont en mode transparente. Nous voyons clairement qu'ils n'établissent pas ou très peu de contacts avec chacun des domaines protéiques.

Le développement de l'approche FBDRNA s'est essentiellement focalisé sur l'étape de sélection. Cette dernière joue un rôle capital dans le succès de l'approche qui peut se définir par la capacité de la méthode à reproduire le mode d'interaction expérimental des chaînes ARN. Cette étape de sélection repose cependant au préalable sur la qualité des résultats du docking, ces derniers ont donc été préalablement analysés de façon systématique pour chacun des trois systèmes.

## **2 Matériels et méthodes**

### **2.1 Sélection du jeu de données**

Pour mettre au point et tester la validité de l'approche prédictive, l'attention a été portée sur trois RBDs appartenant chacun à une des trois familles les plus étudiées et les plus représentées chez l'Homme : les domaines RRM, KH et Zn-CCCH. Trois complexes RBD-ARNsb ont été recherchés dans la PDB, chacun devant représenter l'une de ces trois familles de RBDs. D'une manière générale, les recherches dans la PDB ont été basées sur les critères suivants :

- la méthode développée ayant pour objet la prédiction du mode d'interaction entre RBD et ARNsb, les complexes à sélectionner doivent mettre en jeu uniquement un module unitaire d'interaction avec une chaîne d'ARNsb (un domaine pour RRM et KH, un groupe de trois domaines pour Zn-CCCH)
- la chaîne d'ARNsb doit inclure au moins trois nucléotides en interaction directe avec le domaine
- la méthode développée ayant également pour objet la prédiction de séquences ARN préférentiellement reconnues par un RBD donné, des données d'affinité relatives à différentes séquences ARN doivent être disponibles dans la littérature
- des complexes cristallographiques résolus à haute résolution ( $\leq 2 \text{ \AA}$ ) ont été privilégiés
- dans le cas où plusieurs complexes répondent aux critères énumérés ci-dessus, pour réduire les temps de calculs, seuls ceux pour lesquels le RBD mis en jeu est également disponible sous forme non-liée ont été privilégiés de manière à pouvoir analyser l'influence potentielle de changements conformationnels sur les résultats (travaux qui n'ont pas encore été réalisés au moment de la rédaction de ce manuscrit et qui n'y seront donc pas présentés).

Pour la recherche de complexes impliquant les domaines RRM et KH, les requêtes PDB ont plus précisément été réalisées d'après les règles suivantes :

- Numéro d'accession PFAM : PF00076 (pour RRM) / PF00013 (pour KH)

- Nombre de chaînes : entre 1 et 2
- Type de macromolécule :
  - contient une protéine : Oui
  - contient de l'ADN: Non
  - contient de l'ARN : Oui
  - contient un hybride ADN/ARN : Non
- Résolution rayon X : entre 0 et 2 Å
- Longueur de chaîne : entre 1 et 100 résidus

A la date du 09/06/2018, cette requête a conduit à identifier six domaines RRM : 2XS5, 2XS7, 2XNR, 2X1F et 1URN. A l'exception de 1URN, tous ces complexes RBD-ARN contiennent des données d'affinité pour des variants ARN et sont accessibles sous forme non-liée. La structure 2XNR a été privilégiée *a posteriori* de manière à diversifier les séquences d'ARN du jeu de données final. La chaîne ARN de ce complexe contient en effet un C déterminant dans la spécificité de reconnaissance, nucléotide qui n'est pas présent dans les chaînes des deux autres complexes sélectionnés.

La requête associée au domaine KH a quant à elle conduit à identifier deux complexes : 5WWW et 5WWX. Des données d'affinité sont disponibles pour ces deux complexes, cependant, seule la structure 5WWX bénéficie également d'une structure résolue sous forme non-liée ; elle a donc été privilégiée sur 5WWW.

Le nombre de complexes résolus impliquant les domaines Zn-CCCH étant moins important, tous les numéros d'accèsion PFAM associés au clan de ce domaine ont été donnés en entrée dans la requête : PF00642, PF14608, PF15663, PF18260, PF16131, PF18044, PF18585, PF18586, PF18633, PF18345, PF18384. Par ailleurs, ce domaine étant souvent répété en tandem, aucun critère de longueur de chaîne n'a été imposé. Trois complexes ont été identifiés : 5L2L, 5ELH et 3D2S. Le complexe 5ELH a été sélectionné puisqu'il est le seul bénéficiant de données d'affinité pour différentes séquences ARN.

Les informations relatives aux trois complexes ainsi retenus sont résumées dans le tableau 4. La chaîne nucléotidique cristallisée avec les domaines RRM, KH et Zn-CCCH contient respectivement trois, sept et cinq nucléotides.

Les données d'affinité de liaison pour 2XNR concernent trois variants de séquence et proviennent de mesures de variation d'anisotropie de fluorescence (Lunde, Hörner, & Meinhart, 2011) ; elle peuvent se résumer ainsi : UCU (110  $\mu\text{M}$ ) > CCC (250  $\mu\text{M}$ ) > UUU (350  $\mu\text{M}$ ).

Des affinités de liaison pour cinq variants de séquence sont disponibles pour 5ELH (Murn, Teplova, Zarnack, Shi, & Patel, 2016). Les données, issues de titrages par calorimétrie isotherme peuvent se résumer ainsi : UUA (0,6  $\mu\text{M}$ ) > GUU (1,2  $\mu\text{M}$ ) > AAU (2,1  $\mu\text{M}$ ) > GUG (mesure non déterminée) > GCG (pas de liaison observée).

Pour 5WWX, des données d'affinité de liaison, également issues de titrages par calorimétrie isotherme, sont accessibles pour 10 variants de séquence (Murn, Teplova, Zarnack, Shi, & Patel, 2016) : AGA (0,17  $\mu\text{M}$ ) > UGA (0,35  $\mu\text{M}$ ) > GGA (0,36  $\mu\text{M}$ ) > AUA (0,40  $\mu\text{M}$ ) > CGA (0,52  $\mu\text{M}$ ) > ACA (0,56  $\mu\text{M}$ ) > AAA (0,58  $\mu\text{M}$ ) > AGG (1,48  $\mu\text{M}$ ) > AGU (1,69  $\mu\text{M}$ ) > AGC (mesure non déterminée).

Les valeurs données entre parenthèses indiquent les Kd mesurés. Pour plus de détails sur l'ensemble de ces données d'affinité, le lecteur est renvoyé aux articles correspondants.

Les variants de séquence pour lesquels des données d'affinité sont disponibles n'impliquent que trois nucléotides. Par conséquent seuls les triplets correspondants ont été retenus comme référence pour chaque complexe (indiqués en gras dans le tableau 4). A noter que les autres nucléotides qui ne sont pas considérés n'établissent que peu voire pas de contacts avec l'unité asymétrique du domaine (Fig. 31).

*Tableau 4: Liste des complexes RBD-ARNsb sélectionnés. Les nucléotides en gras dans la séquence du ligand cristallisé sont ceux conservés et utilisés comme référence.*

Type	code PDB	Résolution (Å)	Ligand dans le cristal (5' → 3')	Référence
RRM	2XNR	1,6	<b>UCU</b>	(Lunde, Hörner, & Meinhart, 2011)
KH	5WWX	2,0	<b>CAGAGCU</b>	(Yang et al., 2017)
Zn-CCCH	5ELH	1,8	<b>UUAUU</b>	(Murn, Teplova, Zarnack, Shi, & Patel, 2016)

## 2.2 Simulations de docking

Chaque chaîne ARN de référence étant composée de deux nucléotides de bases différentes, deux calculs de docking ont été réalisés indépendamment sur chaque domaine : les nucléotides U et C ont



ainsi été dockés pour le domaine RRM, les nucléotides A et G pour le domaine KH, et les nucléotides U et A pour le groupe de domaines Zn-CCCH. Les structures nucléotidiques RAXN010, RCXN010, RGXN010 et RUXN010 ont été utilisées par défaut. Chaque structure correspond à un ribonucléotide présentant un plissement du ribose C3'-endo, une base orientée en anti, un groupement phosphate PO<sub>2</sub><sup>-</sup> en 5'-ter et une extrémité 3'-ter définie par un O3' protoné. Pour chaque domaine, les nucléotides d'intérêt ont été dockés à l'intérieur d'une boîte parallélépipédique englobant le site d'interaction avec la chaîne ARN tri-nucléotidique. Cette boîte a été élargie de 5 Å dans les coordonnées x, y et z par rapport aux extrémités des chaînes (Fig. 34).

La préparation des protéines, les paramètres utilisés pour les simulations et l'évaluation énergétique des poses sont tels qu'indiqué dans la section 1 de la partie "Méthodes générales".

## 2.3 Molpy

### 2.3.1 Recherche de chaînes

La recherche de chaînes est effectuée par le programme Molpy développé par Manuel Simoes (<https://github.com/mianuel/Molpy>). Molpy prend en entrée une liste de poses préalablement sélectionnées. Ces dernières sont d'abord parcourues par un algorithme qui recense dans une matrice les paires de nucléotides pouvant être connectées et ne présentant pas de conflits stériques. La connexion est considérée possible si quatre contraintes de distances sont respectées. Ces distances sont précisées dans le tableau 5.

Tableau 5: Contraintes de distances utilisées pour autoriser une connexion entre deux nucléotides

Nucléotide i	Nucléotide j	Distance minimale (Å)	Distance maximale (Å)
Base	Base	4	15
Ribose	Ribose	4	8,5
Ribose	Base	4	-
O3'	C5'	2	7

Lorsque la base et/ou le ribose sont considérés, leur distance est mesurée à partir de leur centre géométrique. Les intervalles de distance ont été définis sur la base de mesures effectuées sur diverses ARN (aussi bien sous la forme simple-brin que double-brin) tirés de 276 complexes protéine-ARN (les critères utilisés pour sélectionner ces complexes n'étant pas précisés dans Molpy, ils me sont inconnus) : 1I5L, 1AQ3, 1AQ4, 1B23, 1C9S, 1CWP, 1D6K, 1DFU, 1DK1, 1E7K, 1E8O, 1E1Y, 1F7Y, 1FEU, 1GTF, 1GTN, 1H3E, 1I6U, 1KQ2, 1KUQ, 1L9A, 1LAJ, 1LNG, 1MJI, 1MMS, 1MZP, 1OB2, 1OB5, 1OLN, 1P6V, 1PVO, 1QF6, 1QZW, 1RC7, 1RLG, 1RPU, 1S03, 1SDS, 1SZ1, 1TFW, 1TFY, 1TTT, 1U1Y, 1U63, 1UTD, 1UTF, 1UTV, 1VOX, 1WMQ, 1WPU, 1WRQ, 1X1L, 1XPO, 1XPR, 1XPU, 1YTU, 1YVP, 1YYK, 1YYO, 1YYW, 1YZ9,

1ZC8, 1ZDH, 1ZDI, 1ZDJ, 1ZDK, 1ZHO, 1ZSE, 2B2D, 2B2E, 2B2G, 2BGG, 2BNY, 2BQ5, 2BS0, 2BS1, 2BU1, 2C4Q, 2C4Y, 2C4Z, 2C50, 2C51, 2DER, 2DET, 2DEU, 2DR5, 2DR7, 2DR8, 2DR9, 2DRA, 2DRB, 2DVI, 2EZ6, 2GIC, 2GTT, 2HT1, 2HVY, 2HW8, 2I91, 2IX1, 2IZ8, 2IZ9, 2IZM, 2IZN, 2JEA, 2JPP, 2JQ7, 2MF0, 2MF1, 2MFC, 2MFE, 2MFF, 2MFG, 2MFH, 2NOQ, 2NUE, 2NUF, 2NUG, 2OB7, 2OGM, 2OGN, 2OGO, 2OZB, 2Q66, 2R7R, 2R7S, 2R7T, 2R7U, 2R7V, 2R7W, 2R7X, 2VAZ, 2WJ8, 2WYY, 2X7N, 2XLI, 2XLJ, 2XLK, 2XXA, 2Y8W, 2Y8Y, 2Y9H, 2YHM, 2ZH1, 2ZH2, 2ZH3, 2ZH4, 2ZH5, 2ZH6, 2ZH7, 2ZH8, 2ZH9, 2ZHA, 2ZHB, 2ZKR, 3A6P, 3AEV, 3AHU, 3AVT, 3AVU, 3AVV, 3AVW, 3AVX, 3AVY, 3BOY, 3DH3, 3FTE, 3FTF, 3HAX, 3HAY, 3HHZ, 3HJW, 3HL2, 3HSB, 3HTX, 3ICE, 3ICQ, 3IE1, 3IEM, 3IEV, 3IYQ, 3IYR, 3IZZ, 3KTW, 3LWO, 3LWP, 3LWQ, 3LWR, 3LWV, 3MDG, 3MDI, 3NVI, 3OHJ, 3OHZ, 3OI1, 3OI3, 3OIJ, 3OIN, 3OV7, 3OVA, 3OVB, 3OVS, 3PKM, 3PTX, 3PU0, 3PU1, 3PU4, 3Q1Q, 3Q1R, 3Q2T, 3QJJ, 3QJL, 3QJP, 3QRP, 3QRR, 3QSU, 3QSY, 3R2C, 3R2D, 3R9W, 3SIU, 3SIV, 3SN2, 3SNP, 3U4M, 3U56, 3UMY, 3V7E, 3VNU, 3VNV, 3WBM, 3ZLA, 4AL5, 4AL6, 4AL7, 4AM3, 4ANG, 4BA2, 4BBL, 4BHH, 4BKK, 4BW0, 4BYQ, 4C4W, 4EYA, 4F3T, 4FWT, 4G0A, 4HOR, 4HOS, 4HOT, 4IJS, 4J7L, 4J7M, 4JNG, 4KR6, 4KR7, 4KR9, 4KRE, 4KRF, 4KXT, 4KZX, 4KZZ, 4L8H, 4LCK, 4M2Z, 4M30, 4MDX, 4OAU, 4OAV, 4OLA, 4OLB, 4OO8, 5MSF, 6MSF, 7MSF.

Dans un second temps, la matrice de connexion est parcourue pour caractériser l'espace des chaînes possibles en identifiant les paires de poses ayant un nucléotide commun et ne présentant pas de conflits stériques. Les longueurs minimale et maximale de la chaîne à modéliser sont définies par l'utilisateur qui peut aussi choisir de générer une séquence d'intérêt en spécifiant sa composition en nucléotides. Pour les chaînes d'au moins trois nucléotides, un angle de courbure est évalué de manière à éviter que les chaînes ne se replient sur elles-mêmes.

### 2.3.2 Optimisation des chaînes

Les chaînes identifiées à l'étape précédente sont constituées de nucléotides se trouvant chacun dans un minimum énergétique local indépendant. Par ailleurs, les intervalles relativement larges des distances utilisées pour autoriser une connexion inter-nucléotidique implique que certains atomes séparant deux nucléotides contigus dans une chaîne peuvent être séparés par une distance aberrante. Molpy intègre donc une étape finale d'optimisation afin de corriger ces distances (Fig. 32).

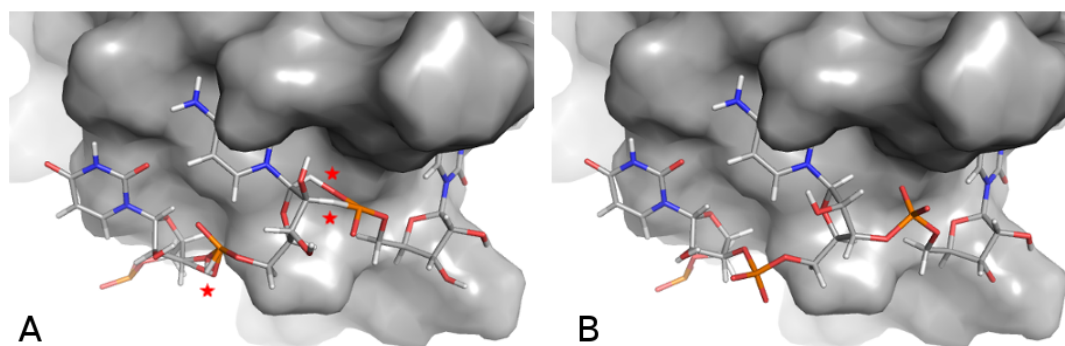


Figure 32: Illustration de la correction des distances aberrantes approtées par la minimisation. (A) Chaîne avant minimisation ; les distances aberrantes sont mises en évidence par une étoile rouge. (B) Chaîne après minimisation ; les distances sont corrigées.

Cette optimisation est réalisée par plusieurs cycles de minimisation permettant également d'optimiser l'ensemble des contacts protéine-ARN. Le processus se déroule comme suit :

- La connexion entre nucléotides est d'abord corrigée par 1000 étapes de minimisation selon la méthode de plus grande pente (SD), suivis de 1000 étapes de minimisation selon la méthode de Newton-Raphson (ABNR). Durant ces étapes, tous les atomes de la protéine sont rigides. Pour l'ARN, seuls les atomes du groupement phosphate ainsi que les atomes C3' et C5' sont libres de tout mouvement.
- La chaîne d'ARN est ensuite optimisée par 1000 étapes de minimisation selon la méthode SD suivis de 3000 étapes de minimisation selon la méthode ABNR. Durant ces étapes, tous les atomes de l'ARN sont flexibles, ceux de la protéine sont maintenus fixes.
- Les contacts entre la chaîne d'ARN et la protéine sont finalement optimisés par 1000 étapes de minimisation selon la méthode ABNR durant laquelle et l'ARN et la protéine sont flexibles.

Les chaînes sont ensuite triées selon leur énergie d'interaction, de la plus favorable à la moins favorable. Précisons que l'optimisation et l'évaluation énergétique sont effectuées avec CHARMM et les paramètres du champ de force CHARMM27. L'énergie d'interaction est estimée selon le même schéma que celui adopté dans le scoring MCSS (section 1.5), à l'exception qu'aucune correction par rapport à la conformation de la chaîne n'est ici apportée.

## **2.4 Analyse de la conformation des nucléotides**

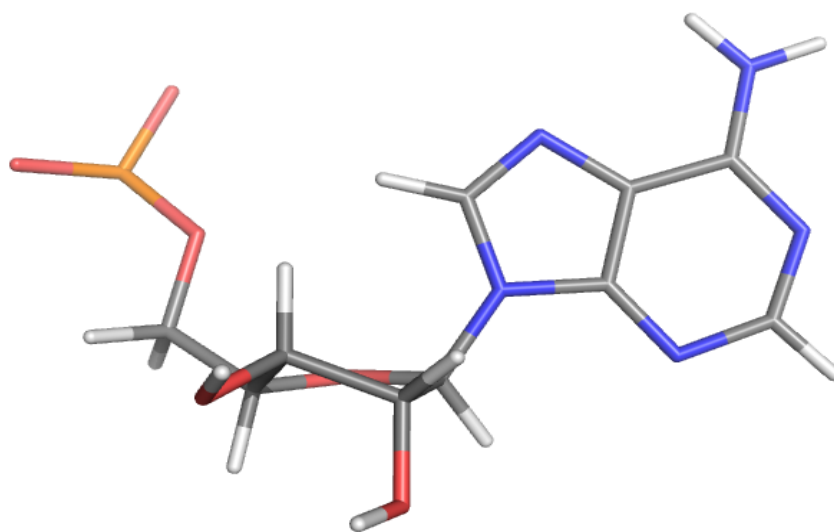
Le programme 3DNA (Lu & Olson, 2008) a été utilisé pour définir la conformation des nucléotides constituant les chaînes ARN tri-nucléotidiques expérimentales ainsi que les poses issues du docking. Le type de plissement du ribose et l'orientation de la base sont définis d'après les paramètres par défaut de 3DNA.

# **3 Résultats**

## **3.1 Analyse des performances de l'étape de docking**

Pour chacun des trois systèmes utilisés comme référence, la séquence ARN à modéliser est composée de nucléotides de deux types de bases différents. Deux calculs de docking indépendants ont donc été réalisés avec MCSS, un pour chaque type de base (Fig. 30). Au cours du docking, un nucléotide dans une conformation pré-définie est distribué aléatoirement dans une boîte parallélépipédique qui englobe le site d'interaction du domaine considéré. Cette conformation présente un plissement du ribose en C3'-endo et la base est orientée en anti (Fig. 33); elle est utilisée

par défaut dans MCSS et correspond à une conformation standard des nucléotides observés dans les ARN simple-brin structurés en hélice de forme A.



*Figure 33: Illustration de la conformation nucléotidique utilisée pour le docking. Une adénine est ici représentée à titre d'exemple, les autres bases présentant la même conformation. Le ribose présente un plissement en C3'-endo et la base est orientée en anti. Les atomes de carbone sont en gris, d'hydrogène en blanc, d'oxygène en rouge et le phosphore en orange.*

La conformation pré-définie du nucléotide docké est la même quel que soit le type de base utilisé.

La boîte parallélépipédique a été définie par rapport aux extrémités des chaînes 3-nt utilisées comme référence (Fig. 34), extrémités auxquelles ont été ajoutées un espace de 5 Å dans les coordonnées x, y, et z. Cette définition représente une situation où le site d'interaction est parfaitement déterminé. En pratique, sans connaissance précise de ce dernier et en l'absence de données expérimentales, des programmes de prédiction de site d'interaction peuvent être utilisés (Si, Cui, Cheng, & Wu, 2015b).

Le premier gage de succès de l'approche FBDRNA repose sur la qualité des résultats de docking qui peut se décliner en deux étapes : l'échantillonnage des poses et l'évaluation de leur énergie d'interaction (ou scoring). Ces deux aspects ont donc été analysés.

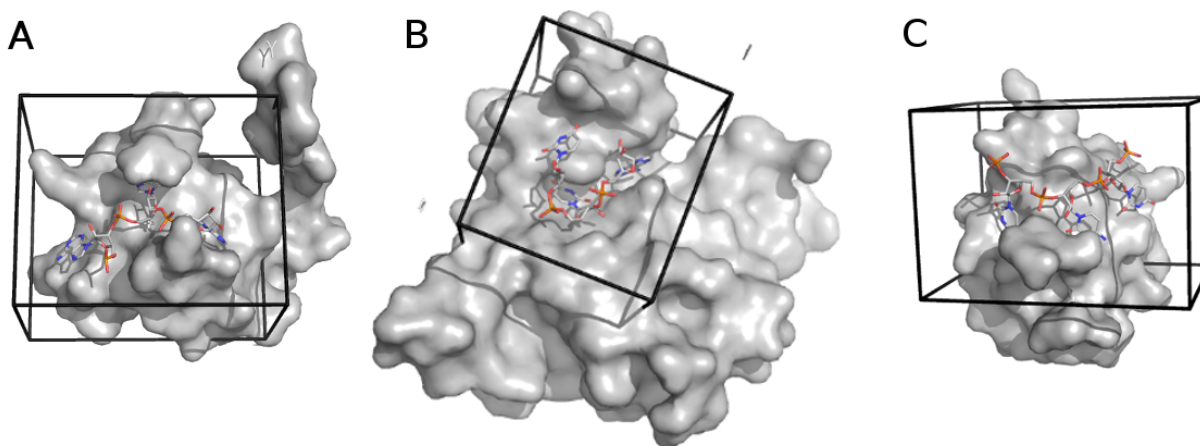


Figure 34: Illustration de l'espace d'échantillonnage utilisé sur les trois domaines: 2XNR (A), 5ELH (B) et 5WWX (C).

### 3.1.1 Evaluation de l'échantillonnage

Pour pouvoir être en mesure de reproduire le mode d'interaction natif des chaînes ARN, l'étape d'échantillonnage du docking doit en premier lieu reproduire le mode d'interaction de chacun des nucléotides composant leur séquence. Les poses natives sont définies par un RMSD  $\leq 2$  Å par rapport au nucléotide expérimental, seuil standard couramment utilisé dans le docking protéine-ligand pour valider la reproduction du mode d'interaction natif. Le RMSD a été calculé sur tous les atomes lourds à l'exception des oxygènes du groupement phosphate. Le tableau 6 montre que la phase d'échantillonnage permet de générer des poses natives pour chacun des nucléotides qui composent les trois chaînes d'ARN testées.

Comme mentionné précédemment, chacun des calculs de docking a été réalisé à partir d'une conformation pré-définie du nucléotide où le plissement du ribose est en C3'-endo et la base orientée en anti. Cette conformation initiale ne correspond pas systématiquement à la conformation expérimentale du nucléotide (tableau 6) : les trois nucléotides de la chaîne ARN de 2XNR sont dans une conformation C2'-endo/anti, de même que les nucléotides A1 et G2 de 5WWX ; seuls les nucléotides A3 de 5WWX et U2 de 5ELH sont en C3'-endo/anti ; les nucléotides U1 et A3 de 5ELH sont dans une conformation moins standard : C1'-exo/anti et C2'-endo/syn, respectivement. Au cours de l'échantillonnage, chaque pose est optimisée dans le champ de force de la protéine par plusieurs cycles de minimisation durant lesquels les degrés de liberté de chacun des angles de torsion nucléotidiques peuvent être explorés. Afin de pouvoir évaluer dans quelle mesure cette procédure permet l'exploration de diverses conformations, les conformations de la totalité des poses

issues de l'ensemble des calculs de docking effectués sur les trois RBDs (deux calculs de docking par RBD) ont été recensées.

*Tableau 6: Caractéristiques des poses natives obtenues pour chaque nucléotide composant la chaîne ARN des trois RBDs. Le nombre total de poses natives est indiqué, avec entre parenthèses leur proportion par rapport au nombre total de poses générées. Parmi ces poses natives, le nombre de poses présentant une conformation (plissement du ribose et orientation de la base) identique au nucléotide expérimental est précisé. La conformation de la pose native ayant le plus petit RMSD par rapport au nucléotide expérimental est également indiqué ; est mise en évidence en gras italique la situation où son plissement du ribose et/ou l'orientation de sa base est/sont conforme(s) au nucléotide expérimental.*

Code PDB	Nucléotide expérimental		Pose native ( $\leq 2 \text{ \AA}$ )			Total de poses
	Position	Conformation	Total (%)	Nb dans la conformation expérimentale	RMSD minimal ( $\text{\AA}$ )	
2XNR	U1	C2'-endo/anti	18 (3,51)	3	1,25 (C3'-endo/ <i>anti</i> )	5128
	C2	C2'-endo/anti	23 (5,83)	4	0,66 ( <i>C2'-endo/anti</i> )	3947
	U3	C2'-endo/anti	13 (2,54)	4	1,06 ( <i>C2'-endo/anti</i> )	5128
5ELH	U1	C1'-exo/anti	3 (0,71)	0	1,82 (C3'-endo/ <i>anti</i> )	4239
	U2	C3'-endo/anti	14 (3,30)	5	0,51 ( <i>C3'-endo/anti</i> )	4239
	A3	C2'-endo/syn	2 (0,47)	0	1,54 (C3'-endo/ <i>syn</i> )	4234
5WWX	A1	C2'-endo/anti	12 (1,58)	1	1,42 (C3'-endo/ <i>anti</i> )	7587
	G2	C2'-endo/anti	35 (5,07)	6	0,48 ( <i>C2'-endo/anti</i> )	6897
	A3	C3'-endo/anti	17 (2,24)	17	0,86 ( <i>C3'-endo/anti</i> )	7586

La figure 35 montre que la conformation initiale de la structure utilisée pour le docking est majoritaire, mais elle montre également que 23 autres conformations ont pu être explorées. La diversité des conformations observées montre donc que l'étape de minimisation permet un franchissement des barrières énergétiques séparant différents angles de torsions. L'exploration de l'espace conformationnel du nucléotide n'est ainsi pas réduite à la conformation initiale utilisée pour le docking.

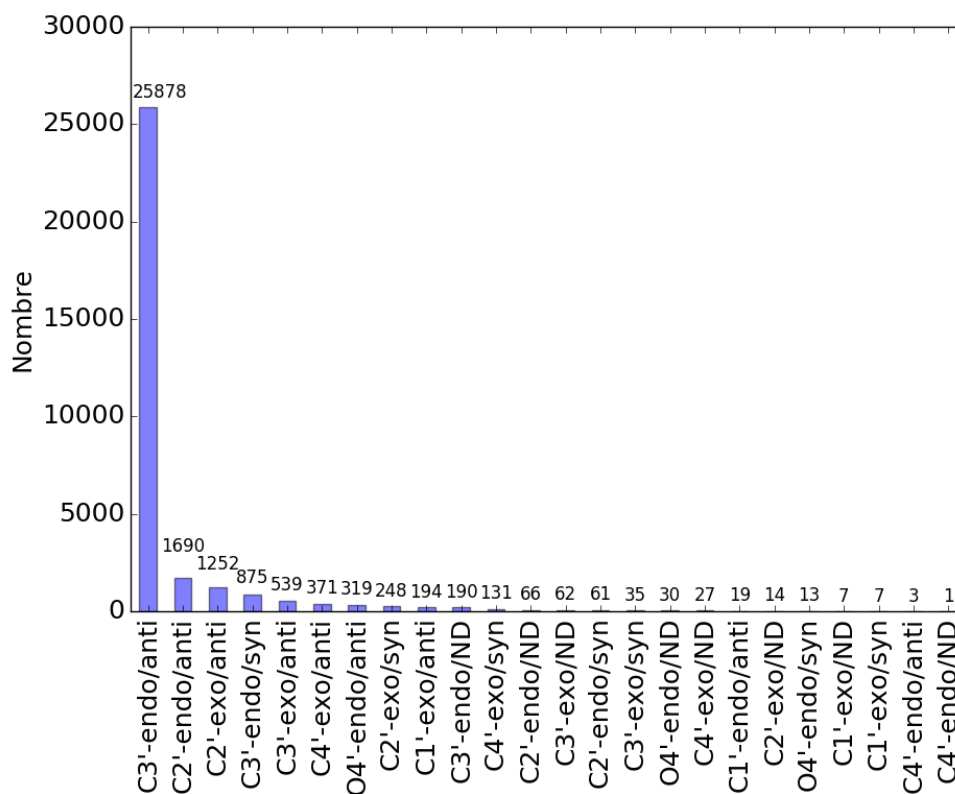


Figure 35: Conformations nucléotidiques observées parmi la totalité des poses générées pour l'ensemble des six calculs de docking effectués (deux nucléotides par domaine). Les conformations correspondent à une combinaison du type de plissement du ribose et de l'orientation de la base. Le dénombrement est effectué sur l'ensemble des calculs de docking réalisés sur les trois RBDs pour chacun des nucléotides de la séquence des chaînes ARN expérimentales. La conformation majoritaire C3'-endo/anti correspond à celle de la structure du ligand nucléotidique utilisé pour le docking. ND indique que l'orientation de la base n'est pas définie selon les critères d'assignation utilisés.

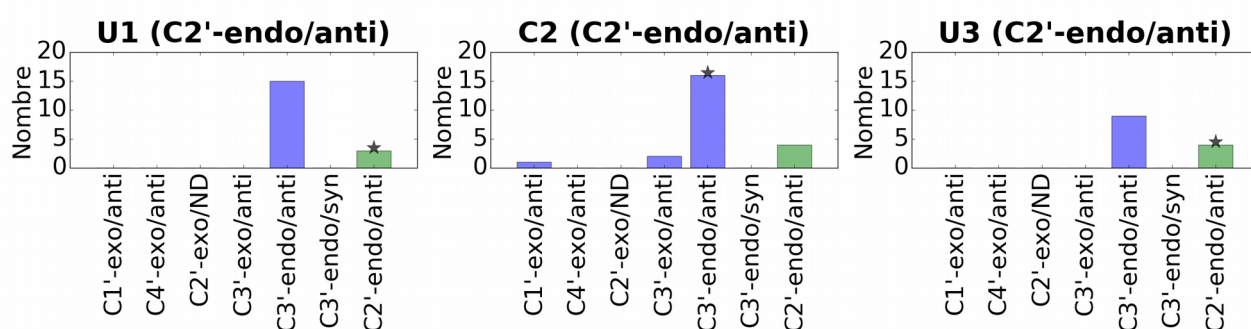
Cette observation est faite sur la totalité des poses, mais qu'en est-il pour les poses natives ? Les conformations retrouvées pour ces dernières correspondent-elles à la conformation des nucléotides expérimentaux ? La figure 36 recense le nombre de chacune des conformations observées parmi l'ensemble des poses natives. La conformation exacte (plissement et orientation de la base corrects) est retrouvée parmi les poses natives pour sept nucléotides sur neuf (barres vertes de la figure 36). Par ailleurs, dans cinq cas sur sept, la pose de plus basse énergie (représentée par une étoile à la figure 36) est dans la conformation exacte. Pour 5ELH, seul le nucléotide U2 a pu être reproduit dans sa conformation expérimentale. Pour ses nucléotides U1 et A3, une seule conformation a été échantillonnée parmi les poses natives ; le plissement du ribose n'est pas retrouvé mais l'orientation de la base est reproduite correctement (tableau 6). Cette observation est particulièrement intéressante pour le nucléotide A3 puisque nous sommes partis d'une orientation anti et que nous

avons réussi à retrouver l'orientation syn après minimisation. La gamme de valeurs RMSD des poses de plus basse énergie (tableau 7) s'étend de 0,49 Å à 1,82 Å, valeurs similaires à celles observées pour les poses de plus bas RMSD (tableau 6). L'ensemble de ces résultats est encourageant dans la perspective d'échantillonner correctement des poses de type natif et de reproduire le mode d'interaction des chaînes expérimentales avec une haute résolution.

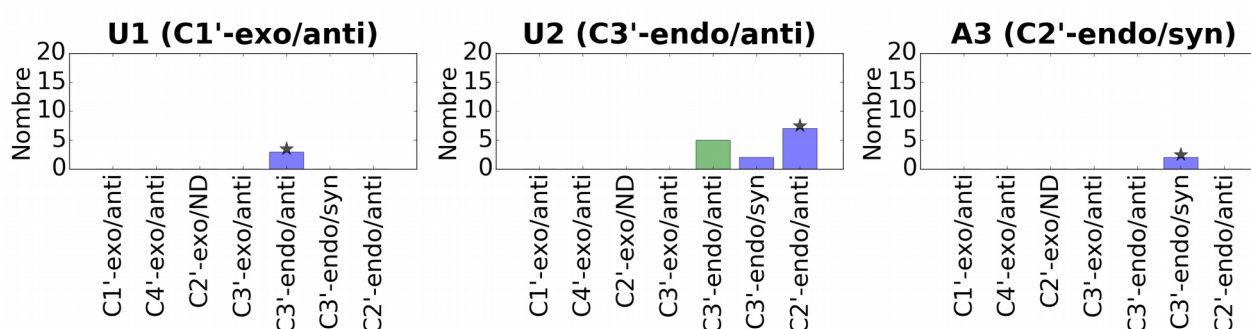
Globalement, l'étape d'échantillonnage du docking n'apparaît pas limitante sur les trois systèmes testés pour pouvoir reconstruire les chaînes d'ARN et reproduire leur mode d'interaction. Notons toutefois que ces poses natives sont en nombre variable pour chaque nucléotide et qu'elles ne représentent globalement qu'une faible proportion de l'ensemble des poses générées à l'issue du docking (entre 0,87 et 5,83 %, tableau 6).



## 2XNR



## 5ELH



## 5WWX

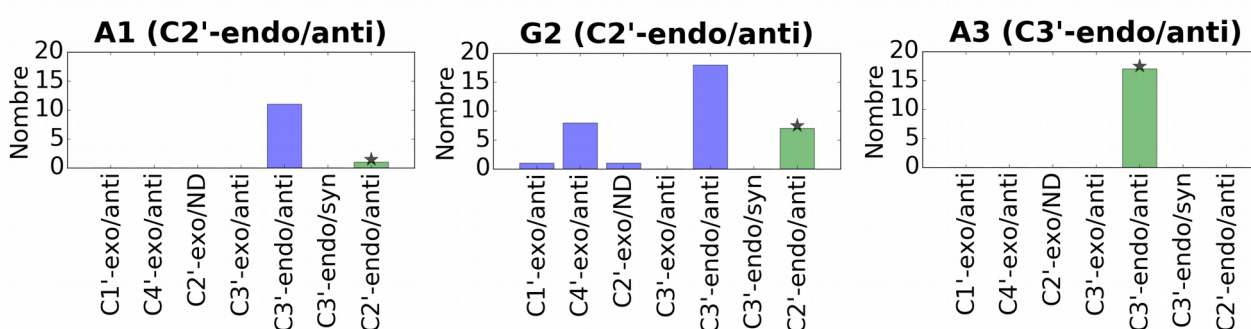


Figure 36: Recensement pour chaque complexe des conformations observées parmi l'ensemble des poses natives de chacun des nucléotides expérimentaux composant les chaînes d'ARN de référence. La conformation du nucléotide expérimental est mentionnée au-dessus de chaque graphique et est représentée par une barre de couleur verte. L'étoile au-dessus des barres indique la conformation de la pose native de plus basse énergie.

### 3.1.2 Evaluation de l'énergie d'interaction et du classement des poses

La deuxième étape essentielle du docking concerne l'estimation de l'énergie d'interaction des poses. Les poses natives doivent présenter une énergie d'interaction suffisamment favorable pour être distinguées de l'ensemble des poses générées. Le tableau 7 montre les caractéristiques des poses

de meilleure énergie d'interaction parmi l'ensemble des poses natives. Pour les nucléotides C2 et U3 de 2XNR, U2 et A3 de 5ELH, et G2 de 5WWX, ces poses natives sont classées parmi les 20 poses de meilleure énergie d'interaction sur l'ensemble des poses générées, ce qui correspond au moins au top 0,4 % de la totalité des poses. Les meilleures poses natives des nucléotides U1 de 2XNR et U1 de 5ELH possèdent une énergie d'interaction bien moins favorable par rapport aux poses natives des autres nucléotides (-12,64 kcal/mol et -11,44 kcal/mol, respectivement) et sont en conséquence beaucoup moins bien classées : tops 15 % et 46 %, respectivement. Cela signifie que pour pouvoir reproduire la chaîne ARN native de ces complexes, il serait nécessaire de sélectionner un nombre considérable de poses pour inclure ces poses : au minimum 759 U (et 759 C) pour 2XNR et 1946 U (et 1946 A) pour 5ELH. Il sera illustré dans la section suivante l'impossibilité d'utiliser une sélection aussi importante pour la recherche de chaînes. Les poses natives de plus basse énergie pour les nucléotides A1 et A3 de 5WWX sont classées dans des tops intermédiaires : tops 5,39 et 2,50 %. Cela pourrait également rendre leur sélection problématique pour la recherche de chaînes. Notons toutefois qu'elles présentent une énergie d'interaction bien plus favorable que celles observées pour les nucléotides U1 de 2XNR et U1 de 5ELH : -18,52 kcal/mol pour A1 et -19,72 kcal/mol pour A3. Ces énergies d'interaction sont comparables à celles observées pour les autres nucléotides du jeu de données classés dans le top 20 des poses meilleure énergie d'interaction.

*Tableau 7: Caractéristiques des poses natives de meilleure énergie d'interaction. Leur rang global parmi l'ensemble des poses générées est indiqué, ainsi que leur énergie d'interaction, leur RMSD par rapport à la conformation expérimentale et leur conformation.*

Code PDB	Nucléotide expérimental		Pose native ( $\leq 2 \text{ \AA}$ ) de meilleure énergie d'interaction			
	Position	Conformation	Energie (kcal/mol)	Rang global (%)	RMSD ( $\text{\AA}$ )	Conformation
2XNR	U1	C2'-endo/anti	-12,64	759 (14,80)	1,42	<b>C2'-endo/anti</b>
	C2	C2'-endo/anti	-20,63	1 (0,03)	1,29	C3'-endo/ <b>anti</b>
	U3	C2'-endo/anti	-17,39	17 (0,33)	1,06	<b>C2'-endo/anti</b>
5ELH	U1	C1'-exo/anti	-11,44	1946 (45,90)	1,82	C3'-endo/ <b>anti</b>
	U2	C3'-endo/anti	-21,19	4 (0,09)	0,81	C2'-endo/ <b>anti</b>
	A3	C2'-endo/syn	-25,58	1 (0,02)	1,65	C3'-endo/ <b>syn</b>
5WWX	A1	C2'-endo/anti	-18,52	409 (5,39)	1,53	<b>C2'-endo/anti</b>
	G2	C2'-endo/anti	-24,03	15 (0,22)	0,49	<b>C2'-endo/anti</b>
	A3	C3'-endo/anti	-19,72	190 (2,50)	1,04	<b>C3'-endo/anti</b>

Afin de comprendre pourquoi les poses natives de certains nucléotides présentent une énergie d'interaction peu favorable, le nombre de contacts a été recensé à partir des structures expérimentales entre chacun des nucléotides et leur domaine respectif. Un contact a été défini entre un atome du nucléotide et un atome du domaine si la distance qui les sépare est inférieure à 4 Å. La figure 37 montre que les poses natives les moins bien classées correspondent aux nucléotides ayant le plus faible nombre d'atomes en contacts avec le domaine. Cette observation suggère que l'énergie d'interaction peu favorable de ces poses natives n'est pas le résultat d'un manque de précision de la fonction de score utilisée mais plutôt le reflet du mode d'interaction des chaînes d'ARNsb. Les nucléotides composant les chaînes ARNsb peuvent en effet contribuer différemment à l'énergie d'interaction, comme l'avaient déjà remarqué les auteurs de l'approche basée sur ATTRACT pour des fragments 3-nt (Chauvot de Beauchene, de Vries, & Zacharias, 2016b; De Beauchene et al., 2016).

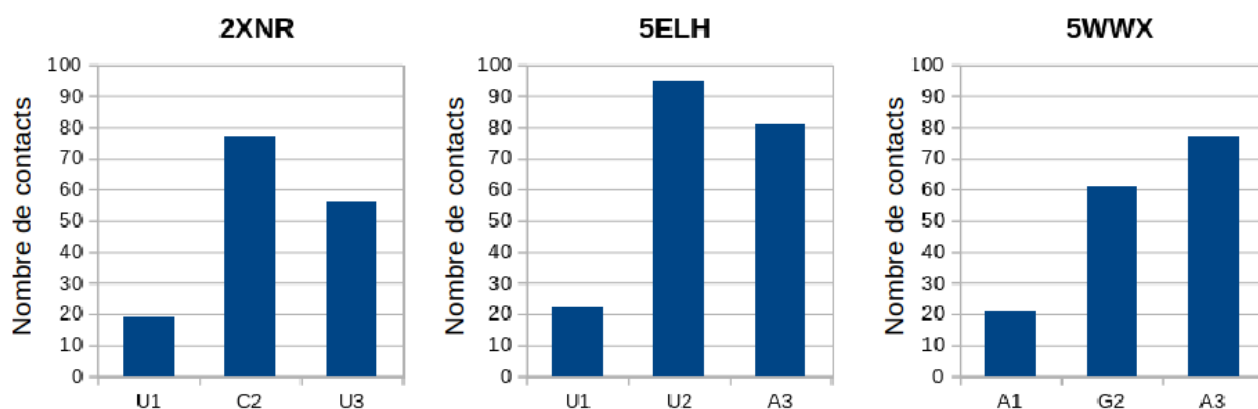


Figure 37: Recensement du nombre d'atomes en contacts entre chaque nucléotide cristallisé et le domaine partenaire. Un contact est défini entre deux atomes si leur distance est inférieure à 4 Å.

Les résultats présentés ci-dessus montrent que certains nucléotides des chaînes d'ARN contribuent plus fortement que d'autres à l'énergie d'interaction. Pour pouvoir reproduire le mode d'interaction expérimental des chaînes d'ARN 3-nt de notre jeu de données, il est nécessaire de sélectionner un nombre important de poses pour retenir les poses natives les moins bien classées correspondant aux nucléotides présentant une moindre contribution à l'énergie d'interaction. Dans la section qui suit, l'influence du nombre de poses sélectionnées sur l'espace de solutions des chaînes recherchées est illustrée. La mise en place d'une stratégie de sélection limitant le phénomène d'explosion combinatoire rencontrée est ensuite décrite.

## **3.2 Mise en place d'une stratégie de sélection**

### **3.2.1 Influence du nombre de poses sélectionnées sur l'espace de solutions des chaînes**

Comme mentionné précédemment, la capacité à reproduire le mode d'interaction expérimental des chaînes d'ARN 3-nt impose de sélectionner un nombre important de poses. Considérons par exemple le cas de 5WWX pour illustrer l'influence d'une sélection trop importante de poses sur le nombre de chaînes 3-nt identifiées. Pour pouvoir reconstruire la chaîne AGA expérimentale de 5WWX, il est nécessaire de retenir au minimum les 409 poses de meilleure énergie d'interaction pour retenir au moins une pose native propre à chaque nucléotide composant la chaîne expérimentale (tableau 7). Dans une procédure standard, la sélection s'applique à chaque type de nucléotides dockés : ce sont donc 409 poses A et 409 poses G qu'il faut sélectionner. La recherche de chaînes AGA par le programme Molpy à partir de cette sélection conduit à identifier 350 326 chaînes potentielles. Une fois ces chaînes déterminées, la suite du protocole consiste à sélectionner une ou des chaînes candidates sur la base de leur énergie d'interaction. Pour cela, une minimisation est requise au préalable pour corriger des distances inter-atomiques aberrantes pouvant séparer des nucléotides contigus ainsi que pour optimiser les contacts des chaînes avec la protéine (Fig. 32). La durée moyenne de minimisation d'une chaîne tri-nucléotidique peut varier de 5s à 8s environ par CPU. Plus de 583h seraient ainsi nécessaires pour minimiser les 350 326 chaînes sur un seul CPU. Ce temps de calcul peut être raisonnable pour une application ponctuelle de l'approche, mais il est prohibitif si cette dernière doit être appliquée à plus grande échelle. De plus, discriminer des chaînes natives sur la base d'une fonction de score relativement simple est d'autant plus difficile que le nombre de chaînes leurres est élevé. Le nombre de poses à sélectionner par type de nucléotides pour pouvoir reproduire les chaînes ARN expérimentales de 2XNR et 5ELH est encore plus important que pour 5WWX : 759 et 1946, respectivement. Il n'est donc pas envisageable pour l'approche FBDRNA de reproduire les chaînes 3-nt expérimentales à partir d'une sélection de poses basée sur l'énergie d'interaction des poses directement issues du docking.

Cette limitation est imposée par une énergie d'interaction peu favorable et donc un mauvais classement des poses natives associées aux nucléotides établissant moins de contacts avec les domaines. Ces nucléotides sont pour les trois systèmes du jeu de données localisés à l'extrémité 5'-terminale de la chaîne d'ARN (Fig. 37). On peut cependant remarquer que les nucléotides aux positions 2 et 3 de chaque chaîne établissent avec leur domaine respectif un nombre de contacts plus important (Fig. 37), et que leur pose native de plus basse énergie est au moins classée parmi les 200 poses de meilleure énergie d'interaction (tableau 7). Ces di-nucléotides (2-nt) peuvent jouer un rôle essentiel en agissant comme point d'ancrage permettant de positionner la chaîne d'ARN de manière

à rendre biologiquement fonctionnel le processus au sein duquel l'interaction intervient. Modéliser le mode d'interaction de telles ancres représente donc un intérêt certain. Pour évaluer l'influence d'une sélection de poses moins importante que celle nécessaire pour reproduire les chaînes 3-nt, les 200 poses de plus basse énergie par type de nucléotides dockés ont été données en entrée au programme Molpy. Pour chaque domaine, les chaînes 2-nt suivantes ont été recherchées à partir de cette sélection : chaînes CU pour 2XNR, chaînes UA pour 5ELH, et chaînes GA pour 5WWX.

La figure 38 résume la procédure de recherche et les résultats obtenus. Pour les trois domaines, près de 5000 chaînes 2-nt ont été identifiées et, pour chacun d'eux, la chaîne de meilleure énergie d'interaction correspond à une chaîne native reproduisant le mode d'interaction expérimental avec une résolution inférieure à 1,5 Å. Si ce résultat pouvait être attendu pour 2XNR et 5ELH au regard du classement de leurs poses natives à l'issue du docking (tableau 7), il peut sembler plus étonnant pour le domaine KH de 5WWX pour qui la pose native associée au nucléotide A3 est seulement classée au rang 190. Cela pourrait s'expliquer par une contribution synergique du di-nucléotide à l'énergie d'interaction qui serait rompue lorsque les nucléotides sont considérés de manière indépendante.

Pour résumer les résultats présentés dans cette section, l'approche FBDRNA peut reproduire le mode d'interaction de chaînes 2-nt avec une grande précision, mais elle présente une limitation pour reproduire des chaînes 3-nt. Cette limitation provient de nucléotides établissant peu de contacts avec les domaines ; leurs poses natives présentent une énergie d'interaction peu favorable qui impose de sélectionner un nombre trop important de poses pour permettre de reproduire les chaînes 3-nt natives. Bien que d'énergie moins favorable, ces nucléotides restent néanmoins importants puisqu'ils peuvent contribuer à l'augmentation de l'affinité pour le domaine partenaire et participer à augmenter la spécificité de reconnaissance. La section qui suit présente une première stratégie envisagée pour réduire l'espace de solutions obtenues pour la reproduction de chaînes 3-nt natives. Cette stratégie repose sur une étape de clustering des poses à l'issue du docking de manière à réduire leur nombre et potentiellement conduire à une sélection moins importante pour la reproduction de chaînes 3-nt natives.

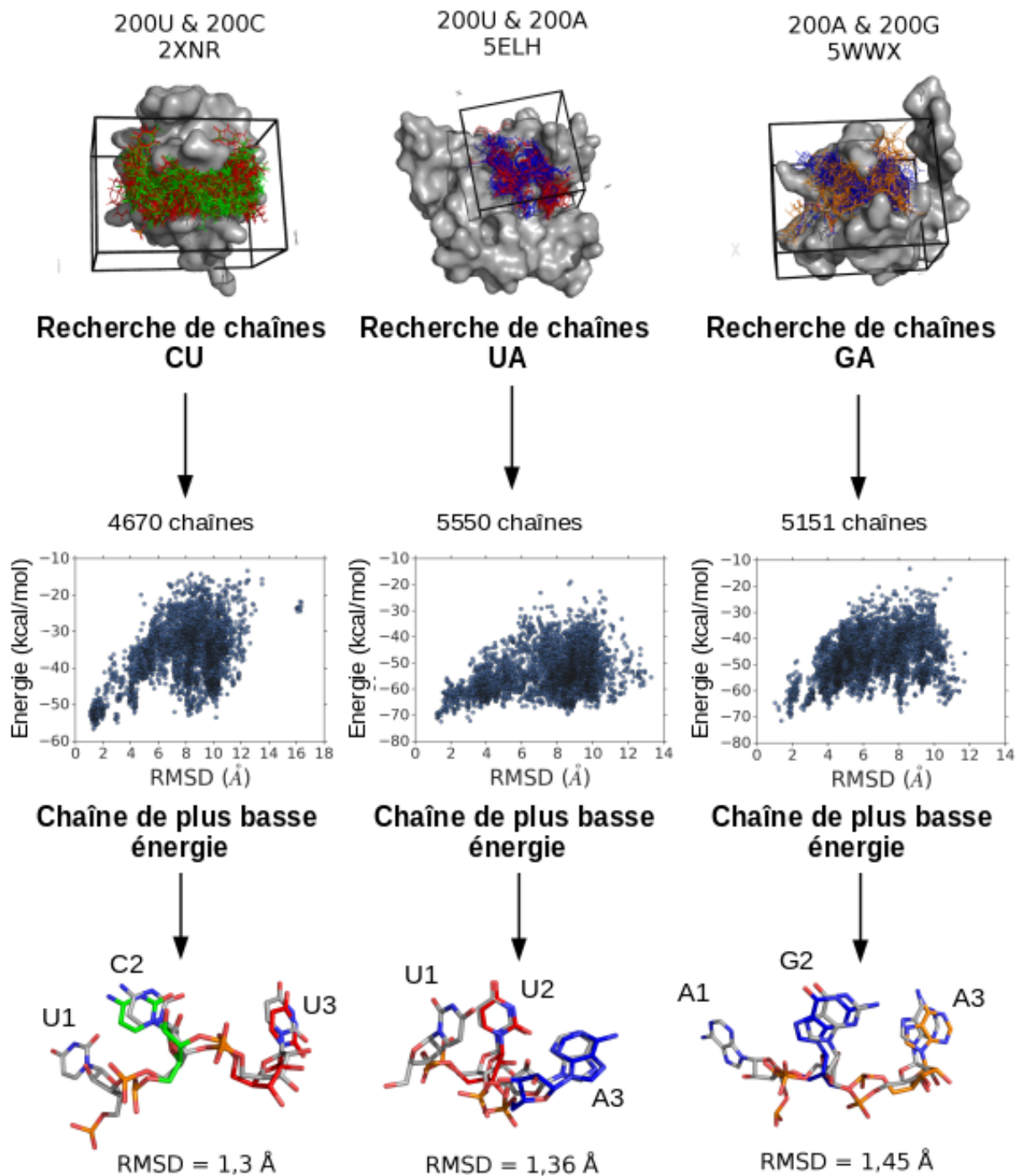


Figure 38: Recherche de chaînes di-nucléotidiques de séquence connue. Pour chaque domaine, les 200 poses de meilleure énergie d'interaction sont sélectionnées pour chaque type de nucléotides composant la séquence ARN recherchée. Les chaînes identifiées sont triées par ordre croissant de leur énergie d'interaction estimée après un protocole d'optimisation. La chaîne de meilleure énergie d'interaction correspond à une chaîne reproduisant correctement le mode d'interaction expérimental pour les trois systèmes testés. Pour chaque système, cette chaîne di-nucléotidique est représentée en bâtonnet et superposée à la chaîne tri-nucléotidique expérimentale. Les nucléotides expérimentaux sont représentés en bâtonnet de couleur grise, les autres sont de couleur rouge (U), bleue (A), verte (C) et orange (G). Le RMSD des chaînes di-nucléotidiques est calculé par rapport aux deux nucléotides expérimentaux correspondants : C2-U3 pour 2XNR, U2-A3 pour 5ELH et G2-A3 pour 5WWX.

### **3.2.2 Effet de la réduction du nombre de poses sur l'espace de solutions des chaînes**

#### ***De la redondance existe parmi les poses générées durant le docking***

L'énergie d'interaction est estimée à partir de la fonction de score MCSS qui repose sur le champ de force CHARMM27. Les fonctions de score basées sur un champ de force présentent l'inconvénient que de faibles variations de coordonnées atomiques peuvent conduire à des écarts importants d'énergie. En conséquence, le paysage énergétique d'interaction qui en découle est extrêmement rugueux. Au cours des différents cycles d'échantillonnage réalisés par MCSS, la redondance est évaluée : si deux poses convergent autour d'une même position, seule celle de meilleure énergie d'interaction est conservée. Le critère de convergence est défini par un seuil RMSD qui a été fixé à 0,5 Å pour tous les calculs de docking réalisés. Ce seuil de 0,5 Å permet une exploration fine du paysage énergétique des poses. En contrepartie, le nombre de poses générées à l'issue du docking demeure élevé puisque cette faible valeur de seuil entraîne nécessairement une forme de redondance entre les poses. Cette redondance affecte négativement l'approche FBDRNA à deux niveaux : le premier concerne le rang global des poses, et plus particulièrement celui des poses dont l'énergie est moins favorable qui verront leur rang augmenté par un simple effet d'empilement (présence à des rangs inférieurs de poses redondantes dont l'énergie d'interaction est plus favorable). Le second concerne les chaînes identifiées dont l'ensemble contiendra inévitablement également de la redondance, augmentant ainsi inutilement l'espace des solutions. Ainsi, pour évaluer dans quelle mesure la réduction du nombre total de poses permet de favoriser la sélection de poses natives d'intérêt et réduire la combinatoire lors de la recherche de chaînes 3-nt, un clustering a été effectué à l'issue du docking sur l'ensemble des poses générées.

#### **Influence de l'élimination de la redondance sur le rang global des poses natives**

La procédure de clustering utilisé (chapitre II, section 3) identifie dans un premier temps la pose ayant le plus de voisins selon un critère RMSD. Cette pose ainsi que toutes ses poses voisines sont regroupées dans un premier cluster. La procédure est répétée itérativement jusqu'à ce que l'ensemble des poses appartiennent à un groupe. Après ce clustering, les groupes formés d'une seule pose sont éliminés ; pour les autres, la pose de meilleure énergie d'interaction est sélectionnée comme pose représentative. L'ensemble des poses représentatives est alors ordonné par ordre croissant des énergies d'interaction, c'est-à-dire de la pose la plus favorable à la moins favorable. Un seuil RMSD de 2 Å a été fixé pour regrouper les poses sur la base du seuil RMSD utilisé pour définir des poses

natives. On suppose que ce seuil offre un bon compromis entre le nombre de clusters générés et la capacité à conserver des poses natives comme poses représentatives. En effet, un seuil RMSD élevé conduira à un faible nombre de clusters, mais le nombre de poses contenu dans chacun des clusters sera important ; en conséquence, dans un cluster contenant des poses natives, le risque qu'une pose native ne soit pas la pose de plus basse énergie du cluster sera plus élevé. Le raisonnement symétrique s'applique pour un seuil RMSD trop bas : le risque mentionné précédemment sera moindre mais en contrepartie le nombre de cluster, et donc le nombre de poses représentatives retenues, sera plus élevé. Un seuil RMSD trop bas n'est donc pas idéal dans la mesure où l'on cherche à réduire le nombre de poses.

Le tableau 8 montre que l'application de ce protocole de clustering permet de réduire d'un facteur 7 environ le nombre total de poses tout en retenant une pose native pour chaque nucléotide. Au regard du scoring, la comparaison du rang relatif (rang global/ nombre total de poses) des poses natives de plus basse énergie avant et après clustering indique que la réduction du nombre de poses n'apporte pas d'amélioration nette dans la discrimination des poses natives. Le clustering conduit toutefois, par un effet de réduction d'échelle, à améliorer le rang global des poses natives de plus basse énergie pour huit des neuf nucléotides expérimentaux. La pose native du nucléotide A1 de 5WWX est la seule à présenter un rang plus élevé après clustering (691 VS 409). Son énergie d'interaction est différente de celle de la pose native de meilleure énergie présente avant clustering, indiquant que cette dernière n'a pas été retenue comme pose représentative ; cela est une conséquence de la procédure de clustering et du seuil RMSD de 2 Å utilisés pour regrouper les poses. Le même constat s'applique également à la pose native du nucléotide U1 de 5ELH. Cependant, à la différence de la pose native de A1 de 5WWX, son rang global après clustering est inférieur à celui observé avant clustering.



Tableau 8: Comparaison des poses natives de meilleure énergie d'interaction obtenues avant et après clustering. Le clustering a été réalisé avec un seuil de 2 Å. La pose de meilleure énergie trouvée dans chaque groupe a été conservée comme pose représentative. Dans la colonne "rang global", le nombre entre parenthèses indique le rang relatif qui correspond au rang global rapporté au nombre total de poses.

Code PDB	Nucléotide expérimental	Sans clustering des poses			Après clustering des poses		
		Total de poses	Pose native de meilleure énergie d'interaction		Total de poses	Pose native de meilleure énergie d'interaction	
			Rang global (%)	Energie (kcal/mol)		Rang global (%)	Energie (kcal/mol)
2XNR	U1	5128	759 (14,80)	-12,64	720	155 (21,53)	-12,64
	C2	3947	1 (0,03)	-20,63	553	1 (0,18)	-20,63
	U3	5128	17 (0,33)	-17,39	720	9 (1,25)	-17,39
5ELH	U1	4239	1946 (45,90)	-11,44	544	490 (90,07)	-7,29
	U2	4239	4 (0,09)	-21,19	544	3 (0,55)	-21,19
	A3	4234	1 (0,02)	-25,58	559	1 (0,18)	-25,58
5WWX	A1	7587	409 (5,39)	-18,52	1168	691 (59,16)	-13,01
	G2	6897	15 (0,22)	-24,03	1034	7 (0,68)	-24,03
	A3	7587	190 (2,50)	-19,72	1168	58 (4,97)	-19,72

### ***Influence de l'élimination de la redondance sur l'espace de solutions des chaînes tri-nucléotidiques***

Pour 2XNR et 5ELH, le nombre minimal de poses à sélectionner par type de nucléotides dockés pour pouvoir reproduire leur chaîne 3-nt native est moins important après clustering : 155 poses par type de nucléotides pour 2XNR et 490 pour 5ELH. Il est en revanche plus important pour 5WWX pour qui 691 poses par type de nucléotides doivent être retenues. Afin d'évaluer l'influence du nombre de poses retenues après clustering sur l'espace de solutions des chaînes, les chaînes d'ARN 3-nt de chaque domaine ont été recherchées à partir des sélections minimales de poses nécessaires pour retrouver les chaînes natives (soit 155, 490 et 691 pour 2XNR, 5ELH et 5WWX respectivement). Les données utilisées en entrée et les résultats sont résumés dans le tableau 9. Sans surprise, au regard de la quantité de poses fournies en entrée et de l'illustration faite sur 5WWX à la section 3.2.1, le nombre de solutions identifiées pour 5ELH et 5WWX demeure trop important pour pouvoir être traité. Pour 2XNR en revanche, le nombre de chaînes identifiées est abordable au regard des temps de calculs. En considérant une durée moyenne de 6s par CPU pour minimiser une chaîne 3-nt, 23h sont nécessaires sur un CPU pour la minimisation des 13782 chaînes UCU identifiées. Parmi ces dernières, sept chaînes UCU (représentées par un cercle bleu à la figure 39-A)

reproduisent le mode d'interaction natif ( $\text{RMSD} \leq 2 \text{ \AA}$ ). Par ailleurs, la chaîne 3-nt de meilleure énergie d'interaction parmi l'ensemble des chaînes générées correspond à une chaîne native dont le RMSD est inférieur à  $1,5 \text{ \AA}$  (tableau 9 et Fig. 39-B) par rapport à la chaîne expérimentale UCU. Ce résultat est assez remarquable compte tenu du nombre élevé de chaînes concurrentes. Les énergies d'interaction des sept chaînes natives se distinguent assez nettement de la majorité des chaînes identifiées dont le RMSD s'étend sur une gamme de 5 à  $15 \text{ \AA}$  (Fig. 40-A). Notons également qu'une dizaine de chaînes présentant un RMSD compris entre 9 et  $13 \text{ \AA}$  ont une énergie d'interaction très proche de la chaîne native de plus basse énergie. Ces chaînes n'ont pas été analysées spécifiquement mais leur observation sur Pymol (non présentée ici) montre que certaines reproduisent le mode d'interaction natif, mais dans une orientation opposée à la chaîne expérimentale.

*Tableau 9: Données associées à la recherche de chaînes tri-nucléotidiques sur les trois domaines à partir d'une sélection de poses issues d'un clustering. Le nombre de poses données en entrée pour l'identification de chaînes est le nombre minimal permettant de reproduire le mode d'interaction natif de la séquence recherchée. ND indique que la procédure d'optimisation des chaînes identifiées n'a pas été réalisée.*

Code PDB	Nombre de poses sélectionnées				Séquence recherchée	Nombre de chaînes identifiées	Chaîne native ( $\leq 2 \text{ \AA}$ ) de meilleure énergie d'interaction	
	A	C	G	U			Rang global	RMSD ( $\text{\AA}$ )
2XNR	/	155	/	155	UCU	13 782	1	1,48
5ELH	490	/	/	490	UUA	1 584 619	ND	ND
5WWX	691	/	691	/	AGA	2 433 270	ND	ND

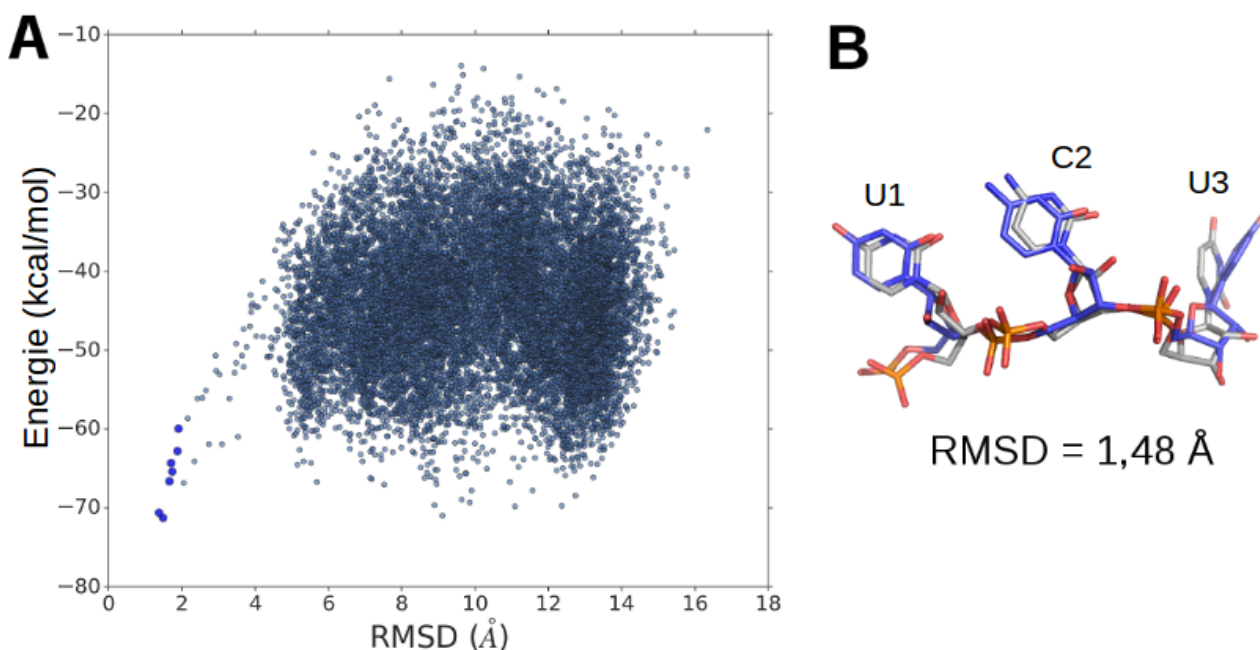


Figure 39: Chaînes UCU identifiées sur 2XNR à partir de 155 poses U et 155 poses C issues d'un clustering. (A) Projection du RMSD des chaînes identifiées en fonction de leur énergie d'interaction. Les cercles bleus représentent les sept chaînes UCU natives ( $\text{RMSD} \leq 2 \text{ \AA}$ ). (B) Superposition de la chaîne de plus basse énergie (en bleu) à la chaîne expérimentale (en gris).

### ***Influence de l'élimination de la redondance sur l'espace de solutions des chaînes di-nucléotidiques***

La prédiction du mode d'interaction de chaînes 2-nt s'est déjà révélée concluante avant clustering à partir d'une sélection des 200 poses de meilleure énergie d'interaction par nucléotide (Fig. 38). Après clustering, le tableau 8 montre qu'une sélection des 60 poses de plus basse énergie par type de nucléotides est suffisante pour inclure la totalité des poses natives nécessaires à la reconstruction des chaînes 2-nt expérimentales (formées des nucléotides aux positions 2 et 3 des chaînes 3-nt) pour les trois domaines. Leur recherche à partir de cette sélection conduit à identifier pour les trois domaines entre 350 et 500 solutions potentielles (tableau 10), soit une réduction de plus d'un facteur 10 par rapport aux chaînes identifiées avant clustering (Fig. 38). Tout comme pour la procédure réalisée sans clustering (Fig. 38), les chaînes de meilleure énergie d'interaction parmi l'ensemble des chaînes identifiées correspondent à une chaîne 2-nt native pour les trois domaines (tableau 10). Notons toutefois une variation notable entre le RMSD de la chaîne UA de 5ELH trouvée avant (1,36 Å, Fig. 39) et après clustering (1,89 Å, tableau 10 et Fig. 40). Cela s'explique par le fait que la pose native associée au site A3 est différente dans les deux chaînes prédites. Celle issue du clustering présente une rotation de 180° de sa base mais conserve malgré tout des contacts natifs au niveau de ses azotes, notamment le N6 (Fig. 40).

Tableau 10: Données associées à la recherche de chaînes di-nucléotidiques sur les trois domaines à partir d'une sélection des 60 poses de plus basse énergie issues d'un clustering.

Code PDB	Nombre de poses sélectionnées				Séquence recherchée	Nombre de chaînes identifiées	Chaîne native ( $\leq 2$ Å) de meilleure énergie d'interaction	
	A	C	G	U			Rang global	RMSD (Å)
2XNR	/	60	/	60	CU	352	1	1,48
5ELH	60	/	/	60	UA	470	1	1,89
5WWX	60	/	60	/	GA	466	1	1,25

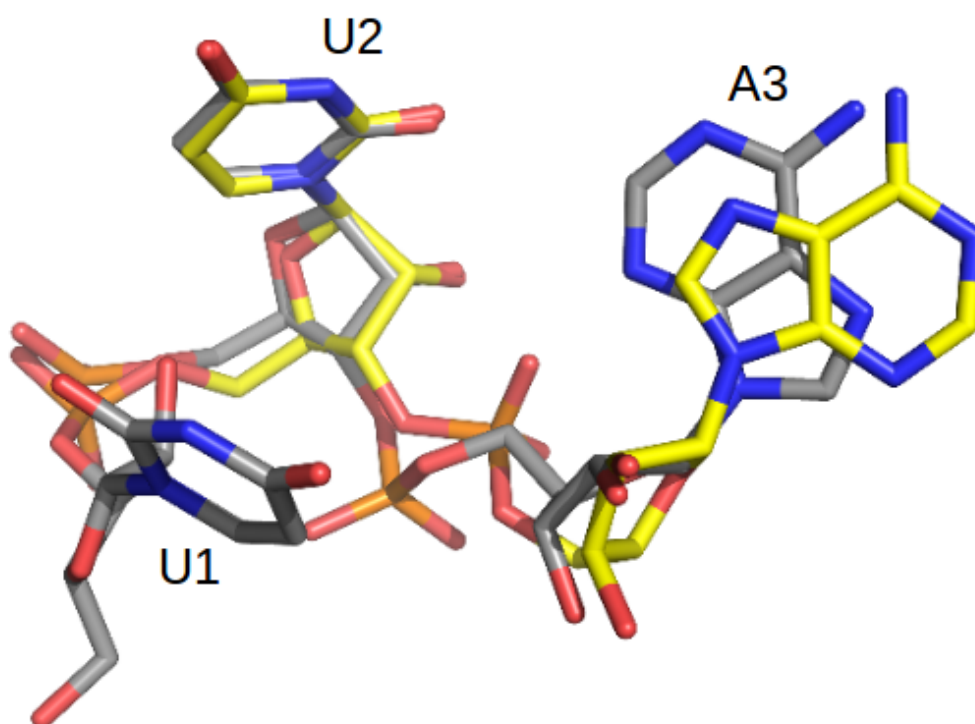


Figure 40: Superposition de la chaîne di-nucléotide UA de plus basse énergie trouvée pour 5ELH à la chaîne UUA expérimentale. La chaîne expérimentale UUA est représentée par des atomes de carbone en gris et la chaîne modélisée UA par des atomes de carbone jaune. Les atomes d'oxygène sont en rouge, d'azote en bleu et les atomes de phosphore en orange. L'adénine A3 de la chaîne modélisée présente sa base orientée en anti alors que la base du nucléotide A3 expérimentale est orientée en syn.

Globalement, les résultats présentés dans cette section montrent que la réduction du nombre de poses résultant d'un clustering apporte une amélioration de l'approche FBDRNA. Cette réduction

du nombre de poses a permis de réduire l'espace de solutions d'un facteur 10 pour la recherche de chaînes 2-nt tout en conservant la capacité à discriminer le mode d'interaction natif. La réduction du nombre de poses a également permis la prédiction à une haute résolution du mode d'interaction de la chaîne UCU du domaine RRM (2XNR) alors que cela n'était pas envisageable sans clustering. L'étape de clustering apparaît en revanche insuffisante pour la prédiction des chaînes 3-nt natives des domaines Zn-CCCH et KH (5ELH et 5WWX, respectivement) pour qui la sélection de la pose native du nucléotide en 5' reste problématique. La section suivante décrit la mise au point d'un protocole de sélection visant à retenir ces poses natives tout en évitant le phénomène d'explosion combinatoire lors de la recherche de chaînes 3-nt. Contrairement aux stratégies testées jusqu'ici dont la sélection des poses se base uniquement sur un critère énergétique, le protocole de sélection suivant inclut un critère supplémentaire : le nombre de poses trouvées autour de chaque site d'interaction nucléotidique.

### **3.2.3 Sélection de poses basées sur une approche "diviser pour mieux régner"**

Comme on l'a déjà vu, les chaînes d'ARN expérimentales de notre jeu de données sont constituées de nucléotides qui, en fonction de leur nombre de contacts avec le domaine partenaire, ne contribuent pas tous de la même façon à l'interaction. Les calculs de docking étant réalisés sur une région qui englobe chacun des sites d'interaction nucléotidiques, une compétition existe naturellement entre les poses retrouvées à chacun de ces sites. Cette compétition est particulièrement pénalisante pour les poses natives des nucléotides établissant moins de contacts avec le domaine puisque leur énergie d'interaction ne leur permet pas d'être bien classées par rapport aux autres nucléotides établissant plus de contacts avec le domaine partenaire. Pour contourner cette concurrence et faciliter la sélection de l'ensemble des poses d'intérêt, il conviendrait de considérer chacun des sites séparément en les isolant les uns des autres. La figure 41 illustre la stratégie mise en place pour y parvenir. Elle s'appuie sur l'hypothèse qu'à chaque site de liaison nucléotidique doit être associé un bassin énergétique (étape 1 – Fig. 41) à l'intérieur duquel on s'attend à retrouver des poses en plus grand nombre (Comeau, Gatchell, Vajda, & Camacho, 2004; Fernández-Recio, Totrov, & Abagyan, 2004; Kozakov, Clodfelter, Vajda, & Camacho, 2005)). L'idée est de tirer parti de cette propriété en sélectionnant dans un premier temps des groupes de poses les plus peuplés puis ensuite, sélectionner celles de plus basse énergie au sein de ces sous-ensembles. Pour cela, une étape de clustering est réalisée afin de regrouper les poses voisines dans l'espace selon un critère RMSD (étape 2, Fig. 41). Un seuil de clustering pertinent doit conduire à regrouper les poses natives propres à chaque site d'interaction nucléotidique dans des clusters différents. On s'attend à ce que ces derniers figurent parmi ceux les plus peuplés

(Comeau, Gatchell, Vajda, & Camacho, 2004; Fernández-Recio, Totrov, & Abagyan, 2004; Kozakov, Clodfelter, Vajda, & Camacho, 2005). L'étape 3 consiste donc à sélectionner les  $N_c$  clusters les plus peuplés (étape 3, Fig. 41). A ce stade, puisque les poses natives propres à chaque nucléotide se trouvent dans des groupes distincts, elles n'entrent plus directement en compétition au regard de leur énergie d'interaction. La sélection des  $N_p$  poses au sein de chaque cluster peut alors se faire selon leur énergie d'interaction (étape 4, Fig. 41). Elle doit permettre d'inclure des poses natives initialement présentes dans des bassins énergétiques moins favorables. L'ensemble  $N_f$  des poses finalement sélectionnées est alors égal au produit de  $N_c$  par  $N_p$  et peut être utilisé pour la recherche de chaînes (étape 5, Fig. 41) :

$$N_f = N_c \times N_p \quad (23)$$

L'objectif de cette stratégie est de réduire le nombre de poses à sélectionner tout en retenant les poses natives d'énergie moins favorable. Elle ne peut être efficace que si le nombre  $N_c$  de clusters à sélectionner et le nombre  $N_p$  de poses à retenir dans chacun de ces clusters sont peu élevés. Les valeurs  $N_c$  et  $N_p$  sont dépendantes du seuil RMSD de clustering utilisé. Un seuil RMSD trop grand conduira à peu de clusters, et donc à un faible nombre  $N_c$  de clusters à sélectionner. Ces clusters seront en revanche plus peuplés que ceux générés avec un seuil RMSD plus bas ; en conséquence, un seuil RMSD trop grand risquerait de fusionner des ensemble de solutions très différents et conduira à une valeur  $N_p$  élevée. En effet, les poses natives au sein des clusters d'intérêt seront plus difficiles à discriminer, particulièrement les poses natives dont l'énergie est moins favorable. A l'inverse, l'utilisation d'un seuil RMSD trop bas pour le clustering conduira à un nombre trop important de clusters. Si la valeur  $N_p$  diminuera puisque les clusters seront moins peuplés en poses, le nombre de clusters générés sera élevé et la valeur  $N_c$  des clusters à sélectionner risque d'être trop grande. Afin de définir un seuil de clustering offrant le meilleur compromis entre les valeurs  $N_c$  et  $N_p$ , et donc la plus faible valeur  $N_f$ , quatre seuils différents ont été testés sur les trois systèmes : 4 Å, 5 Å, 6 Å et 7 Å. Le seuil de clustering conduisant à la plus faible valeur  $N_f$  de poses à sélectionner et permettant de retenir l'ensemble des poses natives des nucléotides composant les trois chaînes ARN de référence est un seuil à 5 Å. Les résultats de ces analyses sont présentés dans la partie "Annexes", section 1. Dans les sections qui suivent, le protocole de sélection mis en place à partir d'un seuil de clustering à 5 Å est décrit avant d'en illustrer les résultats sur les trois systèmes de référence.

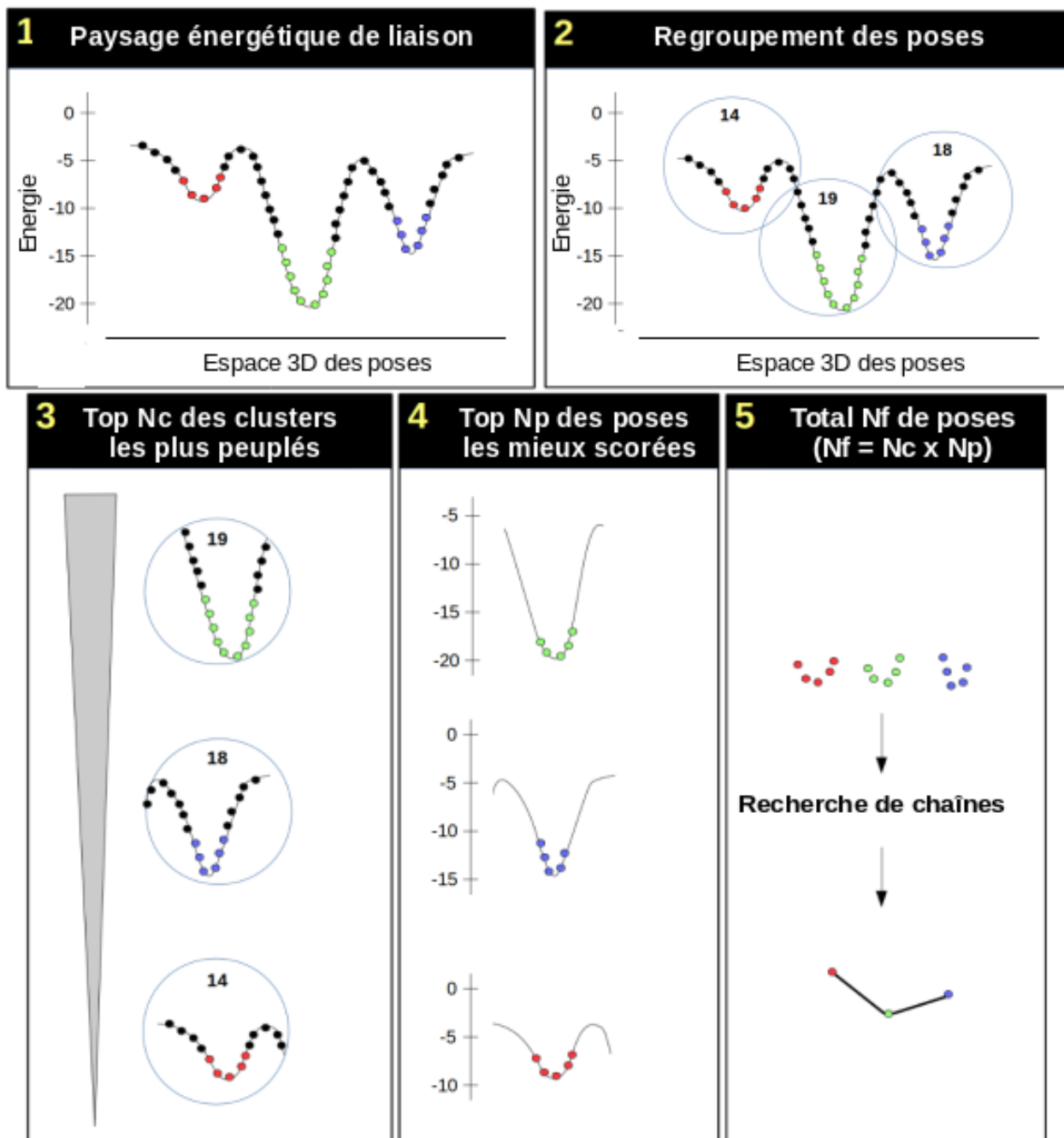


Figure 41: Illustration du principe de la stratégie de sélection de poses "diviser pour mieux régner". 1 - Représentation schématique d'un paysage énergétique de liaison. Les poses trouvées dans des bassins énergétiques différents sont représentées par des couleurs distinctes. 2 - Les poses voisines sont regroupées dans des clusters (cercles bleus). Le nombre de poses dans chaque cluster est indiqué. 3 - Les clusters sont ordonnés par ordre décroissant de leur taille et les Nc clusters les plus peuplés sont sélectionnés. 4 - Les Np poses de meilleure énergie d'interaction sont sélectionnées au sein de chaque cluster retenu. 5 - Les Nf poses finalement retenues peuvent être utilisées comme source d'entrée pour la recherche de chaînes.

## Protocole de sélection à partir d'un seuil de clustering à 5 Å

Pour chaque ensemble de poses issues des calculs de docking réalisés sur les trois systèmes, une matrice est calculée de manière à recenser les valeurs RMSD entre toutes les paires de poses. A partir de cette matrice, les poses sont regroupées entre elles d'après un seuil de clustering égal à 5 Å. Les clusters issus de ce regroupement sont ordonnés du cluster le plus peuplé au moins peuplé. Etant donnés les résultats présentés à la section 3.2.2 montrant les bénéfices de l'élimination de la redondance parmi les poses issues du docking sur le rang des poses natives, une deuxième étape de clustering est effectuée avec un seuil de 2 Å de manière à éliminer la redondance entre les poses trouvés dans chacun des clusters. Le tableau 11 recense, pour chaque nucléotide expérimental, les valeurs  $N_c$  et  $N_p$  minimales pour retenir dans la sélection leur pose native. La valeur  $N_c$  correspond au rang du cluster le plus peuplé contenant au moins une pose native ; la valeur  $N_p$  correspond au rang de la pose native de plus basse énergie trouvée dans le cluster de rang  $N_c$ . La valeur  $N_f$  indique enfin le nombre minimal de poses totales à sélectionner pour retenir au moins une pose native ; ce nombre est défini par le produit de  $N_c$  par  $N_p$ .

On peut voir dans le tableau 11 qu'avec ce protocole de sélection, seulement huit poses par type de nucléotides dockés doivent être sélectionnées pour retrouver une pose native pour chacun des nucléotides expérimentaux de 2XNR, et seulement 14 poses par type de nucléotides dockés pour 5WWX. A titre de comparaison, après une sélection basée uniquement sur l'énergie d'interaction, le nombre minimal de poses à sélectionner par type de nucléotides était de 155 pour 2XNR et 691 pour 5WWX lorsque la redondance entre poses était éliminée (tableau 8). Pour 5ELH, l'écart est moins impressionnant mais reste néanmoins important : 231 poses par type de nucléotides avec le protocole "diviser pour mieux régner" contre 490 après une simple élimination de la redondance (tableau 8).



Tableau 11: Détails des valeurs  $N_c$  et  $N_p$  obtenues à partir du protocole "diviser pour mieux régner" établi avec un seuil de clustering à 5 Å. Pour chaque nucléotide expérimental constituant les chaînes ARN du jeu de données, les valeurs  $N_c$ ,  $N_p$  et  $N_f$  sont indiquées. La valeur  $N_c$  correspond au rang du cluster le plus peuplé contenant au moins une pose native ; la valeur  $N_p$  correspond au rang de la pose native de plus basse énergie trouvée dans le cluster de rang  $N_c$ . La valeur  $N_f$  indique le nombre minimal de poses totales à sélectionner pour retenir au moins une pose native ;  $N_f$  est le produit de  $N_c$  par  $N_p$ .

Code PDB	Nucléotide expérimental	$N_c$	$N_p$	$N_f = N_c \times N_p$
2XNR	U1	5	1	5
	C2	1	1	1
	U3	8	1	8
5ELH	U1	11	21	231
	U2	6	1	6
	A3	10	1	10
5WWX	A1	9	1	9
	G2	2	1	2
	A3	2	7	14

### Recherche de chaînes 3-nt à partir d'une sélection minimale de poses permettant de reproduire les chaînes natives

La stratégie de sélection mise en place apporte donc un gain considérable dans la réduction du nombre de poses à retenir pour pouvoir reproduire les chaînes ARN 3-nt expérimentales. Concrètement, suivant le protocole défini, le tableau 11 montre qu'en retenant les 11 clusters les plus peuplés puis en sélectionnant pour chacun d'eux les 21 poses de plus basse énergie, les poses natives permettant de reproduire la chaînes 3-nt native peuvent être retenues pour les trois systèmes. Cette sélection conduit à retenir 231 poses par type de nucléotides dockés. Afin d'évaluer l'influence du nombre de poses sélectionnées sur l'espace de solutions des chaînes identifiées, une recherche de chaînes 3-nt a été réalisée pour les trois systèmes à partir de cette sélection. Les résultats montrés au tableau 12 indiquent qu'entre 107 000 et 166 000 chaînes sont identifiées. La minimisation nécessaire à leur optimisation et l'évaluation de leur énergie d'interaction demande environ 213 heures en moyenne par CPU et par système ( $\sim 6s$  / chaîne 3-nt / CPU). Des temps de calculs aussi importants ne sont pas appropriés pour une application à grande échelle, mais ils sont néanmoins envisageables pour une application ponctuelle. Chaque chaîne identifiée a donc été optimisée et leur énergie d'interaction évaluée. La chaîne de meilleure énergie d'interaction retrouvée pour 2XNR et 5WWX correspond à une chaîne native dont le RMSD est autour de 1,5 Å

(tableau 12). Ce résultat est remarquable au regard du nombre important de chaînes concurrentes. Il souligne le pouvoir de la fonction de score utilisée à discriminer le mode d'interaction natif des chaînes 3-nt. La chaîne native de plus basse énergie pour 5ELH est quant à elle retrouvée en 94ème position. Bien que ce rang soit moins bon que le rang des chaînes natives de plus basse énergie trouvées pour 2XNR et 5WWX, il reste néanmoins très satisfaisant puisque cette chaîne native figure dans le top 0,06 % de l'ensemble des 165 780 chaînes.

*Tableau 12: Résultats de la recherche de chaînes tri-nucléotidiques à partir d'une sélection de poses minimale permettant de reproduire les chaînes natives. Cette sélection inclut, pour chaque type de nucléotides dockés, les 11 clusters les plus peuplés (Nc), et pour chacun d'eux les 21 poses de plus basse énergie (Np).*

Code PDB	Paramètres de sélection des poses								Séquence recherchée	Nombre de chaînes identifiées	Chaîne native ( $\leq 2 \text{ \AA}$ ) de meilleure énergie d'interaction	
	A		C		G		U				Rang global	RMSD ( $\text{\AA}$ )
	Nc	Np	Nc	Np	Nc	Np	Nc	Np				
2XNR	/	/	11	21	/	/	11	21	UCU	110 983	1	1,48
5ELH	11	21	/	/	/	/	11	21	UUA	165 780	94	1,87
5WWX	11	21	/	/	11	21	/	/	AGA	107 558	1	1,56

### **Recherche de chaînes 3-nt à partir d'une proposition de sélection standard de poses pour une application à plus grande échelle**

Les résultats précédents ont été obtenus à partir d'une sélection minimale de poses pour reproduire les chaînes 3-nt natives. Malgré des prédictions très encourageantes, les chaînes identifiées par cette sélection sont si nombreuses qu'elles ne peuvent être traitées que dans le cadre d'une application ponctuelle. Afin d'envisager une application à plus grande échelle, une proposition de paramètres de sélection plus standard est ici testée. Le tableau 11 montre qu'en sélectionnant les 10 clusters les plus peuplés et dans chacun d'eux les 10 poses de plus basse énergie, des poses natives sont retenues pour huit des neuf nucléotides composant les chaînes ARN de notre jeu de données. Les résultats de la recherche de chaînes 3-nt effectuée à partir de cette sélection sont résumés au tableau 13. Le nombre de chaînes identifiées est réduit de plus d'un facteur 10 par rapport à la sélection précédente (tableau 12) : environ 7500 chaînes sont prédites pour 2XNR et 5WWX, et environ 12000 chaînes pour 5ELH. Cette quantité est abordable puisque entre 12 et 20 heures sont nécessaires pour optimiser ces chaînes sur un seul CPU. Parmi l'ensemble des chaînes générées, la chaîne de plus basse énergie reproduit le mode d'interaction natif de la chaîne

3-nt de 2XNR et 5WWX avec une précision de 1,5 Å. Pour 5ELH, la chaîne de plus basse énergie présente un RMSD de 5,48 Å par rapport à la chaîne 3-nt de référence UUA. En revanche, lorsque le RMSD de cette chaîne est calculé en ne considérant que les nucléotides U2 et A3, le RMSD est de 2 Å. Cela indique que le mode d'interaction natif est correctement reproduit pour ces deux nucléotides, mais pas pour le nucléotide U1. Cette observation est attendue pour le nucléotide U1 puisque les paramètres utilisés pour la sélection des poses ( $N_c = 10$  et  $N_p = 10$ ) ne permettent pas de retenir des poses natives pour ce nucléotide (tableau 11).

*Tableau 13: Résultats de la recherche de chaînes tri-nucléotidiques à partir d'un exemple de sélection standard de poses. Cette sélection inclut, pour chaque type de nucléotides dockés, les 10 clusters les plus peuplés ( $N_c$ ), et pour chacun d'eux les 10 poses de plus basse énergie ( $N_p$ ). L'astérisque pour 5ELH indique la valeur RMSD calculée en ne considérant que les nucléotides U2 et A3 de la chaîne UUA expérimentale.*

Code PDB	Paramètres de sélection des poses								Séquence recherchée	Nombre de chaînes identifiées	Chaîne de meilleure énergie d'interaction	
	A		C		G		U				Rang global	RMSD (Å)
	$N_c$	$N_p$	$N_c$	$N_p$	$N_c$	$N_p$	$N_c$	$N_p$				
2XNR	/		10	10	/		10	10	UCU	7 528	1	1,48
5ELH	10	10	/		/		10	10	UUA	12 135	1	5,48 (*2,00)
5WWX	10	10	/		10	10	/		AGA	7 418	1	1,56

Une sélection des 10 clusters les plus peuplés puis, dans chacun d'eux, des 10 poses de plus basse énergie conduit donc à un compromis intéressant entre temps de calculs et qualité des prédictions. Ces valeurs seront utilisées comme paramètres de sélection par défaut pour la suite des travaux présentés au chapitre IV.

## 4 Discussion

Prédire la conformation d'un ARNsb en interaction avec une protéine, lorsque cet ARN ne présente pas d'éléments de structures secondaires dans sa forme liée, est particulièrement difficile. L'échantillonnage de l'espace conformationnel est en effet délicat à traiter lorsque le nombre de degrés de liberté à considérer est important. Dans ce chapitre, nous avons établi la preuve de concept d'une approche de docking par fragments (FBDRNA) mono-nucléotidiques (1-nt) permettant de traiter la flexibilité de courtes chaînes ARNsb composées de trois nucléotides (3-nt). A partir de la structure d'un RBD, de la connaissance de son site d'interaction et de la séquence d'un ARN, la méthode a pour objectif de prédire le mode d'interaction natif de l'ARN avec le RBD.

L'approche a été développée et appliquée à partir de la forme liée de trois RBDs, chacun représentant une des familles les plus représentées chez l'Homme : les domaines RRM (code PDB : 2XNR), KH (code PDB : 5WWX) et Zn-CCCH (code PDB : 5ELH). Pour les domaines RRM et KH, l'approche FBDRNA a permis de prédire correctement le mode d'interaction natif de leur chaîne ARN 3-nt avec une très grande précision : parmi près de 7500 modèles générés, la chaîne 3-nt de plus basse énergie reproduit la conformation native avec une déviation RMSD avoisinant les 1,5 Å par rapport à la conformation expérimentale (tableau 13). Pour le domaine Zn-CCCH, parmi plus de 12 000 modèles identifiés, la chaîne 3-nt de plus basse énergie reproduit le mode d'interaction natif pour les deux nucléotides en 3' de la chaînes 3-nt expérimentale avec une déviation RMSD de 2 Å (tableau 13). Nous pensons que les difficultés rencontrées pour reproduire le mode d'interaction natif du nucléotide situé en 5' de la chaîne sont imputables à des contacts cristallins ne pouvant plus être établis à partir de la structure utilisée pour le docking.

L'approche FBDRNA parcourt l'espace conformationnel d'une séquence ARN en restreignant l'exploration à l'échelle d'un nucléotide. La première étape consiste à docker indépendamment chacun des nucléotides composant la séquence ARN donnée en entrée. Les calculs de docking, effectués par MCSS, reposent initialement sur une unique structure d'un nucléotide. Durant la procédure d'échantillonnage implémentée dans MCSS, l'exploration conformationnelle du nucléotide est permise en considérant ce dernier comme entièrement flexible pendant une phase d'optimisation. L'analyse des poses générées sur les trois RBDs a en effet montré que les barrières énergétiques séparant les différents angles de torsion peuvent être franchies et ainsi conduire à des conformations de nucléotides différentes de la conformation de départ. La conformation du nucléotide qui a été docké sur les trois RBDs est une conformation utilisée par défaut présentant un plissement du ribose en C3'-endo et une base orientée en anti. Cette conformation est différente de la conformation de certains des nucléotides expérimentaux composant les chaînes 3-nt utilisées comme référence. Pour sept des neuf nucléotides expérimentaux, la conformation native exacte (plissement du ribose et orientation de la base) a pu être échantillonnée avec succès. Par ailleurs, pour cinq nucléotides, cette conformation native exacte est aussi celle de plus basse énergie parmi l'ensemble des poses natives générées. Pour les quatre autres nucléotides, seul le plissement du ribose diffère entre la conformation expérimentale et celle de la pose native de plus basse énergie ; l'orientation de la base est quant à elle correcte dans tous les cas. Plusieurs facteurs pourraient expliquer les échecs observés concernant le plissement du ribose, aussi bien dans son échantillonnage que dans l'identification de sa conformation native sur la base de l'énergie d'interaction. Tout d'abord, la conformation expérimentale du ribose des nucléotides peut résulter

non seulement de son interaction avec la protéine, mais aussi des contraintes géométriques imposées par les nucléotides voisins dans la chaîne ARN. Ces contraintes sont absentes lorsque le nucléotide est considéré individuellement comme c'est le cas lors du docking de fragments 1-nt. Par ailleurs, le docking est effectué à partir de structures de RBDs pour lesquelles tous les hétéroatomes ont été retirés. Ces derniers incluent des molécules d'eau et des ions (*e.g.* Mg<sup>2+</sup>) qui peuvent aussi influencer la conformation du ribose. Cette représentation simplifiée du système peut donc également influencer la capacité à reproduire la conformation native du ribose (et *a fortiori* celle de l'ensemble du nucléotide). Dans ces conditions, s'attaquer à identifier précisément le plissement natif du ribose représente un challenge difficile à atteindre. Précisons cependant que l'approche FBDRNA n'affiche pas pour ambition d'atteindre un tel niveau de résolution. Nous considérons donc que les quelques échecs observés pour la prédiction de la conformation du ribose ne sont pas pénalisants. La capacité à prédire correctement l'orientation de la base est en revanche non négligeable puisque dans les interactions protéine-ARNs où l'ARN ne présente pas d'éléments structuraux dans sa forme liée, les contacts établis entre la base et la protéine sont directement impliqués dans la reconnaissance sélective de l'ARN. A cet égard, il est très encourageant d'avoir retrouvé l'orientation *syn* de la base du nucléotide A3 de 5ELH dans la pose native de plus basse énergie à partir de la conformation du nucléotide docké dont la base est orientée en *anti*.

L'approche basée sur ATTRACT et la méthode Rosetta *RNP-denovo* reposent toutes les deux sur une bibliothèque de conformères 3-nt tirés de structures expérimentales pour échantillonner l'espace conformationnel de l'ARN. Comparée à ces méthodes, l'approche FBDRNA, par l'utilisation d'un fragment 1-nt traité d'une manière flexible durant le docking, permet d'une part une exploration de l'espace conformationnel plus large et résolutive, et d'autre part, offre le potentiel de prédire des conformations de chaînes 3-nt encore jamais observées expérimentalement. L'utilisation d'un fragment 1-nt entraîne néanmoins une difficulté de taille liée aux contacts établis par les nucléotides natifs des chaînes expérimentales : certains nucléotides établissent peu de contacts avec la protéine et présentent en conséquence une énergie d'interaction peu favorable. La sélection des poses natives reproduisant le mode d'interaction de ces nucléotides sur la base de leur énergie impose de retenir un nombre important de poses qui conduit à une explosion combinatoire pénalisante pour la recherche de chaînes. Cette limitation a déjà été rencontrée et mise en évidence dans les travaux présentant l'approche basée sur ATTRACT (De Beauchene et al., 2016). Notons toutefois que pour cette dernière, l'utilisation de fragments 3-nt offre un avantage conséquent par rapport à l'utilisation d'un fragment 1-nt : dans le cas d'un fragment 3-nt, l'énergie d'interaction d'un nucléotide établissant peu de contacts avec la protéine peut être compensée par une énergie

plus favorable des autres nucléotides du même fragment si ces derniers établissent des contacts suffisamment importants. Pour faire face aux difficultés associées au nombre de contacts établis entre la protéine et certains nucléotides, nous avons développé pour l'approche FBDRNA un protocole de sélection basée sur une approche de clustering permettant de réduire drastiquement le nombre total de poses sélectionnées tout en retenant les poses natives d'énergie moins favorables. L'application de ce protocole avec des paramètres de sélection proposés comme potentiellement standard s'est montrée efficace sur les trois RBDs en générant un nombre de solutions raisonnables (~ 10 000 chaînes 3-nt) parmi lesquelles, pour deux RBDs, la chaîne de plus basse énergie reproduit avec une grande précision la chaîne 3-nt native. Les paramètres de sélection utilisés (seuils de clustering, nombre de clusters retenus, nombre de poses représentatives retenues) demandent néanmoins à être testés sur un plus grand jeu de données pour s'assurer qu'ils soient transférables à d'autres systèmes.

Les approches existantes pour la modélisation des interactions protéine-ARNsb font toutes appel à une réduction de la complexité combinatoire pour la recherche de chaînes : la méthode RNA-LIM (chapitre IV de la partie "Introduction", section 3.1) restreint cette recherche autour de résidus présentant une certaine probabilité à interagir avec des nucléotides, tandis que la méthode Rosetta *RNP-denovo* (chapitre IV de la partie "Introduction", section 3.3) et l'approche basée sur ATTRACT (chapitre IV de la partie "Introduction", section 3.2) reposent respectivement sur la connaissance exacte ou la prédiction de quelques nucléotides d'ancrage utilisés comme contraintes pour modéliser des chaînes en interaction avec la protéine. Dans l'approche FBDRNA, l'exploration de l'espace conformationnel de l'ARN est restreinte au site d'interaction des RBDs. L'ensemble des résultats présentés dans ce chapitre a été obtenu dans des conditions où le site d'interaction des RBDs a été considéré comme parfaitement caractérisé. En pratique, sans connaissance précise de ce dernier et en l'absence de données expérimentales (RMN, alanine scanning...), des programmes de prédiction peuvent être utilisés pour définir des résidus de la protéine pouvant interagir avec l'ARN. La comparaison de plusieurs de ces programmes sur différents jeux de données protéine-ARN a montré que la plupart offre d'excellentes performances prédictives (Miao & Westhof, 2015). Par ailleurs, la méthode RNA-LIM a déjà illustré avec succès sur un domaine RRM comment l'utilisation d'un programme de prédiction d'interaction nucléotide-acide aminé peut être appliqué pour restreindre l'espace de recherche de chaînes ARN. Cela suggère que les résultats obtenus par l'approche FBDRNA pourraient être transférables dans le cas où le site d'interaction n'est pas connu à l'avance, mais des tests doivent néanmoins être effectués pour s'en assurer.

Un autre élément contribuant à réduire la complexité combinatoire dans l'approche FBDRNA réside dans la restriction de son application à des RBDs. Les RBDs sont les modules de liaison primaires des RBPs capables de reconnaître de courtes chaînes ARN, généralement de manière spécifique. En restreignant l'exploration conformationnelle au niveau du site d'interaction de RBDs, l'espace de recherche des chaînes est moins important qu'il ne le serait si l'exploration était réalisée au niveau du site d'interaction de RBPs. Le choix de se restreindre à des RBDs n'a toutefois pas uniquement été orienté par le souci de réduire la complexité combinatoire ; ce choix provient aussi de considérations pratiques dans l'optique d'une application de l'approche FBDRNA dans le cas où la structure du récepteur n'aurait pas encore été résolue expérimentalement. Dans ces conditions, la stratégie la plus commode consiste à obtenir un modèle de la structure par homologie de séquences. Une centaine de structures est accessible dans la PDB pour les domaines RRM et KH ; ces structures montrent un repliement tertiaire qui est très conservé, même entre des séquences homologues pouvant présenter moins de 30 % d'identité. Ces caractéristiques suggèrent que la structure 3D de ces domaines peut être prédite par homologie avec une précision globale potentiellement raisonnable pour une application de docking. L'obtention de tels modèles pour des RBPs apparaît en revanche beaucoup plus délicat. Tout d'abord, beaucoup de RBPs présentent une architecture composée de plusieurs RBDs arrangés en tandem. La majorité des structures de RBPs ayant été résolue expérimentalement concernent des RBPs constituées tout au plus de deux RBDs. Cela limite donc la possibilité de prédire par homologie la structure de RBPs composée d'au moins trois domaines. Ensuite, l'arrangement inter-domaine, responsable de l'architecture globale des RBPs et pouvant jouer un rôle important dans la fixation de l'ARN, est particulièrement difficile à déterminer en raison du caractère variable des résidus connectant les différents domaines. D'une manière générale, indépendamment du type de protéines, la prédiction de l'assemblage de protéines composées de plusieurs domaines est un challenge non résolu et fait l'objet d'un travail constant au sein de la communauté des bio-informaticiens structuraux (Lensink et al., 2018). Développer une méthode générale permettant de modéliser les interactions entre ARNs et RBPs constitue donc un obstacle extrêmement complexe à surmonter et sa réalisation à moyen terme ne pourra probablement s'envisager que par la combinaison de plusieurs approches et au prix de simplifications dans la représentation des modèles. Le développement de FBDRNA s'est donc restreint à des RBDs car nous pensons que son application est plus accessible dans le cas où la structure du RBD d'intérêt n'aurait pas encore été résolue expérimentalement.

Une conséquence inhérente au fait de restreindre les prédictions à des RBDs est que la longueur des chaînes ARN à prédire est courte. Cet aspect contribue également à réduire la complexité

combinatoire pour la recherche de chaînes. L'approche FBDRNA a ici été développée et testée pour prédire des chaînes 3-nt. Cette longueur est typique de la longueur des motifs ARN généralement reconnus par les RBDs les plus représentés (Dominguez et al., 2018). Toutefois, certains de ces RBDs peuvent présenter une interface élargie permettant de lier des chaînes plus grandes (Cléry & Allain, 2013). On s'attend logiquement à ce que la combinatoire augmente pour la prédiction de telles chaînes par l'approche FBDRNA. Des tests additionnels sont nécessaires pour évaluer à partir de quelle longueur de chaînes cette augmentation devient limitante pour l'approche. Pour contourner les potentielles difficultés rencontrées, une stratégie de prédiction en deux étapes pourrait être envisagée. Dans cette stratégie de type *fragment-growing*, une ou quelques courtes chaînes d'ancrage seraient prédites dans une première étape de recherche de chaînes ; ces chaînes seraient ensuite étendues dans une deuxième étape de reconstruction de chaînes en utilisant les chaînes d'ancrage prédites comme contraintes pour limiter l'espace de solutions des chaînes identifiées.

Des programmes de docking comme AutoDock Vina (Trott & Olson, 2010) traitent la flexibilité du ligand directement à la volée au cours de l'échantillonnage. L'application d'AutoDock Vina sur la forme liée de structures prélevées de complexes protéine-peptide a montré des performances variables dans la capacité à reproduire le mode d'interaction de peptides présentant de 10 à 20 degrés de liberté (Rentzsch & Renard, 2015). Les prédictions se sont montrées globalement meilleures pour de courts peptides allant de trois à quatre résidus, et dépendantes des paramètres utilisés pour définir la profondeur d'échantillonnage. Une chaîne 3-nt comporte 21 degrés de liberté (sept par nucléotide), nombre qui pourrait potentiellement être abordable par AutoDock Vina. En conséquence, il est indispensable par la suite de valider l'approche FBDRNA en comparant ses performances à celles d'AutoDock Vina.

L'approche FBDRNA a été développée et appliquée sur la forme liée de trois RBDs. La forme liée de structures est souvent utilisée en première approximation dans le développement d'approches de docking. Dans ces conditions, le caractère transférable des résultats observés demande à être évalué, généralement à partir de la forme non-liée de la structure si elle est disponible. Cette évaluation est nécessaire en raison des changements conformationnels pouvant exister entre les formes liée et non-liée. Ces changements peuvent avoir une influence négative sur les résultats de docking, par exemple en obstruant l'accès d'un ligand à son site d'interaction. Des travaux réalisés sur douze RBPs ont montré que les résidus participant à l'interaction dévient en moyenne de 2 Å entre les formes liée et non-liée des protéines comparées (J. E. and S. J. Jones, 2008). Par ailleurs, l'approche basée sur ATTRACT a été capable de reproduire des modèles à haute



résolution à partir de la forme non-liée de deux RBPs à domaines RRM (l'orientation des domaines a été considérée connue). Pour ces deux protéines, la déviation RMSD des résidus est inférieure à 2 Å entre les formes liée et non-liée (De Beauchene et al., 2016). Ces résultats suggèrent que seuls de légers changements conformationnels se produisent suite à la liaison de l'ARN à des RBPs et que ces changements peuvent être tolérés dans une approche de docking. Il est néanmoins important de préciser d'une part que ces observations sont issues d'un faible nombre de RBPs et d'autre part que l'énergie d'interaction estimée par ATTRACT repose sur un modèle statistique gros-grain qui peut expliquer la tolérance observée face à de faibles changements conformationnels. La fonction de score utilisée dans l'approche FBDRNA repose quant à elle sur un modèle physique basé sur un champ de force et les molécules sont représentées en tout-atome. En conséquence, le paysage énergétique de liaison est particulièrement rugueux et sensible à de faibles changements de coordonnées. Il est donc indispensable d'évaluer l'approche FBDRNA sur la forme non-liée de RBDs pour s'assurer du caractère transférable des résultats observés sur la forme liée des trois RBDs du jeu de données.

Dans l'ensemble, l'approche FBDRNA a montré d'excellentes capacités prédictives du mode d'interaction de chaînes 3-nt à partir de leur séquence, de la forme liée de trois RBDs et de la connaissance du site d'interaction. Ces résultats encourageants ouvrent la voie vers une utilisation de l'approche FBDRNA pour prédire le mode d'interaction de courts motifs ARN identifiés par des approches expérimentales à haut-débit (*e.g.* CLIP-seq, RNAcontext, ...). Des validations sont cependant de l'approche sont néanmoins nécessaires, notamment pour s'assurer du caractère transférable de ses performances sur la forme non-liée de RBDs. La prédiction du mode d'interaction de courtes chaînes ARNs par l'approche FBDRNA pourrait aussi bénéficier à l'approche Rosetta *RNP-denovo* et l'approche basée sur ATTRACT qui reposent toutes les deux sur la connaissance/prédiction de courtes chaînes d'ancrage pour contraindre la recherche de chaînes ARN plus longues.

## IV Développement de l'approche FBDRNA sans *a priori* sur la séquence ARN

### 1 Introduction

Les interactions protéine-ARN interviennent dans une variété de processus cellulaires. Grand nombre de ces interactions passent par la reconnaissance spécifique de courtes séquences ARN par des RBDs composant les protéines de liaison à l'ARN. Déchiffrer le code de reconnaissance protéine-ARN représente un intérêt majeur pour la compréhension des implications fonctionnelles de ces interactions et permettrait d'ouvrir la voie au *design* de protéines pour aller cibler spécifiquement des partenaires d'intérêt. De nombreuses méthodes computationnelles ont été développées pour prédire des séquences ARN spécifiquement reconnues par des RBPs. La plupart de ces approches reposent sur des méthodes d'apprentissage profond à partir de données de séquençage à haut-débit (Alipanahi et al., 2015; X. Li, Kazan, Lipshitz, & Morris, 2014; X. Pan, Rijnbeek, Yan, & Shen, 2018). Malgré des résultats intéressants dans la prédiction de séquence ARN préférentiellement reconnue, aucune de ces méthodes ne permet de renseigner sur le mode d'interaction des séquences identifiées. Ces informations sont pourtant essentielles pour rationaliser les observations et aider au déchiffrement d'un code de reconnaissance. Les travaux présentés dans ce chapitre ont été réalisés avec l'objectif de pouvoir prédire correctement la séquence ARN préférentiellement reconnue par un RBD en même temps que son mode d'interaction. Pour ce faire, l'approche FBDRNA présentée au chapitre précédent a été adaptée pour répondre au problème de la recherche de chaînes sans connaissance *a priori* de la séquence ARN.

Les travaux présentés ci-dessous décrivent la mise en place d'une stratégie pour adapter l'approche FBDRNA à la prédiction d'une séquence ARN préférentiellement reconnue par un RBD simultanément à leur mode d'interaction. Pour répondre à cette problématique, l'approche développée suit le même principe générale que l'approche présentée dans le chapitre III (Fig. 30). Une différence importante nécessite cependant d'être précisée : puisque aucune séquence n'est ici donnée en *a priori* et donc imposée comme contrainte pour la recherche de chaînes, les quatre types de nucléotides doivent être dockés. L'augmentation de la combinatoire qui en résulte est d'abord illustrée dans les résultats qui suivent. Les travaux montrent ensuite la mise en place d'une stratégie pour réduire la complexité de la recherche et les résultats de son application sur la réduction de l'espace de solutions des chaînes identifiées sans connaissance *a priori* d'une séquence ARN. La capacité de l'approche à prédire une séquence préférentiellement reconnue par un RBD donné en même temps que leur mode d'interaction est finalement évaluée.

## 2 Matériels et méthodes

### 2.1 Jeu de données

Les trois complexes RBD-ARN utilisés et présentés dans le chapitre précédent ont été utilisés comme systèmes test. Pour chaque domaine du jeu de données, les séquences suivantes ont été prises comme référence pour définir la séquence préférentiellement reconnue :

- UCU pour le domaine RRM (code PDB : 2XNR)
- NUA pour le domaine Zn-CCCH (code PDB : 5ELH)
- AGA pour le domaine KH (code PDB : 5WWX)

Ces séquences sont celles observées dans la structure cristallographique des trois complexes et sont aussi celles présentant la meilleure affinité de liaison par rapport aux différents variants de séquences disponibles (chapitre III, section 2.1). Notons que pour 5ELH, le premier nucléotide de la chaîne tri-nucléotidique n'a pas été considéré en raison des doutes sur la pertinence biologique de son mode d'interaction avec le domaine Zn-CCCH, doutes soulevés *a posteriori* des résultats du chapitre III suite à l'observation de contacts cristallins pour ce nucléotide. Par ailleurs, en raison de contraintes de temps, les analyses n'ont pas considéré l'affinité de liaison disponible pour les différents variants de séquences. En conséquence, seule la séquence listée ci-dessus pour chaque domaine est considérée comme la séquence préférentiellement reconnue et donc comme séquence à prédire correctement.

### 2.2 Simulations de docking

Pour chaque domaine du jeu de données, les quatre types de nucléotides (A, C, G et U) ont été dockés à la surface de leur site d'interaction. Les structures nucléotidiques RAXN010, RCXN010, RGXN010 et RUXN010 ont été utilisées par défaut. Chaque structure correspond à un ribonucléotide présentant un plissement du ribose C3'-endo, une base orientée en anti, un groupement phosphate PO<sub>2</sub><sup>-</sup> en 5'-ter et une extrémité 3'-ter définie par un O3' protoné. Le site d'interaction est défini pour chaque domaine par une boîte parallélépipédique englobant le site d'interaction avec la chaîne ARN tri-nucléotidique. Cette boîte a été élargie de 5 Å dans les coordonnées x, y et z par rapport aux extrémités des chaînes (Fig. 34).

La préparation des protéines, les paramètres utilisés pour les simulations et l'évaluation énergétique des poses sont tels qu'indiqués au chapitre II.

## 2.3 Recherche de chaînes

Toutes les recherches de chaînes ont été faites avec le programme Molpy (chapitre III, section 2.3) avec comme seule contrainte une longueur de trois nucléotides.

# 3 Résultats

## 3.1 Augmentation de la combinatoire

L'approche FBDRNA peut se résumer en trois grandes étapes : (i) des nucléotides sont dockés, (ii) une stratégie de sélection est appliquée sur l'ensemble des poses résultantes pour retenir les poses d'intérêt, et (iii) des chaînes sont recherchées à partir de cette sélection identifiant les nucléotides qui peuvent être connectés entre eux par satisfaction de contraintes (distances, et séquence si le mode d'interaction d'une séquence connue est recherchée). Sans *a priori* sur la séquence, les quatre types de nucléotides doivent être dockés : A, C, G et U. La stratégie de sélection des poses "diviser pour mieux régner" mise au point dans le chapitre III se déroule de la manière suivante : après un clustering à 5 Å des poses issues du docking, les 10 clusters les plus peuplés sont sélectionnés. Après élimination de la redondance des poses au sein des 10 clusters sélectionnés, les 10 poses de plus basse énergie sont retenues pour chaque cluster. Le nombre total de poses ainsi retenues est de 100 par type de nucléotides dockés. Sans *a priori* sur la séquence ARN, le nombre total de poses sélectionnées s'élève à 400. On a déjà vu que le nombre de poses sélectionnées a une influence directe sur l'espace de solutions des chaînes recherchées. On s'attend donc à une augmentation évidente du nombre de chaînes lorsque la séquence n'est pas connue sur le simple fait que le nombre de poses sélectionnées sera plus important. Par ailleurs, l'absence de contrainte de séquence pour la recherche de chaînes conduit aussi à augmenter la complexité combinatoire. Plusieurs poses issues du docking de types de nucléotides différents peuvent en effet être retrouvées dans des positions et conformations très proches (elles peuvent par exemple présenter entre elles des valeurs  $RMSD \leq 2$  Å). Ainsi, si l'une de ces poses répond aux critères de distances et d'angles autorisant sa connexion à une pose voisine, les autres poses superposables à la première pourront également y répondre. Ces poses superposables issues du docking de différents types de nucléotides conduisent donc naturellement à augmenter le nombre de chaînes candidates. Notons que cette caractéristique s'applique également lorsque la séquence est connue et donnée en entrée, et qu'au moins deux types de nucléotides sont dockés ; mais dans ce cas, les chaînes recherchées devant répondre à une contrainte de séquence imposée, l'influence des poses superposables sur l'espace de solutions est limitée.

Pour illustrer l'augmentation de la complexité combinatoire découlant d'une recherche de chaînes sans contrainte de séquence imposée, les quatre types de nucléotides (A, C, G, et U) ont été dockés au niveau du site d'interaction des trois domaines de notre jeu de données (section 2.2). Pour chaque type de nucléotides, 100 poses ont été retenues d'après la stratégie de sélection "diviser pour mieux régner" mise en place dans le chapitre III. Des chaînes tri-nucléotidiques ont ensuite été recherchées à partir des 400 poses retenues (100 poses par type de nucléotide) et sans imposée de contrainte de séquence. Le tableau 14 montre que le nombre de chaînes identifiées à partir de ces poses est compris entre 400 000 et 900 000 environ selon les domaines. Cette quantité est trop importante pour pouvoir être traitée. La stratégie de sélection doit donc être adaptée pour prédire le mode d'interaction de chaînes lorsque leur séquence n'est pas connue.

*Tableau 14: Résultats d'une recherche de chaînes tri-nucléotidiques sans contrainte de séquences. La sélection des poses utilisées pour la recherche de chaînes inclut pour les quatre types de nucléotides dockés, les 10 clusters les plus peuplés (Nc), et pour chacun d'eux les 10 poses de plus basse énergie (Np).*

Code PDB	Paramètres de sélection des poses								Séquence recherchée	Nombre de chaînes identifiées
	A		C		G		U			
	Nc	Np	Nc	Np	Nc	Np	Nc	Np		
2XNR	10	10	10	10	10	10	10	10	Aucune	398 943
5ELH	10	10	10	10	10	10	10	10	Aucune	868 312
5WWX	10	10	10	10	10	10	10	10	Aucune	472 103

Les résultats présentés ci-dessous décrivent la mise en place d'une stratégie de sélection de poses pour une recherche de chaînes sans contrainte de séquence imposée. L'objectif est de pouvoir réduire l'espace de solutions des chaînes et de prédire, en plus du mode d'interaction natif d'une chaîne ARNsb, une séquence préférentiellement reconnue par un RBD.

### **3.2 Comparaison de l'énergie d'interaction entre poses natives et native-like**

Comme mentionné ci-dessus, lorsque des nucléotides de type différent (A, C, G ou U) sont dockés, des poses de base différente peuvent être retrouvées dans des positions et conformations superposables. Pour une recherche de chaînes sans contrainte de séquence imposée, ces poses superposables conduisent à augmenter la complexité combinatoire et donc à augmenter l'espace de solutions des chaînes. Une stratégie qui pourrait permettre de réduire cette complexité consisterait idéalement à sélectionner, autour de chaque position et conformation (définies par un "rayon" RMSD donné) de poses générées durant le docking, une seule pose correspondant au type de nucléotide de plus basse énergie. Pour que cette stratégie puisse être efficace et donc applicable, la

fonction de score utilisée doit être capable de correctement classer les types de nucléotides en fonction de leur préférence de liaison à un site d'interaction donné.

Afin d'évaluer la capacité de la fonction de score utilisée dans MCSS à classer correctement les types de nucléotides en fonction de leur préférence de liaison, l'énergie d'interaction des poses de plus basse énergie trouvées autour de chaque nucléotide expérimental composant les chaînes ARN de référence a été comparée pour chaque type de nucléotide docké. D'après les données d'affinité de liaison disponibles pour chaque système du jeu de données, les nucléotides expérimentaux composant les chaînes ARN de référence ont été considérés comme les nucléotides préférentiellement reconnus (section 2.1). Un critère RMSD de 2 Å a été utilisé pour définir une pose trouvée autour de chaque nucléotide expérimental. Les mesures RMSD impliquent des comparaisons entre nucléotides dont la base peut être différente de celle du nucléotide expérimental de référence ; afin que ces mesures puissent être comparables entre différents nucléotides, les calculs RMSD ont été réalisés à partir de l'ensemble des atomes lourds du squelette ribose-phosphate (à l'exception des atomes d'oxygène du groupement phosphate écartés pour des raisons de symétrie) et de seulement quatre atomes de la base (Fig. 42).

	Paire d'atomes			
<b>Purine</b>	N9	C8	N7	C4
<b>Pyrimidine</b>	N1	C6	C5	C2

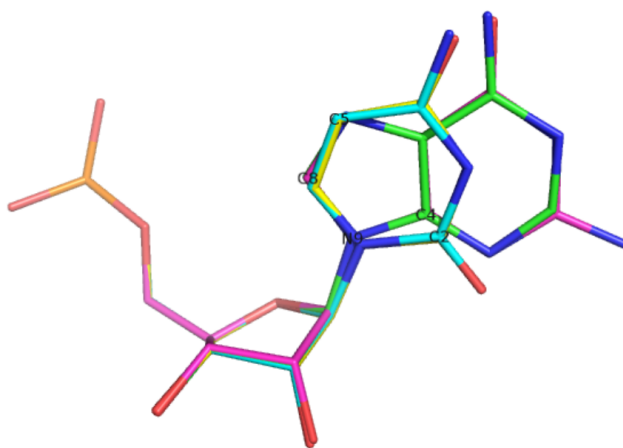


Figure 42: Paires d'atomes de la base comparés pour un calcul RMSD entre un nucléotide de base purique et un nucléotide de base pyrimidique.

D'une manière générale, pour s'y référer plus facilement, toute pose ayant un RMSD par rapport au nucléotide expérimental inférieur ou égal à 2 Å sera nommée pose native ou pose native-like :

une pose native fera référence à une pose dont la base est identique à celle du nucléotide expérimental, une pose native-like définira une pose de base différente. Pour chaque système, le classement établi par la fonction de score est considéré correct si la pose de plus basse énergie correspond à une pose native, c'est-à-dire une pose dont la base est identique à celle du nucléotide expérimental. Le nucléotide U1 de 5ELH n'est pas considéré dans les analyses (voir section 2.1).

*Tableau 15: Comparaison de l'énergie d'interaction des poses native et native-like de plus basse énergie pour chaque nucléotide expérimental. Les poses native et native-like sont définies par un RMSD  $\leq 2$  Å par rapport au nucléotide expérimental ; une pose native possède une base identique à celle du nucléotide expérimental, une pose native-like possède une base différente. Le classement des nucléotides établi en fonction de l'énergie d'interaction observée pour les poses native et native-like de plus basse énergie est précisé ; le nucléotide expérimental attendu au meilleur rang est mis en évidence en gras.*

Code PDB	Nucléotide expérimental	Energie d'interaction (kcal/mol) de la pose native/native-like de plus basse énergie par type de nucléotide docké				Classement des poses native et native-like selon l'énergie d'interaction
		A	C	G	U	
2XNR	U1	-12,02	-12,49	-13,84	-12,64	G > U > C > A
	C2	-20,67	-20,63	-18,85	-17,61	A > C > G > U
	U3	-18,61	-16,67	-17,24	-17,39	A > U > G > C
5ELH	U2	-16,95	-12,88	-25,02	-21,19	G > U > A > C
	A3	-25,58	-19,81	-24,33	-17,57	A > G > C > U
5WWX	A1	-18,52	-12,69	-19,53	-14,14	G > A > U > C
	G2	-20,48	-18,23	-24,03	-16,87	G > A > C > U
	A3	-19,72	-17,43	-19,67	-16,91	A > G > C > U

Le tableau 15 ci-dessus montre que le classement est correct pour seulement trois nucléotides (A3 de 5ELH, et G2 et A3 de 5WWX) et que pour les cinq autres nucléotides, la pose native de plus basse énergie est classée en deuxième position. Deux remarques importantes peuvent être faites au regard des classements observés. La première concerne l'énergie d'interaction qui sépare les poses dont l'écart est parfois très faible : seul 0,04 kcal/mol sépare par exemple la pose native C de 2XNR classée en deuxième position de la pose native-like A classée en tête de liste ; ou bien encore pour le nucléotide A3 de 5WWX, la pose native A classée au premier rang a une énergie d'interaction plus favorable de seulement 0,05 kcal/mol par rapport à la pose G. Le classement observé pour ces poses peut donc être discutable. La deuxième remarque concerne le type de nucléotides observé en première position : pour les huit nucléotides expérimentaux du jeu de données, ce nucléotide est une

purine. La présence systématique d'une purine en tant que pose de plus basse énergie d'interaction reflète un biais de la fonction de score utilisée. L'énergie d'interaction estimée par la fonction de score MCSS est issue de la somme de différents termes énergétiques évalués entre paires d'atomes du ligand et de la protéine. En conséquence, l'énergie d'interaction de nucléotides puriques tend à être globalement plus favorable que l'énergie d'interaction de nucléotides pyrimidiques puisque ces derniers présentent un nombre d'atomes moins important (voir "Annexes", section 3). Compte tenu de ce biais, il est assez remarquable d'observer que les seuls cas où une pyrimidine est classée en deuxième position correspondent au cas où le nucléotide expérimental est aussi une pyrimidine : U1, C2 et U3 de 2XNR, et U2 de 5ELH. Pour ces quatre nucléotides, l'écart d'énergie séparant la pose native classée au 2ème rang de la pose classée au 3ème rang varie entre 0,15 kcal/mol (nucléotides U1 et U3 de 2XNR) et 4,24 kcal/mol (nucléotide U2 de 5ELH).

Malgré des écarts d'énergie qui peuvent être faibles entre poses native et native-like, nous pensons que le classement systématique de poses natives parmi les deux poses de plus basse énergie reflète une certaine capacité de la fonction de score MCSS à favoriser énergétiquement des contacts impliqués dans la reconnaissance sélective de nucléotides. Cette caractéristique peut être suffisante pour réduire le nombre de poses à sélectionner pour la prédiction d'une séquence ARN préférentiellement reconnue par un RBD. La section suivante décrit une adaptation du protocole de sélection "diviser pour mieux régner" tirant partie de la capacité de la fonction de score utilisée à classer parmi les deux poses de plus basse énergie un nucléotide préférentiellement reconnu au niveau d'un site d'interaction donné.

### **3.3 Adaptation de la stratégie de sélection des poses "diviser pour mieux régner"**

Les observations décrites ci-dessus permettent d'envisager une stratégie de sélection permettant de réduire la complexité combinatoire inhérente à une recherche de chaînes effectuée sans contraintes de séquence. Le principe de cette sélection, qui peut se décliner en trois grandes étapes, est illustrée à la figure 43. Après un docking indépendant des quatre nucléotides au niveau du site d'interaction d'un RBD, le protocole de sélection "diviser pour mieux régner" mis en place dans le chapitre précédent est appliqué à chaque ensemble de poses issu des calculs de docking (étape 1, Fig. 43). Ce protocole consiste en une première étape de clustering à 5 Å visant à regrouper entre elles les poses trouvées autour de chaque site d'interaction nucléotidique. Parmi l'ensemble des groupes résultants, les 10 clusters les plus peuplés sont retenus. Les poses trouvées dans chacun de ces 10 clusters sont alors regroupées par un clustering à 2 Å de manière à éliminer la redondance ; les 10 poses représentatives (pose de plus basse énergie dans chaque sous-groupe) sont finalement



retenues dans chacun des 10 clusters à 5 Å. A ce stade, cette première étape de sélection conduit à retenir 400 poses : 100 poses A, 100 pose C, 100 poses G et 100 poses U. La suite du protocole consiste à joindre les quatre ensembles de 100 poses en une seule distribution, puis à appliquer un clustering à 2 Å sur l'ensemble de cette distribution (étape 2, Fig. 43). Chaque groupe résultant est alors potentiellement composé de plusieurs types de nucléotides (poses A, C, G et/ou U) dont les conformations et positions sont très proches (Fig. 54 en "Annexes" ). Cette configuration est alors comparable à la configuration à partir de laquelle les observations décrites à la section 3.2 ci-dessus ont été réalisées. Ces observations ont montré qu'en sélectionnant les deux nucléotides de plus basse énergie trouvés autour de la position du ligand expérimental (selon un  $\text{RMSD} \leq 2 \text{ \AA}$ ), la pose native (pose de type nucléotidique attendu) est retenue. La troisième et dernière étape du protocole de sélection consiste donc à sélectionner dans chacun des groupes les deux nucléotides de plus basse énergie (étape 3, Fig. 43 et Fig. 54 pour une illustration plus détaillée donnée en "Annexes"). Précisons que si un seul type de nucléotides est trouvé dans un groupe de poses, alors seule la pose de plus basse énergie est retenue. Par ailleurs, les clusters ne contenant qu'une seule pose sont éliminés en partant de l'hypothèse que les positions et conformations où au moins deux poses ont convergé sont plus probables. Cela permet également de réduire le nombre total de poses retenues.

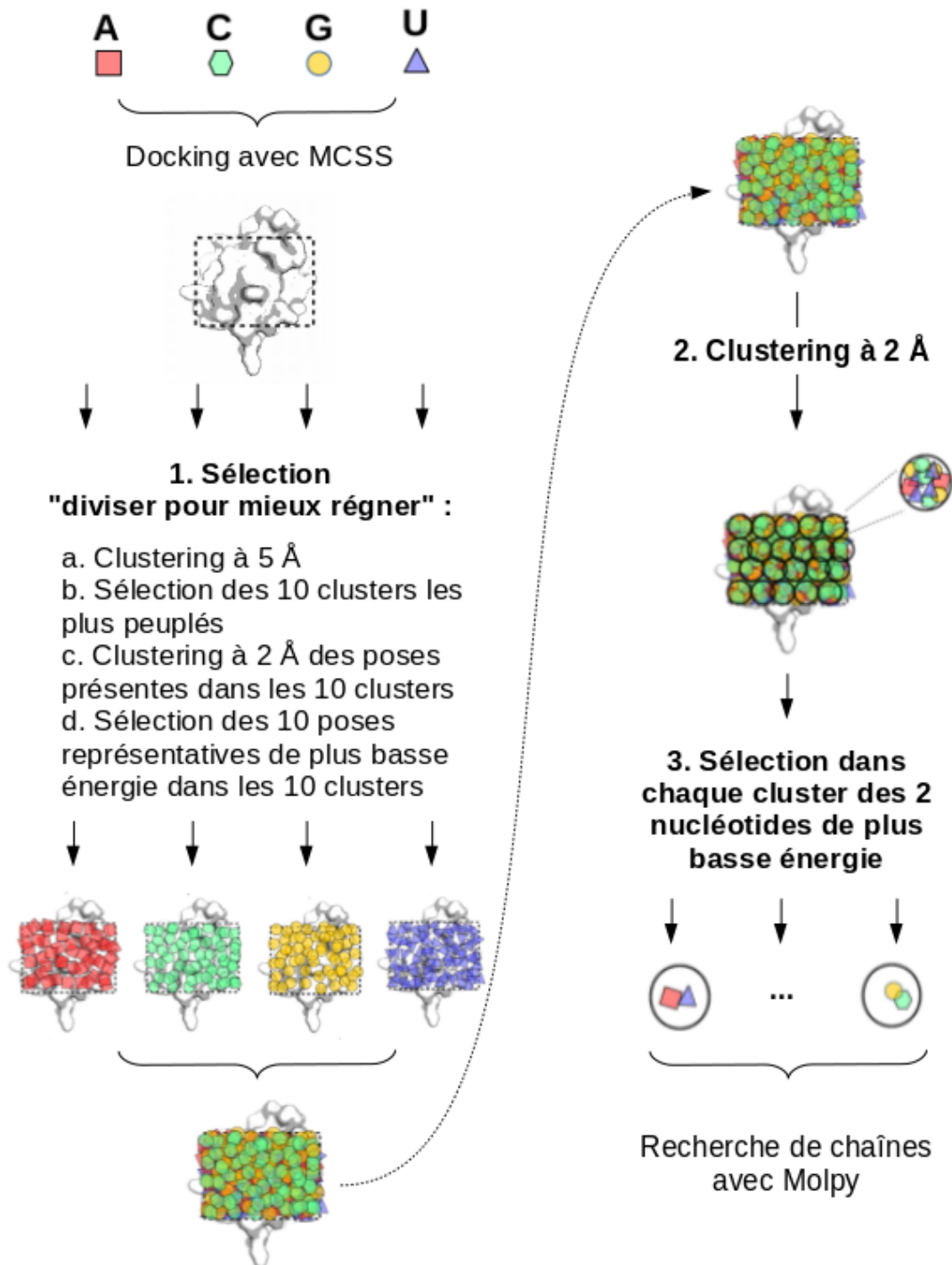


Figure 43: Illustration du principe de la stratégie de sélection des poses "diviser pour mieux régner" adaptée pour la recherche de chaînes sans contrainte de séquence imposée. Après docking des quatre nucléotides au niveau du site d'interaction d'un RBD, la procédure de sélection des poses peut se décliner en trois grandes étapes : (1) la stratégie de sélection "diviser pour mieux régner" est appliquée sur chacune des distributions de poses (distributions de poses A, C, G et U) et conduit à retenir 100 poses par type de nucléotides dockés ; après regroupement de ces quatre distributions, (2) l'ensemble des poses résultant est soumis à un clustering à 2 Å ; (3) pour chaque groupe obtenu, les deux nucléotides de plus basse énergie sont retenues. L'ensemble des poses ainsi sélectionnées peut alors être donné en entrée pour une recherche de chaînes ARN sans contrainte de séquence imposée.

Le tableau 16 montre le nombre de poses retenues pour chaque type de nucléotide docké après application de ce protocole de sélection sur les trois systèmes du jeu de données. Le nombre total de poses retenues pour chaque complexe est réduit de plus d'un facteur trois par rapport à une sélection établie uniquement d'après le protocole "diviser pour mieux régner" qui conduit à retenir 400 poses au total. On peut remarquer que le nombre de purines retenues est à peu près deux fois supérieur au nombre de pyrimidines. Cela reflète à la fois la stratégie de sélection suivie (sélection des deux nucléotides de plus basse énergie - étape 3 de la figure 43) et le biais de l'estimation de l'énergie d'interaction par la fonction de score MCSS qui tend à favoriser les purines par rapport aux pyrimidines (voir "Annexes" section 3).

*Tableau 16: Nombre de poses retenues à partir de la stratégie de sélection pour chaque nucléotide docké.*

Code PDB	Nombre de poses retenues par type de nucléotide				Nombre total de poses
	A	C	G	U	
2XNR	41	29	41	26	137
5ELH	48	26	52	21	147
5WWX	52	19	45	24	140

En plus de réduire le nombre total de poses, l'application du protocole de sélection permet de retenir des poses native et native-like pour chaque nucléotide expérimental des chaînes ARN. Le tableau 17 montre pour chaque nucléotide expérimental l'énergie d'interaction des poses native et native-like de plus basse énergie. Le classement des poses native et native-like observé pour chaque nucléotide expérimental montre que la procédure de sélection suivie reproduit dans sept cas sur huit le classement observé lorsque les poses native et native-like sont directement définies à partir de la connaissance de la position du nucléotide expérimental (tableau 15). La seule différence concerne le nucléotide A1 de 5WWX pour qui une seule pose native est retrouvée. Le classement observé dans le tableau 15 montre que cette pose était classée en deuxième position derrière une pose native-like G. Le fait que cette pose G ne soit pas retrouvée après la procédure de sélection s'explique par l'étape de clustering à 2 Å (étape 3 – Fig. 43) qui peut conduire à regrouper des poses natives et native-like dans des clusters différents. Cela explique également pourquoi trois poses native et native-like sont retrouvées pour le nucléotide U1 de 2XNR, et pourquoi une seule pose native est aussi retrouvée pour le nucléotide A3 de 5WWX.

Tableau 17: Comparaison de l'énergie d'interaction des poses native et native-like de plus basse énergie retenues après la stratégie de sélection. Les poses native et native-like sont définies par un RMSD  $\leq 2$  Å par rapport au nucléotide expérimental ; une pose native possède une base identique à celle du nucléotide expérimental, une pose native-like possède une base différente. Le classement des nucléotides établi en fonction de l'énergie d'interaction observée pour les poses native et native-like de plus basse énergie est précisé ; le nucléotide expérimental attendu au meilleur rang est mis en évidence en gras.

Code PDB	Nucléotide expérimental	Energie d'interaction (kcal/mol) de la pose native/native-like de plus basse énergie par type de nucléotide docké				Classement des poses native et native-like selon l'énergie d'interaction
		A	C	G	U	
2XNR	U1	/	-12,49	-13,84	-12,64	<b>G &gt; U &gt; C</b>
	C2	-20,67	-20,63	/	/	<b>A &gt; C</b>
	U3	-18,61	/	/	-17,39	<b>A &gt; U</b>
5ELH	U2	/	/	-25,02	-21,19	<b>G &gt; U</b>
	A3	-25,58	/	-24,33	/	<b>A &gt; G</b>
5WWX	A1	-18,52	/	/	/	<b>A</b>
	G2	-20,48	/	-24,03	/	<b>G &gt; A</b>
	A3	-19,72	/	/	/	<b>A</b>

### 3.4 Recherche de chaînes à partir des poses retenues par la procédure de sélection adaptée "diviser pour mieux régner"

Globalement, la procédure de sélection mise en place permet de réduire le nombre de poses tout en conservant les poses natives. Cela suggère que le mode d'interaction de la séquence préférentiellement reconnue par les RBDs de notre jeu de données peut être prédit. La sélection des poses retenues est-elle néanmoins suffisamment réduite pour limiter la complexité combinatoire et donc l'espace de solutions inhérente à une recherche de chaînes faite sans imposer une séquence comme contrainte ? Pour répondre à cette question, une recherche de chaînes a été faite sans contrainte de séquence imposée à partir des poses sélectionnées dont le nombre par type de nucléotides est précisé dans le tableau 16. Des chaînes tri-nucléotidiques (3-nt) ont été recherchées pour les trois domaines, y compris pour le domaine 5ELH pour qui le nucléotide U1 de la chaîne 3-nt n'est pas considéré. Le tableau 18 résume les résultats de cette recherche. Entre 17000 et 53000 chaînes ont été identifiées selon les domaines. Comparé au nombre de chaînes identifiées à partir d'une sélection de 100 poses par type de nucléotides (tableau 14), le nombre de chaînes trouvées est réduit d'un facteur 16 pour 2XNR et 5WWX, et d'un facteur 23 pour 5ELH. La procédure de

sélection des poses permet donc de réduire d'une manière considérable l'espace de solutions des chaînes lorsque aucune contrainte de séquence n'est imposée.

*Tableau 18: Résultats de recherches de chaînes tri-nucléotidiques effectuées sans contrainte de séquence sur les trois domaines. Les poses utilisées en entrée pour la recherche de chaînes sont issues de la procédure de sélection décrite à la section 2.3. Les chaînes native et native-like sont définies comme toute chaîne présentant un RMSD  $\leq 2$  Å par rapport à la chaîne ARN de référence ; une chaîne native correspond à une chaîne dont la séquence correspond à la chaîne ARN de référence, et une chaîne native-like une chaîne dont la séquence est différente. La séquence de la chaîne ARN de référence est UCU pour 2XNR, NUA pour 5ELH et AGA pour 5WWX. Le premier nucléotide de la chaîne UUA de 5ELH n'étant pas considéré, le RMSD calculé pour définir des chaînes natives et native-like pour ce domaine ne considère en référence que le di-nucléotide UA. Le N indique que n'importe quel nucléotide peut être trouvé à la position U1.*

Code PDB	Nombre total de chaînes identifiées	Chaînes native-like				Chaînes natives			
		Nb	Chaîne de plus basse énergie			Nb	Chaîne de plus basse énergie		
			Rang (‰)	RMSD (Å)	Séquence		Rang (‰)	RMSD (Å)	Séquence
2XNR	17 251	17	6 (0,35)	1,96	CCA	2	32 (1,85)	1,63	UCU
5ELH	52 929	81	27 (0,51)	1,43	NUG	14	34 (0,64)	1,77	NUA
5WWX	29 021	15	19 (0,65)	1,43	UGA	6	2 (0,07)	1,43	AGA

En considérant une durée moyenne de 6s par CPU pour optimiser une chaîne 3-nt (suite du protocole de recherche de chaînes après leur identification), 28h sont nécessaires pour optimiser les 17 251 chaînes identifiées pour 2XNR sur un CPU, et 88h pour les 52 929 chaînes de 5ELH. Les temps de calculs requis sont importants dans l'optique d'une application à plus grande échelle, mais nous considérons cependant qu'ils restent raisonnables pour une première version de l'approche. Chaque chaîne 3-nt a donc été minimisée pour optimiser leur interaction avec leur domaine respectif.

Connaissant à l'avance que des poses natives et native-like sont présentes pour chaque nucléotide expérimental dans la sélection de poses utilisées pour la recherche de chaînes (tableau 17), on s'attend à ce que des chaînes natives et native-like soient générées. Ces chaînes ont été définies comme des chaînes présentant un RMSD  $\leq 2$  Å par rapport aux chaînes ARN de référence ; une chaîne native est définie par une chaîne dont la séquence correspond à la séquence attendue (celle préférentiellement reconnue), et une chaîne native-like correspond à une chaîne de séquence différente. Le premier nucléotide de la chaîne UUA de 5ELH n'étant pas considéré, les chaînes 3-nt générées pour ce domaine sont définies comme natives ou native-like uniquement par rapport au di-nucléotide UA. Le tableau 18 montre pour les trois domaines que le nombre de chaînes native-like

générées est plus important que le nombre de chaînes natives. On peut également remarquer que les chaînes natives et native-like de plus basse énergie sont toutes classées au moins parmi les 34 chaînes de meilleure énergie d'interaction sur l'ensemble des chaînes générées pour les trois domaines. Ce rang est excellent compte tenu du nombre total de chaînes ; il correspond au moins au top 0,65 % de l'ensemble des chaînes pour 5ELH et 5WWX, et au top 1,85 % pour 2XNR. La chaîne native-like de plus basse énergie présente un rang plus bas que celui de la chaîne native pour 2XNR et 5ELH. Pour 5WWX en revanche, la chaîne native AGA de plus basse énergie possède un rang inférieur à celui de la chaîne native-like ; cette chaîne native est par ailleurs classé au deuxième rang parmi les 29021 chaînes identifiées. D'une manière générale, ces résultats montrent que la fonction de score utilisée pour l'estimation de l'énergie d'interaction des chaînes permet de discriminer avec une excellente précision le mode d'interaction natif (top 1,85 % dans le pire des cas), mais présentent une moindre performance pour discriminer la séquence ARN préférentiellement reconnue parmi l'ensemble des chaînes natives et native-like : seule la séquence attendue AGA de 5WWX présente une énergie d'interaction plus favorable que les séquences des chaînes native-like.

### **3.5 Analyse de la composition nucléotidique des chaînes natives et native-like**

Cette dernière section du chapitre présente les résultats de l'analyse de la composition en nucléotides de l'ensemble des chaînes natives et native-like générées pour les trois domaines. Ainsi, pour chaque domaine, toutes les chaînes dont le RMSD par rapport aux chaînes ARN de référence est inférieur ou égal à 2 Å ont été regroupées, puis les nucléotides A, C, G et U trouvés à chaque position de la séquence des chaînes ont été dénombrés. Un motif logo a alors été construit à partir de ce dénombrement. La figure 44 montre que la séquence consensus qui ressort des motifs obtenus correspond à la séquence attendue pour les domaines 2XNR et 5WWX mais pas pour 5ELH. Pour ce dernier, la séquence consensus qui ressort est NUG alors que la séquence NUA est attendue. On peut remarquer pour 5ELH qu'un A est également observé à la position du G, mais en nombre moins important. Il est important de préciser que le nombre d'un nucléotide donné observé à une position ne se distingue pas toujours clairement du nombre d'un nucléotide différent. Par exemple, 6 C, 5 G et 8 U sont trouvés à la première position de la séquence des chaînes natives et native-like de 2XNR, et 10 A et 11 G sont trouvés en seconde position pour 5WWX. Ces faibles différences imposent d'interpréter avec précaution les séquences consensus obtenues.

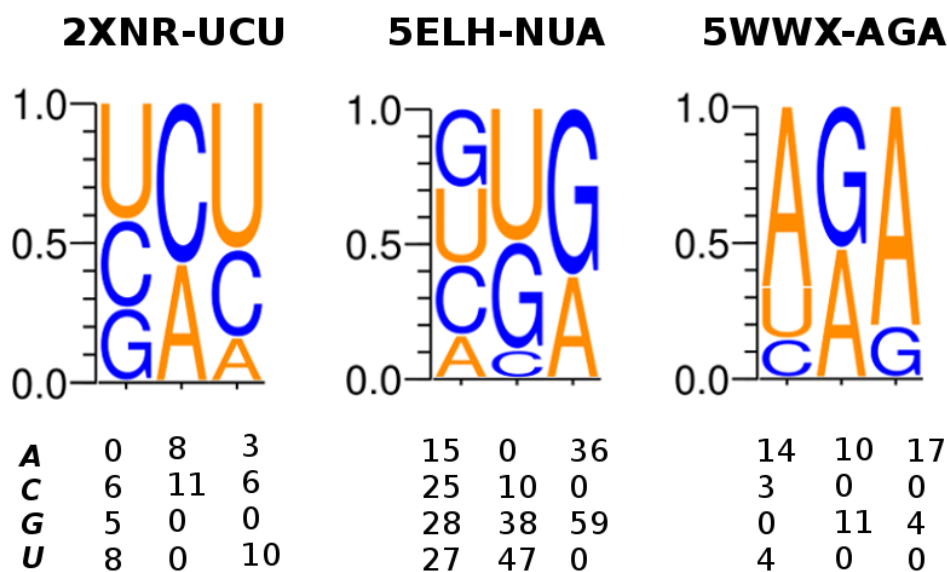


Figure 44: Motifs logo issus du dénombrement des nucléotides A, C, G et U trouvés à chacune des positions dans les séquences des chaînes natives et native-like générées pour les trois domaines. Au-dessus de chaque motif logo est indiqué le nom du domaine et la séquence qu'il reconnaît préférentiellement. Pour le domaine 5ELH, le premier nucléotide n'étant pas considéré, n'importe quel nucléotide peut être trouvée à cette position représentée par un N. Au-dessous des motifs est indiqué le nombre de A, C, G et U trouvés à chaque position.

## 4 Discussion

Déchiffrer le code de reconnaissance protéine-ARN représente un intérêt capital qui permettrait d'avoir une meilleure compréhension des interactions mises en jeu et d'ouvrir la voie à la construction rationnelle de protéines pour aller cibler des ARN d'intérêt. Le meilleur moyen de rationaliser les mécanismes impliqués dans la reconnaissance spécifique entre protéines et ARN repose certainement sur l'accessibilité à des structures 3D de leur complexe. L'obtention de ces structures par les méthodes de cristallographie aux rayons X ou RMN peut être longue et laborieuse et aucune approche computationnelle ne permet à ce jour de modéliser le mode d'interaction protéine-ARN et de prédire en même temps une séquence ARN préférentiellement reconnue par une protéine ou un domaine de liaison à l'ARN. Les travaux présentés dans ce chapitre ont été réalisés avec l'objectif de pouvoir prédire correctement la séquence ARN préférentiellement reconnue par un RBD en même temps que son mode d'interaction. Pour ce faire, l'approche FBDRNA présentée au chapitre précédent a été adaptée pour répondre au problème de la recherche de chaînes sans connaissance *a priori* de la séquence ARN.

L'approche a été développée et appliquée à partir de la forme liée des trois RBDs utilisés au chapitre III : les domaines RRM (code PDB : 2XNR), KH (code PDB : 5WWX) et Zn-CCCH (code PDB : 5ELH). Pour ces trois RBDs, une courte séquence ARN liée dans la structure cristallographique a été considérée comme la séquence préférentiellement reconnue sur la base de données d'affinités de liaison disponibles pour quelques variants de séquences (chapitre III, section 2.1). A partir de la structure du RBD et de la connaissance de son site d'interaction, la séquence native attendue a pu être discriminée des séquences non-natives sur la base de l'énergie d'interaction pour le domaine KH ; de plus son mode d'interaction a été prédit au deuxième rang parmi près de 30 000 chaînes identifiées (tableau 18). Pour les domaines RRM et Zn-CCCH, la séquence native présente une énergie d'interaction moins favorable par rapport aux séquences non-natives ; leur mode d'interaction est néanmoins prédit dans les 35 chaînes de plus basse énergie parmi près de 17 000 chaînes identifiées pour le domaine RRM et 53 000 pour le domaine Zn-CCCH (tableau 18).

La prédiction d'une séquence préférentielle par l'approche FBDRNA impose initialement de docker les quatre nucléotides au niveau du site d'interaction d'un RBD. Puisque la recherche de chaînes est effectuée sans imposer de contrainte de séquence, la combinatoire est inévitablement augmentée par rapport à une recherche de chaînes réalisée à partir d'une séquence ARN connue. Le protocole de sélection mis en place précédemment (chapitre III) a donc dû être adapté afin de limiter l'espace de solutions. L'application de ce protocole s'est montrée efficace puisqu'il a permis de réduire de 16 à 23 fois le nombre de chaînes identifiées par rapport au protocole standard. Pour parvenir à cette réduction, le postulat suivant a dû être suivi : lorsque plusieurs nucléotides se superposent autour d'une position donnée (définie par un rayon RMSD de 2 Å environ), la sélection des deux nucléotides de plus basse énergie est suffisante pour retenir le nucléotide préférentiellement reconnu à cette position. Ce postulat découle d'observations faites sur les huit nucléotides expérimentaux utilisés comme référence. Néanmoins, compte tenu du faible nombre de données de référence, ce postulat demande à être validé sur un jeu de données plus important, d'autant plus que l'énergie d'interaction séparant deux nucléotides est parfois infime. Les analyses ont par ailleurs montré un biais de la fonction de score qui tend à favoriser l'énergie d'interaction des purines par rapport aux pyrimidines. Ce biais pourrait être évité en normalisant les énergies d'interaction par un Z-score.

D'une manière plus générale, la stratégie de sélection mise en place repose sur la capacité de la fonction de score à classer correctement des nucléotides en fonction de leur affinité de liaison. Pour que cette stratégie soit parfaitement efficace, la fonction de score devrait idéalement affecter



l'énergie d'interaction la plus basse au nucléotide présentant la meilleure affinité de liaison. La capacité à classer des ligands se liant à un même site d'interaction avec des affinités différentes est un des critères utilisés pour évaluer et comparer les performances de fonctions de score dans le docking protéine-ligand. Plusieurs fonctions de score ont montré des performances de classement intéressantes sur le jeu de données CASF composé d'une centaine de complexes protéine-ligand variés (Su et al., 2019). Tester ces fonctions de score dans l'approche FBDRNA pourrait également être une solution pour améliorer les prédictions.

Globalement, la fonction de score utilisée pour l'estimation de l'énergie d'interaction des chaînes ARN a montré d'excellentes capacités à discriminer le mode d'interaction natif. Pour les trois RBDs, une chaîne native a pu être systématiquement classée au moins dans le top 35 des chaînes de plus basse énergie. La capacité de la fonction de score à discriminer la séquence préférentielle parmi les chaînes natives et native-like s'est en revanche montrée moins satisfaisante : seule la séquence native du domaine KH présente une énergie d'interaction plus favorable par rapport aux séquences des chaînes native-like. Dans les interactions RBD-ARNsb, les liaisons hydrogène jouent un rôle très important pour la reconnaissance spécifique des bases par la protéine. La fonction de score utilisée ne traite les liaisons hydrogène que d'une manière implicite par les termes coulombique et de van der Waals. Par ailleurs, les molécules d'eau, qui peuvent également contribuer à établir des liaisons hydrogène importantes pour la reconnaissance spécifique des bases, ne sont pas considérées explicitement dans l'approche FBDRNA. La représentation simplifiée du système ainsi que l'absence de traitement explicite des liaisons hydrogène dans la fonction de score utilisée peuvent expliquer les difficultés rencontrées pour identifier les séquences préférentiellement reconnues par les RBDs de notre jeu de données. Néanmoins, l'excellente capacité de la fonction de score à discriminer les chaînes reproduisant le mode d'interaction natif indépendamment de leur séquence permet d'envisager la sélection d'un nombre réduit de candidats pour une estimation plus précise de leur énergie d'interaction, et potentiellement par de courtes simulations de dynamique moléculaire en modèle de solvant explicite.

Le dénombrement des nucléotides trouvés à chaque position dans la séquence de l'ensemble des chaînes natives et native-like a fait ressortir une séquence consensus correspondant à la séquence ARN préférentielle pour les domaines RRM et KH. Cette observation indique que, pour ces deux cas, les nucléotides préférentiels à une position donnée sont présents dans un plus grand nombre de chaînes par rapport aux nucléotides non-natifs. Cela suggère que les nucléotides préférentiels à une position donnée répondent plus favorablement aux contraintes de distances utilisées pour construire les chaînes ARN. Cette propriété demande à être confirmée sur un jeu de données plus important

mais elle suggère que prendre en compte le nombre de nucléotides trouvés à une position donnée peut être une voie à explorer pour améliorer la discrimination d'une séquence préférentiellement reconnue parmi les chaînes natives et native-like, indépendamment de leur énergie d'interaction. Cette information peut être capturée par une étape de clustering visant à regrouper les chaînes les plus proches entre elles.

Globalement, les travaux réalisés et présentés dans ce chapitre ont permis de mettre au point un protocole de sélection pour l'approche FBDRNA permettant de réduire la combinatoire inhérente à la recherche de chaînes sans contrainte de séquence imposée. L'application de ce protocole sur la forme liée de trois RBDs a permis de générer de courtes chaînes ARN reproduisant le mode d'interaction natif classées dans le top 35 des chaînes de plus basse énergie parmi quelques dizaines de milliers de chaînes générées. Bien que la séquence native n'ait pu être discriminée des séquences non-natives que pour un seul cas, les résultats restent encourageants pour la suite au regard des perspectives d'amélioration existantes.

## V Etude de l'influence de la structure du nucléotide sur les performances de docking

### 1 Introduction

L'approche FBDRNA développée durant cette thèse a pour objectif la prédiction du mode d'interaction entre des domaines de liaison à l'ARN (RBD) et des chaînes d'ARN, de séquences connues ou non, et supposées être sans élément de structures secondaires dans leur forme liée. La méthode peut se décliner en trois grandes étapes : (i) des nucléotides sont dockés au niveau du site d'interaction d'un RBD avec le programme MCSS, (ii) une sélection des poses d'intérêt est effectuée et (iii) des chaînes d'ARN sont construites en identifiant à partir de ces poses quelles sont celles pouvant être connectées entre elles par satisfaction de contraintes (distances, séquences). Les chapitres précédents se sont focalisés sur le développement de stratégies de sélection de poses visant à pouvoir reproduire le mode d'interaction natif de chaînes à partir de leur séquence (chapitre III) ou non (chapitre IV). Les stratégies de sélection mises au point et, d'une manière générale, le succès de l'approche FBDRNA, reposent en premier lieu sur les performances de docking. Jusqu'ici, tous les calculs de docking ont été réalisés à partir d'une structure de nucléotide utilisée par défaut qui présente une conformation pré-définie où le plissement du ribose est en C3'-endo et la base orientée en configuration anti. Par ailleurs, la structure utilisée comporte à l'extrémité O5' du ribose un groupement phosphate composé de deux atomes d'oxygène, et un groupement OH à l'extrémité 3' du ribose (Fig. 33). MCSS offre cependant la possibilité à l'utilisateur de choisir parmi une variété de structures nucléotidiques (chapitre II, section 1.3). Ces structures varient en fonction du plissement du ribose (C3'-endo ou C2'-endo), de l'orientation de la base (anti ou syn) ou bien du type d'atomes présents aux extrémités O5', O3' et/ou O2' du ribose. On peut donc s'attendre à observer des performances différentes dans les résultats de docking en fonction de la structure du ligand utilisée. L'une de ses structures est-elle avantageuse par rapport à une autre ? Conduit-elle à un échantillonnage plus efficace en générant plus de poses natives ? Ces poses natives sont-elles plus facilement discriminées par rapport à celles obtenues à partir d'autres structures de ligand ? Ce sont les questions qui ont été traitées dans ce chapitre dans la perspective *in fine* de potentiellement optimiser l'approche FBDRNA en favorisant la sélection de poses d'intérêt pour l'identification de modèles de chaînes ARNs.

Pour des raisons de temps de calculs, toutes les structures implémentées dans MCSS n'ont pu être testées. Les comparaisons se sont focalisées sur seulement cinq structures différentes de nucléotides qui sont représentées schématiquement à la figure 45.

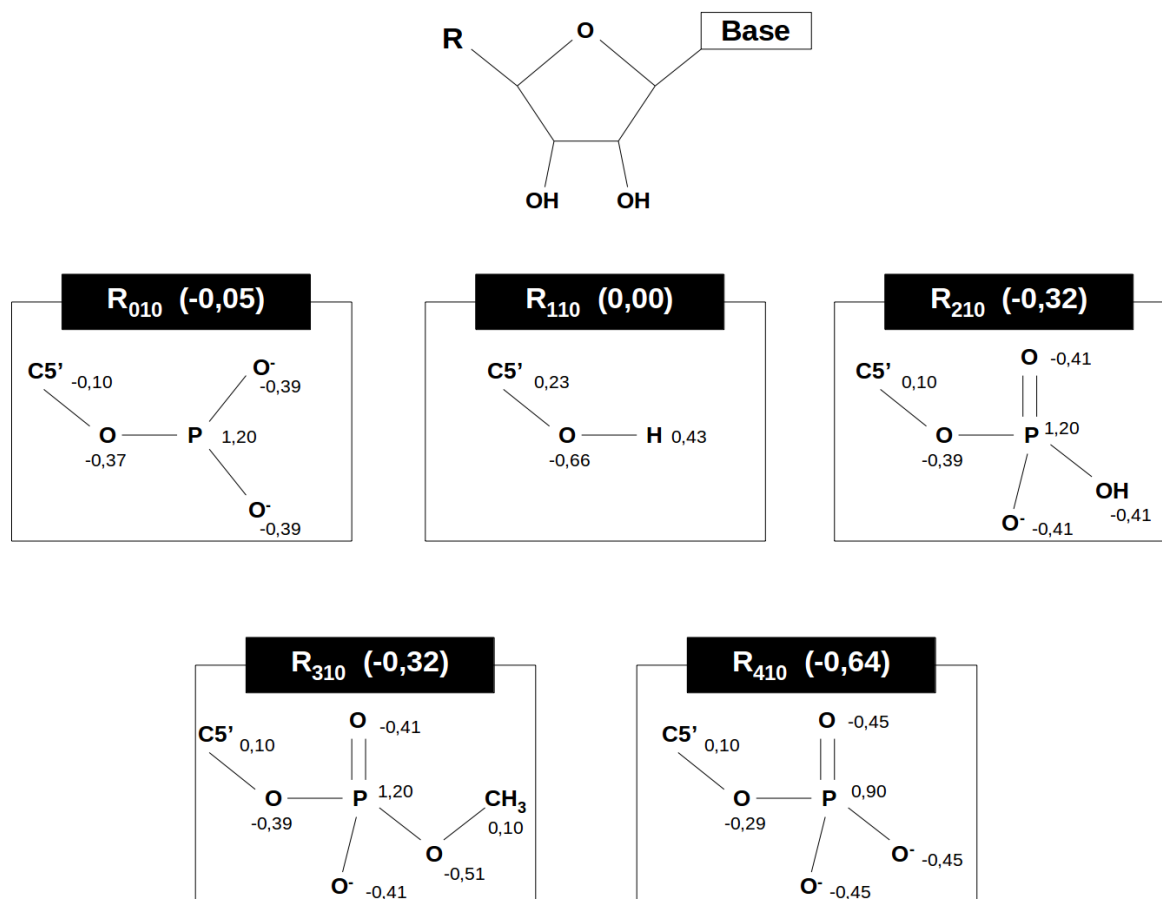


Figure 45 : Description des propriétés des cinq structures nucléotidiques utilisées. Les structures diffèrent par leur groupement chimique attaché à l'extrémité C5'-O5' du ribose, représentée par la lettre R. Les chiffres apposés en indice suivent la nomenclature MCSS. La charge globale des nucléotides est indiquée entre parenthèses, tandis que la charge de chaque atome est précisée à côté de leur nom. Précisons que la charge portée par les atomes du ribose et de la base ne diffère pas entre les différentes structures.

Elles sont accessibles dans MCSS d'après la nomenclature (chapitre II, section 1.3) suivante : RNXN010 (010), RNXN110 (110), RNXN210 (210), RNXN310 (310) et RNXN410 (410), où N définit le type de bases (A, C, G et U). Ces cinq structures présentent un plissement du ribose en C3'-endo et une base orientée en anti. Leur extrémité O3' contient un atome d'hydrogène. Elles diffèrent en revanche au niveau de leur extrémité O5'-terminale et en conséquence de leur charge globale. La structure 010 est la structure par défaut utilisée durant les chapitres précédents. Elle porte un groupement phosphate à l'extrémité O5' du ribose composé de deux atomes d'oxygène, et sa charge globale est de -0,05. La structure 110 ne contient qu'un atome d'hydrogène attaché à l'O5' du ribose ; sa charge est nulle. La structure 210 porte à l'extrémité O5' un groupement phosphate avec trois atomes d'oxygène, l'un d'eux étant protoné ; sa charge est de -0,32. La structure 310 correspond à un nucléotide modifié où un groupement méthyle est attaché à l'un des

trois atomes d'oxygène du groupement phosphate ; sa charge est également de -0,32. Enfin, la structure 410 porte en O5' un groupement phosphate avec trois atomes d'oxygène, tous étant déprotonés ; sa charge est de -0,64.

Afin d'avoir une vision représentative de l'influence de ces cinq structures sur les performances de docking, un jeu de données suffisamment important était requis. Les travaux s'inscrivent directement dans une perspective d'amélioration des performances de l'approche FBDRNA, ce jeu de données doit par ailleurs refléter les modes d'interaction entre protéines et ARN simple-brin où l'ARN ne présente pas d'éléments de structures secondaires dans sa forme liée. Pour répondre à ces deux critères, un ensemble non-redondant de complexes protéine-ribonucléotide-5'-P (nucléotide) a préalablement été sélectionné. Les résultats présentés ci-dessous sont issus de calculs de docking réalisés à partir d'un jeu de données de 120 complexes protéine-nucléotide. La comparaison des cinq structures de nucléotides utilisées pour les calculs de docking s'est focalisée sur les performances d'échantillonnage et du scoring dans sa capacité à discriminer correctement des poses natives. La stratégie de sélection des poses appliquée durant l'approche FBDRNA implique une étape de clustering des poses visant à réduire leur nombre en éliminant les poses redondantes. L'influence de l'élimination de la redondance des poses issues du docking sur les performances de docking a donc également été évaluée.

## **2 Matériels et méthodes**

### **2.1 Sélection du jeu de données protéine-nucléotide**

Un ensemble de complexes protéine-ribonucléotide-5'-P (nucléotide) a été collecté sur la banque de données PDB (Si, Cui, Cheng, & Wu, 2015a). Les requêtes se sont focalisées sur des structures cristallographiques de résolution  $\leq 2\text{\AA}$  correspondant à des protéines en association avec l'un des quatre nucléotides standards. Quatre requêtes indépendantes ont donc été réalisées, une pour chaque nucléotide dont les identifiants chimiques utilisés sont :

- AMP pour l'adénine 5'-P (A)
- C5P pour la cytidine 5'-P (C)
- 5GP pour la guanine 5'-P (G)
- U5P pour l'uracile 5'-P (U)

Pour chacun des groupes résultants (protéine-A, protéine-C, protéine-G et protéine-U), l'ensemble des complexes a été clusterisé pour éliminer la redondance sur le critère suivant : si une chaîne de la protéine d'un complexe partage au moins 30 % d'identité de séquence avec une chaîne

protéique d'un autre complexe, alors ces derniers sont regroupés dans un même cluster. La structure de meilleure résolution est choisie comme représentative du cluster. A la date du 21 octobre 2017, cette sélection a conduit à 188 structures : 122 complexes protéine-A, 18 protéine-C, 21 protéine-G et 27 protéine-U.

Bien que cela ne soit pas présenté dans le manuscrit, l'un des objectifs ayant conduit à la construction de ce jeu de données concerne l'évaluation de fonctions de score dans leur capacité à discriminer une base sélectivement reconnue à un site de liaison. Les 188 structures résultantes ont donc été manuellement nettoyées sur le critère suivant : la préférence d'un site d'interaction à lier un nucléotide donné doit pouvoir être inférée, prioritairement à partir de données de la littérature. Lorsque aucune donnée littéraire n'a pu être trouvée, la préférence nucléotidique a été inférée à partir de la fonction de la protéine (par exemple un C sera considéré comme le nucléotide privilégié au site d'interaction d'une CMP-kinase), ou de la réaction catalysée par la protéine (le nucléotide doit être un substrat ou un produit de la réaction). A l'issue de ce filtrage, 56 complexes ont été éliminés, conduisant à conserver 131 structures.

L'élimination de la redondance basée sur l'identité de séquence ayant été réalisée séparément sur chaque ensemble de structures liées à un type de nucléotide donné, de la similarité peut persister entre deux structures liant un nucléotide différent. Par ailleurs, un seuil de 30 % d'identité de séquence ne garantit pas que les résidus constituant le site d'interaction ne soient pas parfaitement identiques entre deux complexes. Afin d'éliminer cette potentielle redondance résiduelle entre complexes liés à un nucléotide différent, toutes les structures ont été superposées entre elles avec le programme TM-align (Berman et al., 2000). Les sites d'interaction de toutes les structures superposées avec un TM-score  $\geq 0,8$  ont alors été inspectées visuellement de manière à éliminer les sites d'interaction parfaitement identiques. Deux sites d'interaction ont été considérés comme dissemblables s'ils ne différaient que d'un seul acide aminé parmi ceux directement en contact avec le nucléotide. Cette inspection n'a conduit à éliminer qu'une seule structure : 3DXG (qui lie un U), parfaitement identique à 3DJX (qui lie un C). Cette dernière a été choisie de manière à bénéficier d'une structure supplémentaire liant un nucléotide C et ainsi répartir au mieux le nombre de structures par type de nucléotide lié. Au final, ce sont donc 130 complexes non-redondants qui constituent le jeu de données. Ces 130 complexes sont présentées dans la partie "Annexes – section 5".

Après analyses post-docking, 10 complexes ont été retirés en raison de problèmes décrits dans la partie "Annexes – section 6". Les résultats de docking portent donc sur des analyses réalisées sur 120 complexes protéine-nucléotide dont voici la liste des codes PDB : 1EX7, 1HDI, 1IYB, 1JP4,

1KTG, 1NH8, 1QF9, 1QGX, 1RAO, 1S68, 1UA4, 1UCD, 1UJ2, 1UUY, 1WXI, 1XTT, 1Y1P, 1Z4M, 2A7X, 2CNQ, 2EQA, 2FFC, 2FJB, 2G1U, 2GXQ, 2II6, 2J91, 2JB7, 2JBH, 2OUN, 2QRK, 2R85, 2UV4, 2VFK, 2XBU, 2XWM, 2YAB, 2YRX, 2YVO, 3AKE, 3C85, 3CJ9, 3CLS, 3DDJ, 3DJX, 3DLZ, 3EWY, 3FEG, 3FWZ, 3G1Z, 3GLV, 3GRU, 3IB8, 3KD6, 3KGD, 3LFR, 3LKM, 3M84, 3N1S, 3NUA, 3NYQ, 3O0M, 3OMF, 3PLN, 3RL4, 3RPZ, 3SF0, 3TTF, 3UQ8, 3UWQ, 3W07, 4BLW, 4BRQ, 4CO4, 4CS3, 4D05, 4D7A, 4EEI, 4EMD, 4EQL, 4EUM, 4FBC, 4FE3, 4G0P, 4H2W, 4HE2, 4IG1, 4IJN, 4IKE, 4JEM, 4KBF, 4M0K, 4M9D, 4MA0, 4MPO, 4MX2, 4NDF, 4O6M, 4OZL, 4P86, 4PNO, 4R78, 4UUW, 4WW7, 4X9D, 4ZCP, 4ZFN, 5B6D, 5B8F, 5BPH, 5COT, 5D4N, 5ED3, 5GMD, 5JDA, 5K1D, 5T8S, 5V0I, 5V1M, 5X0J.

## **2.2 Simulations de docking**

Pour chaque protéine du jeu de données, les cinq structures de ligand suivantes ont été dockées : RNXN010, RNXN110, RNXN210, RNXN310 et RNXN410, où N correspond à la base du nucléotide lié dans la structure expérimentale. Les structures sont décrites dans la partie "Introduction" de ce chapitre et la nomenclature MCSS qui s'y réfère est précisée dans la partie "Méthodes générales – section 1.3". Toutes ces structures ont été dockées à l'intérieur d'une boîte parallélépipédique englobant le site d'interaction des protéines. Cette boîte est centrée sur le ligand et présente un volume de  $17 \text{ \AA}^3$  (Fig. 46), volume minimal pour amarrer les cinq structures nucléotidiques utilisées comme ligand. La préparation des protéines, les paramètres utilisés pour les simulations et l'évaluation énergétique des poses sont tels qu'indiqué au chapitre II.

### 3 Résultats

#### 3.1 Echantillonnage et scoring des cinq structures de ligand

Tous les calculs de docking ont été réalisés à partir de la forme liée des protéines issues de 120 complexes protéine-nucléotide à haute résolution ( $\leq 2 \text{ \AA}$ ) et non-redondants. Les calculs de docking sont restreints au site d'interaction des protéines. Pour chaque protéine, les cinq structures de ligand testées ont été dockées indépendamment à l'intérieur d'une boîte englobant le site d'interaction (Fig. 46). Les paramètres utilisés pour définir la boîte et pour les calculs de docking sont données dans la partie "Matériels et méthodes, 2.2".

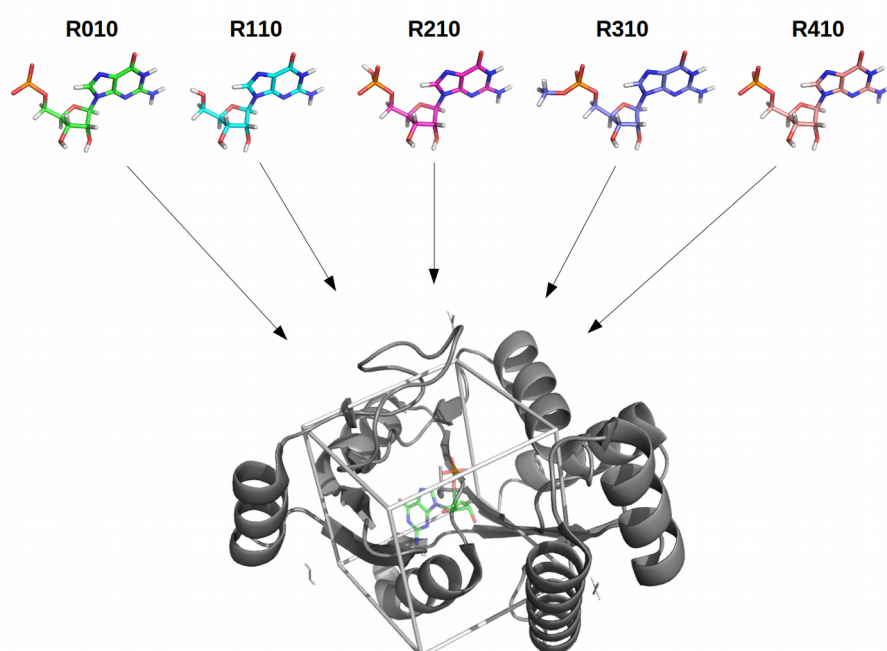


Figure 46 : Illustration de la procédure de docking. Les cinq structures nucléotidiques sont dockées dans un espace défini par une boîte centrée sur le ligand natif (représenté en bâtonnet et transparence). Le volume de cette boîte est de  $17 \text{ \AA}^3$ .

##### 3.1.1 Évaluations faites sur l'ensemble des poses générées à l'issue du docking

Les figures 47A et 47B résument les résultats de l'échantillonnage à partir de l'ensemble des poses issues du docking réalisé sur les 120 complexes. Les structures du ligand nucléotidique qui contiennent un groupement phosphate (010, 210, 310 et 410) génèrent en moyenne autour de 3000 poses au total (Fig. 47A). La structure 110 montre en revanche un nombre total de poses notablement inférieur ( $\sim 2000$ ). Cette dernière porte une charge globale très proche de la structure



010, et est la seule des cinq structures à ne pas porter de groupement phosphate. L'absence de ce dernier pourrait donc expliquer la différence observée entre le nombre de poses générées à partir de la structure 010 et celui obtenu à partir des autres structures.

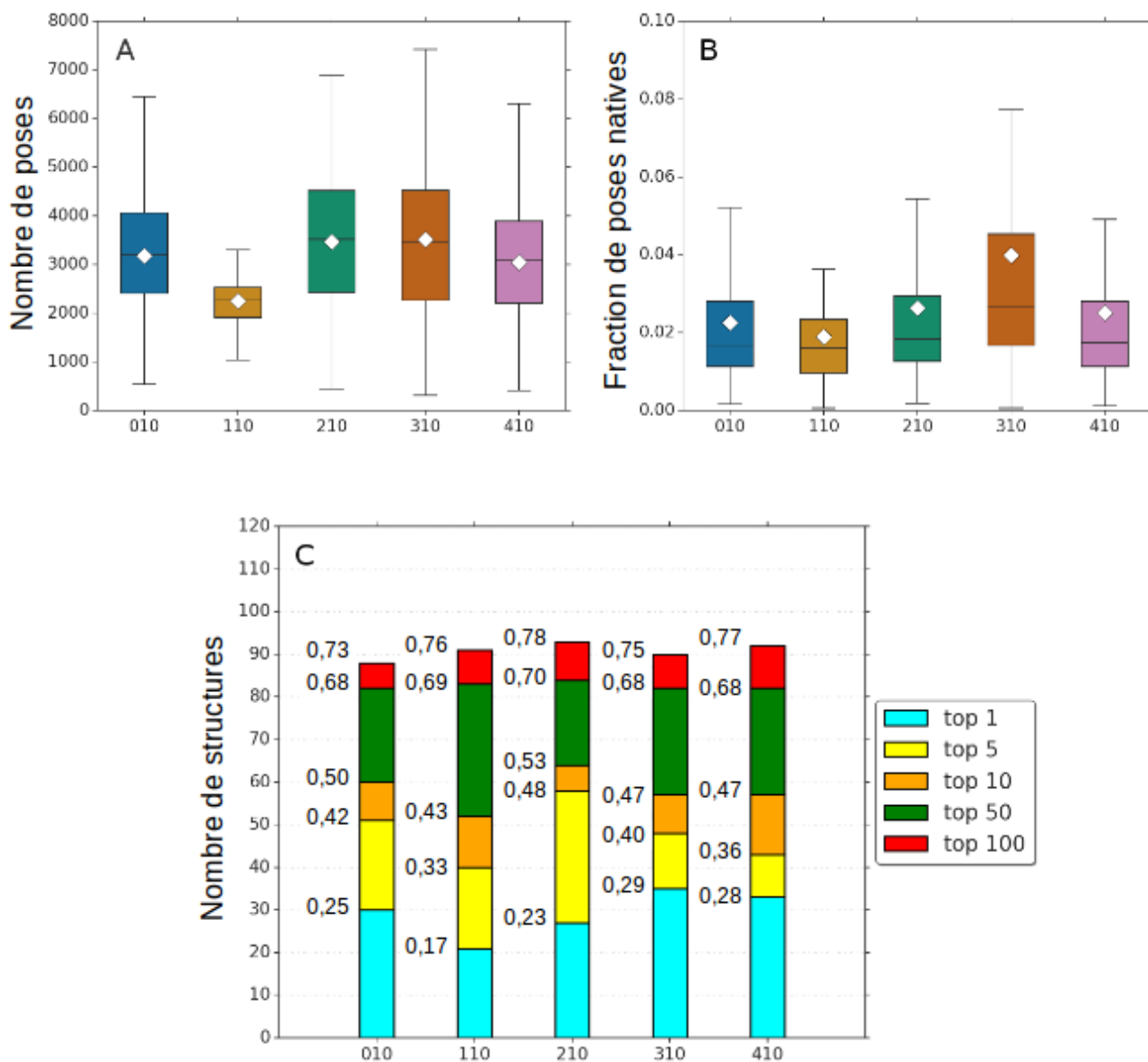


Figure 47: Comparaison des résultats de docking entre les cinq structures de ligand testées. A - Distribution sous forme de boîtes à moustaches des nombre de poses totales générées pour chaque structure. B - Distribution sous forme de boîtes à moustaches des proportions de poses natives obtenues sur la totalité des poses générées pour chaque structure. Pour les boîtes à moustache, la boîte représente les premier et dernier quartiles, la ligne noire la médiane, le losange blanc la moyenne ; les moustaches s'étendent jusqu'aux valeurs les plus extrêmes dans la limite de 1.5 fois la hauteur de la boîte. C - Evaluation du nombre de structures pour lesquelles une pose native est classée dans les tops 1, 5, 10, 50 ou 100. L'évaluation est faite sur l'ensemble des poses issues du docking des différentes structures de ligand testées ; le taux de succès rapporté sur 120 protéines est indiqué à côté de chaque histogramme.

Malgré un nombre total de poses générées inférieur pour la structure 110, la proportion de poses natives obtenues pour cette dernière ne diffère pas significativement (test de la somme des rangs de Wilcoxon – analyse non présentée) de celles observées pour les structures 010, 210 et 410 (Fig. 47B). Seule la structure 310 présente une proportion moyenne de poses natives significativement (test de la somme des rangs de Wilcoxon - analyse non présentée) supérieure aux autres : environ 4 % de poses natives générées en moyenne sur les 120 protéines utilisées contre environ 2 % pour les autres structures.

La discrimination des poses natives parmi l'ensemble des poses générées a ensuite été évaluée pour les différentes structures de ligand testées. L'évaluation a été faite en dénombrant le nombre de protéines pour lesquelles une pose native est classée aux rangs (ou tops) 1, 5, 10, 50 ou 100. Pour chaque top, le taux de succès est calculé en rapportant ce nombre au 120 protéines utilisées comme jeu de données ; précisons que pour ces 120 protéines, au moins une pose native a été générées au cours du docking pour les cinq structures de ligand testées (tableau 19).

La figure 47C montre le nombre de complexes pour lesquels une pose native est classée aux rangs (ou tops) 1, 5, 10, 50 ou 100. Aucune différence nette ne ressort entre les structures pour les tops 50 et 100. la structure 110 montre en revanche un nombre de succès globalement inférieur à celui observé pour les autres structures pour les tops 1, 5 et 10. Il est ensuite difficile d'établir une hiérarchie pour les structures 010, 210, 310 et 410. La structure 210 montre un nombre de succès plus important aux tops 5 et 10, mais il est inférieur à celui des autres structures pour le top 1. Globalement, aucune différence nette n'apparaît dans la discrimination des poses natives obtenues à partir des structures contenant un groupement phosphate (010, 210, 310, et 410). Seule la structure de ligand 010 se distingue par des taux de succès moins importants.

### **3.1.2 Evaluations faites après élimination de la redondance parmi les poses issues du docking**

Les évaluations présentées ci-dessus ont toutes été réalisées à partir de l'ensemble des poses générées durant le docking. Comme on l'a vu dans le chapitre III (section 3.2.2), il existe de la redondance entre les poses générées. Cette redondance conduit à augmenter le rang global des poses natives. Les évaluations de la section précédente ont donc été répétées après élimination de cette redondance. Pour cela, un clustering à 2 Å a été appliqué sur l'ensemble des poses (voir chapitre II, section 3 pour la procédure). Pour chaque cluster engendré, la pose de meilleure énergie d'interaction est choisie comme pose représentative. Ces dernières sont finalement rangées par ordre décroissant de leur énergie, c'est-à-dire de la plus favorable à la moins favorable. La figure 48 montre que le nombre moyen de poses représentatives obtenues à l'issue du clustering est autour de

250 pour les cinq structures, réduisant de près d'un facteur 10 le nombre total de poses issues du docking (Fig. 47A).

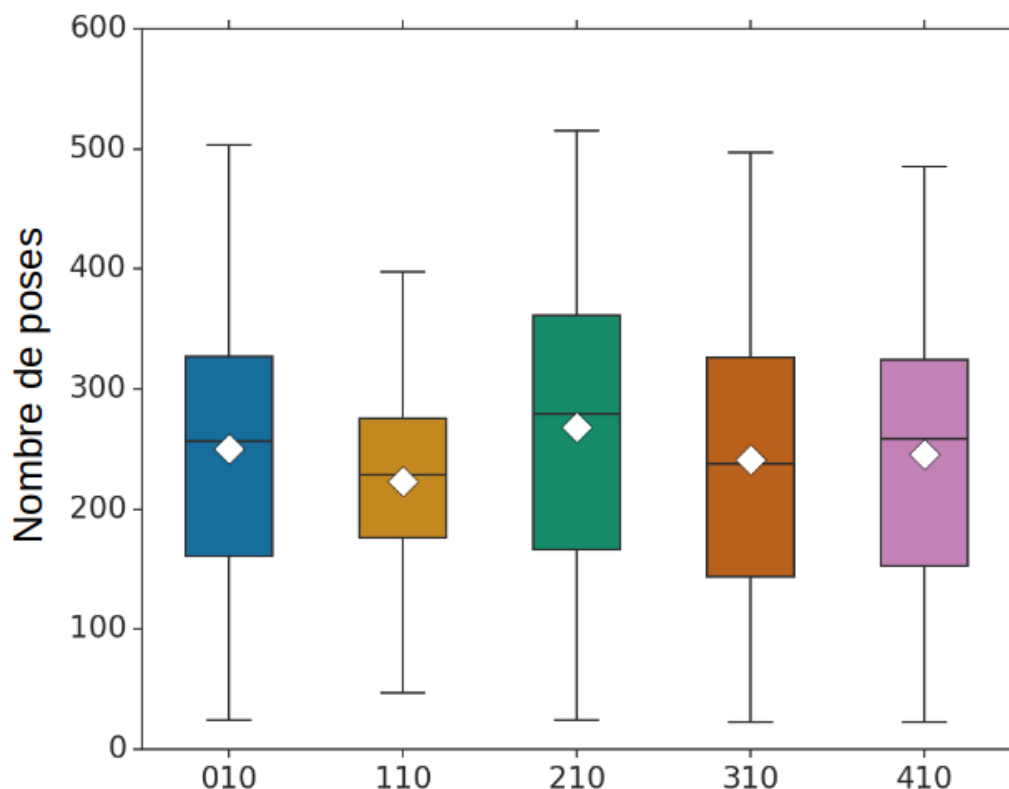


Figure 48: Distribution sous forme de boîtes à moustaches du nombre de poses issues d'un clustering à 2 Å pour chaque structure de ligand testée. La boîte représente les premier et dernier quartiles, la ligne noire la médiane, le losange blanc la moyenne ; les moustaches s'étendent jusqu'aux valeurs les plus extrêmes dans la limite de 1.5 fois la hauteur de la boîte.

Un clustering doit être sensible et spécifique : idéalement, toutes les poses natives doivent être regroupées dans un même cluster, et ce cluster ne devrait contenir aucune poses non-natives. La procédure de regroupement utilisée, le seuil de clustering (2 Å), ainsi que les données en elles-mêmes font que ce n'est évidemment pas le cas : les poses natives peuvent se répartir au sein de plusieurs groupes, chacun d'eux pouvant être constitués de poses non-natives. Dans ces conditions, il n'est pas garanti que la pose de plus basse énergie d'interaction sélectionnée comme représentative d'un cluster corresponde à une pose native. Des poses natives peuvent donc être "perdues" après la procédure de clustering. Le tableau 19 montre que la procédure de clustering réalisée à 2 Å conduit globalement à peu de perte de poses natives : sur les 120 protéines contenant au moins une pose native sans étape de clustering, entre 110 (structure du ligand 010) et 117 (structure du ligand 210) protéines contiennent toujours au moins une pose native. La procédure de clustering utilisée

identifie en premier lieu la pose ayant le plus de voisins parmi l'ensemble des poses issues du docking pour générer le premier cluster. La procédure est ensuite itérativement répétée pour les autres poses. La composition de tous les clusters résultants est donc dépendante de la pose initialement identifiée comme celle ayant le plus de voisins. La figure 47-B montre que la structure du nucléotide 310 présente un enrichissement en poses natives par rapport au quatre autres structures de ligand testées. On pourrait s'attendre à ce qu'une proportion plus importante de poses natives apporte un avantage pour le regroupement des poses natives entre elles en limitant le risque que ces dernières soient regroupées avec des poses non-natives. Si tel était le cas, le risque que l'on perde une pose native pour une protéine donnée suite à la procédure de clustering serait également moindre. Le tableau 19 montre pour la structure du ligand 310 que nous perdons une pose native pour cinq protéines suite au clustering, alors que seulement qu'une pose native est perdue pour seulement quatre protéines avec la structure 110 qui présente en moyenne la plus faible proportion de poses natives (figure 47-B). Cela montre que l'enrichissement en poses natives n'apporte pas d'avantage particulier dans la qualité du partitionnement lors de la procédure de clustering utilisée.

*Tableau 19: Nombre de structures pour lesquelles au moins une pose native a été générée pour chaque structure de ligand testée. Le dénombrement a été opéré sur l'ensemble des poses générées durant le docking (sans clustering) ou bien sur les poses représentatives issues d'un clustering à 2 Å (après clustering). Une pose représentative d'un cluster est la pose de plus basse énergie trouvée dans ce cluster.*

	<b>010</b>	<b>110</b>	<b>210</b>	<b>310</b>	<b>410</b>
<b>Sans clustering</b>	120	120	120	120	120
<b>Après clustering</b>	110	116	117	115	115

La comparaison des performances de scoring mesurées à partir de l'ensemble des poses initiales ou des poses représentatives issues du clustering à 2 Å est représentée à la figure 49 pour chaque structure de ligand testée. Pour effectuer une comparaison pertinente et prendre en considération l'approche de clustering dans sa globalité, tous les taux de succès indiqués se rapportent aux 120 protéines. Pour les cinq structures de ligand, le clustering n'apporte pas de modifications de tendances dans les taux de succès (histogramme B, Fig. 49) par rapport à celles observées à partir de l'ensemble des poses issues du docking (histogramme A, Fig. 49). En effet, après clustering, l'utilisation de la structure 110 pour le docking entraîne toujours un nombre de succès dans les tops 1, 5 et 10 qui est inférieur à celui observé pour les structures de ligand 010, 210, 310 et 410. Pour ces dernières, il est également toujours compliqué d'affirmer que l'une d'elles conduit à une

meilleure discrimination des poses natives. On peut néanmoins remarquer que pour la structure du ligand 010, comme dit plus haut, 110 protéines possèdent au moins une pose native après clustering tandis qu'au moins 115 protéines ont une pose native pour les autres structures de ligand (tableau 19). Ainsi, si les taux de succès indiqués à la figure 49 étaient rapportés sur le nombre de protéines ayant au moins une pose native pour chaque structure de ligand, les taux de succès pour la structure du ligand 010 apparaîtraient, pour chaque top évalué (tops 1, 5, 10, 50 et 100), légèrement supérieurs à ceux des autres structures de ligand testées. Cependant, comme dit précédemment, les taux de succès indiqués se rapportent aux 120 protéines du jeu de données de manière à ce que l'évaluation du scoring considère l'approche de clustering comme partie intégrante de la procédure de docking. Les figures 49 et 38 montrent que malgré un nombre légèrement moins important de protéines contenant au moins une pose native après clustering, la procédure de regroupement des poses avec un seuil RMSD de 2 Å conduit à une augmentation importante des taux de succès aux tops 5, 10, 50 et 100, quelle que soit la structure du ligand testée. Cela est le reflet de l'élimination de la redondance des poses qui conduit à réduire le nombre total de poses (Fig. 48) et ainsi à abaisser le rang global des poses natives retenues.

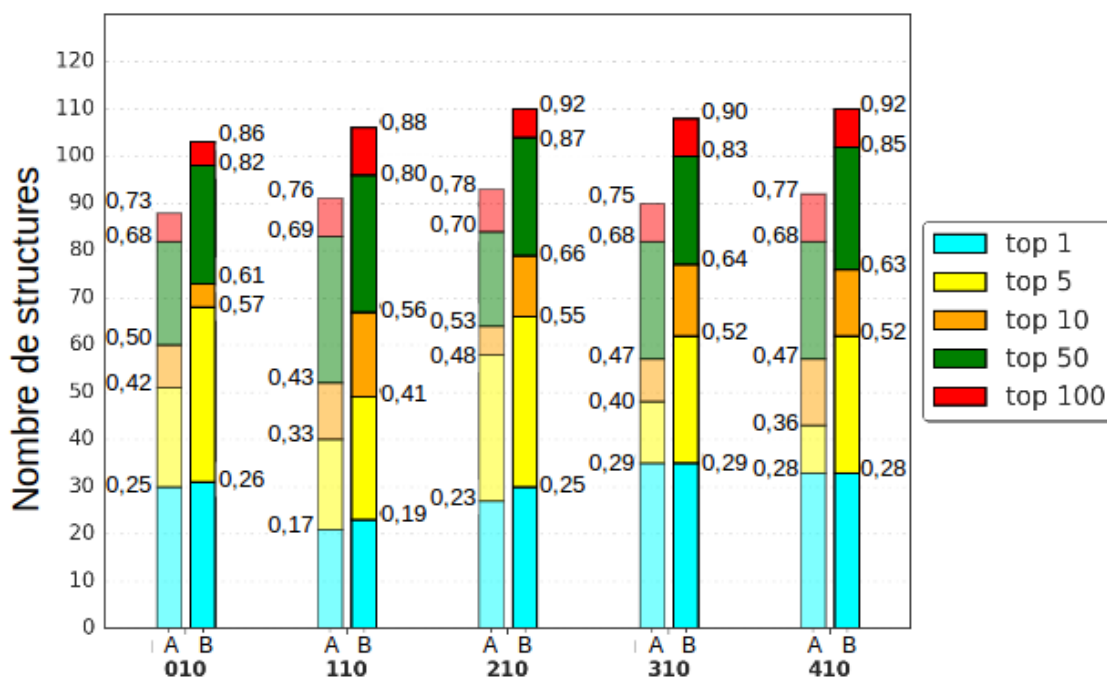


Figure 49: Comparaison de l'influence de différentes structures de ligand sur la capacité de la fonction de score MCSS à discriminer une pose native. Chaque histogramme représente le nombre de structures pour lesquelles une pose native est classée dans les tops 1, 5, 10, 50 ou 100. Les résultats sont comparés à partir de poses clusterisées (B) ou non (A). Les taux de succès, rapportés à 120 structures, sont indiqués.

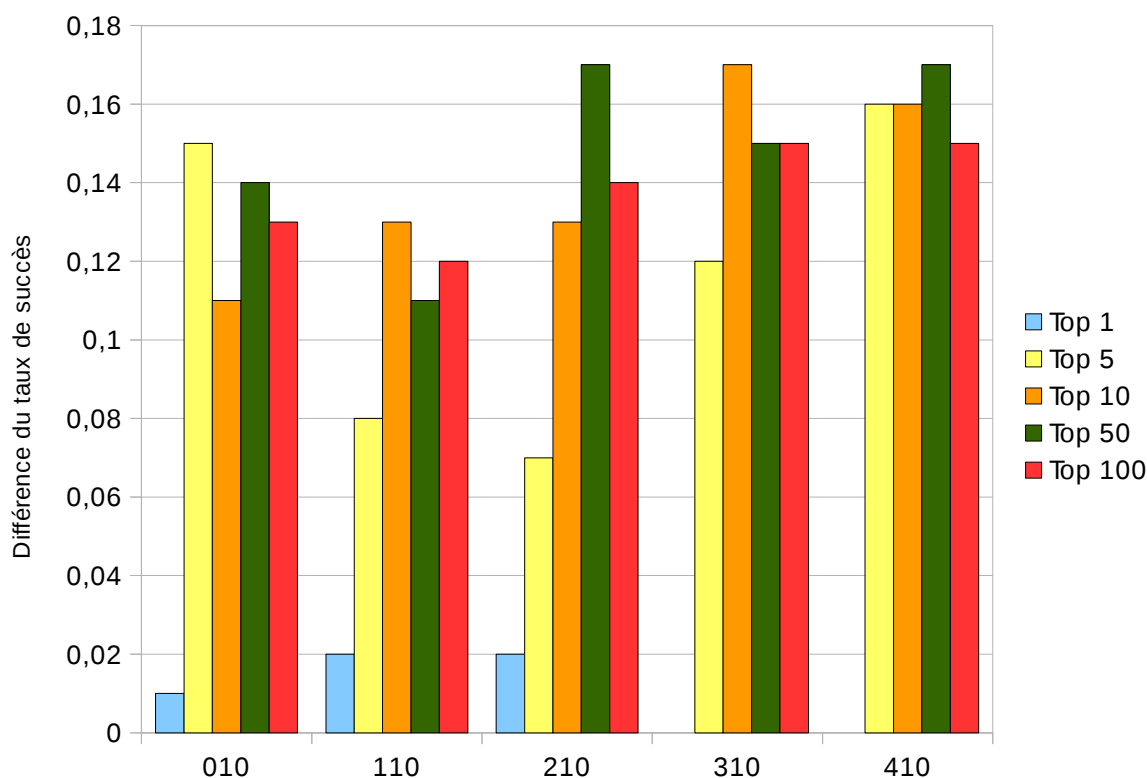


Figure 50: Comparaison pour chaque structure de ligand testée du gain apporté par un clustering à 2 Å sur les taux de succès pour la discrimination des poses natives. Pour chaque top (1, 5, 10, 50 et 100), le taux de succès observé pour l'ensemble des poses issues du docking (Fig. 50, histogrammes A) est soustrait au taux de succès observé pour les poses représentatives issues du clustering à 2 Å (Fig. 50 – histogrammes B)

## 4 Discussion

Différentes structures de ligand nucléotidique sont implémentées dans MCSS. Jusqu'ici, seule la structure 010 a été systématiquement utilisée par défaut dans l'approche FBDRNA pour la modélisation des interactions protéine-ARNsb. Les performances de l'approche FBDRNA sont en premier lieu dépendantes des performances de docking avec MCSS. L'identification d'une structure de ligand nucléotidique implémentée dans MCSS conduisant à de meilleures performances de docking pourrait donc être bénéfique à l'approche FBDRNA en facilitant la sélection de poses d'intérêt pour la construction de modèles de chaînes ARNsb.

Dans l'étude présentée dans ce chapitre, cinq structures différentes de ligand nucléotidique ont été comparées dans leur performance de docking sur un jeu de données non-redondant composé de 120 complexes protéine-nucléotide à haute résolution ( $\leq 2$  Å). En évaluant la capacité de la fonction de score MCSS à discriminer des poses natives générées à partir de chaque structure de ligand

testée, nous n'avons identifié aucune structure conduisant à une meilleure discrimination par rapport à la structure 010 utilisée par défaut. Parmi les cinq structures de ligand testées, la structure du ligand 110 est celle qui présente les moins bonnes performances pour la discrimination des poses natives. Cette structure se distingue des quatre autres structures testées par l'absence de groupement phosphate à l'extrémité O5' du ribose. Les structures 010, 210, 310 et 410 montrent elles des performances comparables malgré qu'elles diffèrent au niveau des atomes composant leur groupement phosphate et au niveau de leur charge globale. Ces résultats suggèrent que le groupement phosphate à l'extrémité O5' du ribose présente, sur le jeu de données utilisé, une contribution à l'énergie d'interaction importante pour la discrimination des poses natives, indépendamment de sa composition et de sa charge.

L'approche FBDRNA a montré à partir de la structure 010 que l'élimination de la redondance des poses issues du docking par une étape de clustering à 2 Å permet de réduire le nombre de poses à sélectionner pour retenir les poses natives d'intérêt. Nous avons donc également évalué l'apport d'un clustering à 2 Å dans l'identification des poses natives générées par chacune des structures de ligand testées sur les 120 protéines du jeu de données. Après clustering, la comparaison des performances obtenues pour la discrimination des poses natives montrent une tendance similaire à celle observée avant clustering : les structures 010, 210, 310 et 410 montrent des performances équivalentes entre elles et supérieures à celle observée pour la structure 110. Les résultats montrent également que la procédure de clustering à 2 Å entraîne une augmentation conséquente des performances dans la discrimination des poses natives pour les cinq structures du ligand utilisées pour le docking. Globalement, ces résultats supportent les observations faites à partir du jeu de données constitué de trois RBDs utilisés pour le développement de l'approche FBDRNA. Ils confirment que l'élimination de la redondance des poses issues du docking par un clustering à 2 Å permet de faciliter l'identification des poses natives, et montrent de plus que le gain apporté est indépendant de la structure du ligand utilisée.

L'évaluation de l'échantillonnage a montré qu'un docking réalisé avec la structure 310 conduit à une proportion moyenne de poses natives plus importante par rapport aux autres structures de ligand. La procédure de clustering utilisée peut conduire à perdre des poses natives si ces dernières ne sont pas retenues comme pose représentative du cluster, c'est-à-dire la pose de plus basse énergie. Puisque la proportion en poses natives obtenues est plus grande lorsque la structure 310 est dockée, on aurait pu s'attendre à ce que le risque de perdre des poses natives soit moins important comparé aux autres structures de ligand testées. Cependant, les résultats ne montrent aucune corrélation apparente entre la structure du ligand utilisée et le nombre de protéines ayant perdu une

pose native après clustering (tableau 19). L'enrichissement en poses natives obtenues à partir de la structure 310 ne montre donc aucun avantage particulier par rapport aux autres structures de ligand. La structure 310 se distingue des autres structures de ligand testées par la présence d'un groupement méthyle attaché à l'un des oxygènes du groupement phosphate. Plusieurs études ont mis en évidence que l'ajout d'un groupement méthyle à un ligand peut conduire à augmenter son affinité de liaison envers un récepteur si ce groupement méthyle se lie dans un environnement hydrophobe (A. M. Davis & Teague, 1999; Hajduk & Sauer, 2008; Leung, Leung, Tirado-Rives, & Jorgensen, 2012). Cependant, compte tenu de la charge négative portée par un groupement phosphate, il est peu probable, pour les 120 protéines du jeu de données, que l'environnement où se fixe le groupement méthyle des poses natives soit hydrophobe. Par ailleurs, l'enrichissement en poses natives observé pour la structure 310 pourrait être le reflet des atomes considérés pour les calculs RMSD, aussi bien ceux effectués au cours de la procédure de docking pour évaluer les poses convergentes, que ceux réalisés pour l'analyse des poses par rapport au ligand expérimental. Dans les deux cas, le RMSD est calculé sur tous les atomes lourds (à l'exception des atomes d'oxygène du groupement phosphate) communs aux nucléotides dockés et au ligand expérimental. L'atome de carbone du groupement méthyle n'est donc pas considéré pour les calculs RMSD. En conséquence, deux poses issues de la structure 310 pourront présenter un RMSD parfaitement identique alors que le positionnement de l'atome de carbone de leur groupement méthyle sera différent. Des analyses complémentaires doivent donc être réalisées afin de déterminer si l'enrichissement en poses natives observé pour la structure 310 résulte d'un effet réel des propriétés physico-chimiques du groupement méthyle ou bien s'il résulte d'un biais de mesures RMSD.

Globalement, l'ensemble des résultats de cette étude suggèrent que parmi les structures 110, 210, 310 et 410, aucune n'est susceptible d'apporter une amélioration substantielle pour l'approche FBDRNA dans la sélection des poses d'intérêt par rapport aux résultats obtenus avec la structure 010. L'approche FBDRNA se destine à être appliquée à des RBDs. Dans cette étude, les résultats proviennent d'un jeu de données composé de protéines liant des nucléotides. Le mode de liaison des nucléotides à leur récepteur s'apparente à celui observé pour les nucléotides d'une chaîne ARNsb sans élément de structures secondaires dans sa forme liée à un RBD, notamment au niveau des bases qui établissent des contacts directs avec la protéine pour être sélectivement reconnue. Nous pensons donc que les observations faites dans cette étude à partir des 120 complexes protéine-nucléotide peuvent être transférables à des sites d'interaction nucléotidique de RBDs. Cette étude pourrait être complétée par la comparaison d'autres structures de ligand nucléotidique implémentées dans MCSS qui n'ont pu être testées faute de temps.



## VI Comparaison de cinq fonctions de score dans leur capacité à discriminer les poses natives

### 1 Introduction

L'approche FBDRNA a pour objectif de prédire le mode d'interaction de courtes chaînes ARNs liées à des RBDs sans élément de structures secondaires, soit à partir de la connaissance de la séquence ARN, soit sans *a priori* sur cette dernière. Dans les deux cas, l'application effective de l'approche passe par une réduction de la complexité pour réduire la combinatoire et parvenir à générer des modèles de chaînes ARN en quantité raisonnable. Les travaux présentés aux chapitres III et IV portent sur la mise au point de protocoles de sélection des poses d'intérêt dans le but de réduire cette combinatoire. Que la séquence ARN soit donnée ou non en entrée pour l'approche FBDRNA, le protocole de sélection passe par une procédure référée par le terme "diviser pour mieux régner". Globalement, cette procédure implique une première étape de sélection basée sur l'accumulation de poses autour de chaque site d'interaction nucléotidique, et une deuxième étape de sélection basée sur l'énergie d'interaction des poses. Alors que l'efficacité de la première étape de sélection est essentiellement dépendante de la procédure d'échantillonnage utilisée par le programme de docking MCSS, l'efficacité de la seconde étape de sélection est elle dépendante de la capacité de la fonction de score à discriminer les poses natives d'intérêt. La fonction de score utilisée dans MCSS a montré des performances de discrimination globalement intéressante, mais les résultats ne concernent qu'un jeu de données limité à neuf sites d'interaction nucléotidiques sur trois RBDs. Les travaux présentés dans ce chapitre ont été réalisés avec l'objectif d'avoir une vision plus représentative des performances de la fonction de score MCSS, d'une part par l'utilisation d'un jeu de données plus important, et d'autre part en la comparant à d'autres fonctions de score.

Pour répondre à cet objectif, les 120 protéines liant des nucléotides présentées précédemment dans le chapitre V ainsi que les poses non-redondantes générées à partir du docking de la structure 310 ont été utilisées comme jeu de données. Quatre fonctions de score ont été comparées à MCSS dans leur capacité à discriminer les poses natives : les fonctions de score Vina (Trott & Olson, 2010), Vinardo (Quiroga & Villarreal, 2016), DVRF20 (C. Wang & Zhang, 2017), et ITscorePR (S.-Y. Huang & Zou, 2014a). Ces fonctions de score ont été choisies en raison des bonnes performances qu'elles ont affichées pour la discrimination du mode d'interaction natif sur des jeux de données variés. Par exemple, parmi 25 fonctions de score testées, Vina et DVRF20 ont montré les meilleures performances pour identifier dans les tops 1, 2 et 3 la pose native de 285 complexes protéine-ligand constituant le jeu de données de la PDBbind version 2016 (Su et al., 2019). Précédemment, Vinardo

avait montré des performances prédictives supérieures à Vina sur 195 complexes protéine-ligand du jeu de données antérieur de la PDBbind version 2013 (Quiroga & Villarreal, 2016). La fonction de score ITscorePR a elle montré des performances systématiquement supérieures à cinq autres fonctions de score pour discriminer le mode d'interaction natif de chaînes ARN sur quatre jeu de données protéine-ARN différents (S.-Y. Huang & Zou, 2014a).

## **2 Matériels et méthodes**

### **2.1 Données**

La comparaison des performances prédictives des différentes fonctions de score a été effectuée à partir des 120 protéines présentées précédemment dans le chapitre V. Les poses générées à partir de la structure du ligand 310 ont été utilisées (voir ici pour les calculs de docking : 2.2). Le choix de cette structure a été orientée par le fait qu'elle conduit à générer une plus grande proportion de poses natives. Bien que cette caractéristique n'a montré aucun apport particulier, nous pensions au moment de la réalisation de ces travaux qu'elle pourrait être bénéfique à l'approche FBDRNA. Seules les poses retenues après un clustering à 2 Å ont été utilisées.

### **2.2 Fonctions de scores comparées**

Pour chacune des poses retenues après clustering à 2 Å, l'énergie d'interaction a été ré-évaluée par quatre fonctions de score : Vina (Trott & Olson, 2010), Vinardo (Quiroga & Villarreal, 2016), DVRF20 (C. Wang & Zhang, 2017), et ITscorePR (Huang & Zou, 2014).

#### **2.2.1 Vina**

La fonction de score Vina peut être classée parmi les fonctions de score empiriques. Elle comprend par défaut cinq termes : trois termes pour décrire les interactions stériques, et deux termes dépendants de la paire d'atomes considérée utilisés pour décrire respectivement les interactions hydrophobe et hydrogène en fonction de la distance séparant les atomes. A chacun de ces termes est affecté un coefficient qui a été ajusté pour reproduire au mieux les affinités de liaison de complexes protéine-ligand de la base de données PDBbind version 2007 (R. Wang, Fang, Lu, & Wang, 2004).

Le programme Smina <http://smina.sf.net/> a été utilisé pour la ré-évaluation de l'énergie d'interaction des poses par Vina et Vinardo (voir ci-dessous). Smina est une version dérivée d'Autodock Vina optimisée pour supporter l'évaluation énergétique de poses à grande échelle et l'utilisation de fonctions de score personnelles (Koes, Baumgartner, & Camacho, 2013). La commande suivante a été appliquée pour estimer l'énergie entre chaque pose et un récepteur par Vina :

- `smina.static -r recepteur.pdbqt -l pose.pdbqt --score_only --scoring vina`

### 2.2.2 Vinardo

La fonction de score Vinardo dérive directement de la fonction utilisée par défaut dans Vina et peut donc aussi être définie comme une fonction de score empirique. Vinardo inclut quatre termes de Vina, l'un des termes stériques ayant été retiré. Elle diffère de Vina dans les coefficients affectés à chaque terme et aussi dans la définition du rayon de van der Waals des atomes. Ces modifications ont été obtenues à partir de différentes formules perturbées de la fonction Vina et évaluées sur 122 structures de la PDBbind version 2013 (Y. Li, Han, Liu, & Wang, 2014).

Le programme Smina a été utilisé pour la ré-évaluation de l'énergie d'interaction des poses et la commande suivante a été appliquée pour estimer l'énergie entre chaque pose et un récepteur par Vinardo :

- `smina.static -r recepteur.pdbqt -l pose.pdbqt --score_only --scoring vinardo`

### 2.2.3 DVRF20

La fonction de score DVRF20 est basée sur un apprentissage automatique reposant sur des forêts d'arbres décisionnels. Elle comprend 20 descripteurs, dont 10 proviennent du code source d'AutoDock Vina (<https://github.com/ProzacR/vina>), et 10 autres se rapportent à l'accessibilité au solvant par type d'atomes. L'apprentissage de DVRF20 a été effectué à partir de données expérimentales issues de la PDBbind version 2014 et d'un ensemble de poses générées artificiellement faisant partie des données CSAR (Dunbar et al., 2011).

La fonction de score DVRF20 a été téléchargée à l'adresse suivante : <https://www.nyu.edu/projects/yzhang/DeltaVina>. La commande ci-dessous a été appliquée pour estimer l'énergie entre chaque pose et un récepteur :

- `dvrf20.py recepteur.pdb pose.pdb`

### 2.2.4 ITscorePR

La fonction ITscorePR est une fonction de score à potentiel statistique. Pour contourner le problème de la définition d'un état de référence inhérent à ce type de fonction de score, le potentiel a été défini par un processus itératif comparant des distributions de paires d'atomes prédites aux distributions de paires d'atomes observées dans des structures cristallines de 110 complexes protéine-ARN. Le processus s'arrête lorsque le potentiel est capable de discriminer correctement le mode d'interaction natif de chaque complexe parmi un ensemble de 1000 modèles leurs générés artificiellement. ITscorePR se distingue donc des fonctions Vina, Vinardo et DVRF20 par le fait qu'elle n'est pas dédiée à l'estimation de l'énergie d'interaction de complexes protéine-ligand.

La fonction de score ITscorePR a été téléchargée à l'adresse suivante : [http://zoulab.dalton.missouri.edu/resources\\_itscorepr.html](http://zoulab.dalton.missouri.edu/resources_itscorepr.html). La commande ci-dessous a été appliquée pour estimer l'énergie entre chaque pose et un récepteur :

- `itscorepr.py recepteur.pdb pose.pdb`

### 2.3 Adaptation du format des fichiers de coordonnées

Les fichiers de coordonnées générés par MCSS sont dans un format pdb étendu. Pour la ré-évaluation de l'énergie d'interaction par les fonctions de score DVRF20 et ITscorePR, ce format a été converti par un script maison dans un format pdb standard.

Le format de fichier reconnu par Smina et nécessaire à la ré-évaluation de l'énergie d'interaction par les fonctions de score Vina et Vinardo est un format pdbqt. Ce format est une extension du format pdb standard qui contient des champs additionnels comme le type d'atome, leur charge partielle, ainsi qu'une information sur les parties rigides distinctes de la molécule permettant de prendre en compte sa flexibilité. Les fichiers au format pdb ont été convertis au format pdbqt à partir de scripts dédiés inclus dans la suite Autodock (Morris et al., 2009) dont l'usage est :

- `prepare_ligand4.py -l pose.pdb -o pose.pdbqt (pour les poses)`
- `prepare_receptor4.py -r recepteur.pdb -o recepteur.pdbqt (pour les récepteurs protéiques)`

## 3 Résultats

Quatre fonctions de score ont été comparées à MCSS dans leur capacité à discriminer les poses natives : les fonctions de score Vina, Vinardo, DVRF20, et ITscorePR. Pour effectuer cette comparaison, le dénombrement de protéines pour lesquelles une pose native est classée dans les tops 1, 5, 10, 50 et 100 a été utilisé pour comparer leurs performances. La figure 51 recense ce nombre de protéines pour chaque fonction de score ainsi que les taux de succès associés. Ces taux correspondent au nombre de protéines ayant une pose native dans un top donné, nombre rapporté à 120 protéines. Les poses ré-évaluées sont les poses non-redondantes obtenues après docking de la structure du ligand 310 et clustering à 2 Å de l'ensemble des poses générées. Après clustering à 2 Å, on a vu dans le chapitre précédent que seules 115 protéines sur les 120 contiennent toujours une pose native (tableau 19). Les 120 protéines ont néanmoins été prises en compte pour déterminer les taux de succès de manière à considérer la ré-évaluation de l'énergie des poses comme une étape faisant partie intégrante d'une stratégie de post-traitement des poses issues du docking.

La figure 51 montre que la fonction de score Vinardo est celle conduisant aux meilleurs taux de succès dans les tops 1, 5 et 10. Elle présente néanmoins des taux de succès identiques à ceux

observés pour MCSS dans les tops 50 et 100. Les fonctions de score MCSS, DVRF20 et Vina montrent des taux de succès globalement équivalents à tous les tops excepté au top 10 pour Vina où le taux de succès apparaît assez nettement inférieur : 0,56 contre 0,64 et 0,63 pour MCSS et DVRF20, respectivement. La fonction de score ITscorePR est celle montrant les moins bonnes performances. Ces performances sont par ailleurs largement inférieures à celles observées pour les autres fonctions de score, et cela à tous les tops considérés.

Globalement, Vinardo, malgré des performances assez proches de celles observées pour MCSS, est la fonction de score montrant les meilleures performances dans la discrimination de poses natives sur le jeu de données protéine-nucléotide utilisé. Elle permet d'envisager une amélioration pour la sélection des poses d'intérêt dans l'approche FBDRNA et donc potentiellement une réduction de la combinatoire pour la recherche de chaînes.

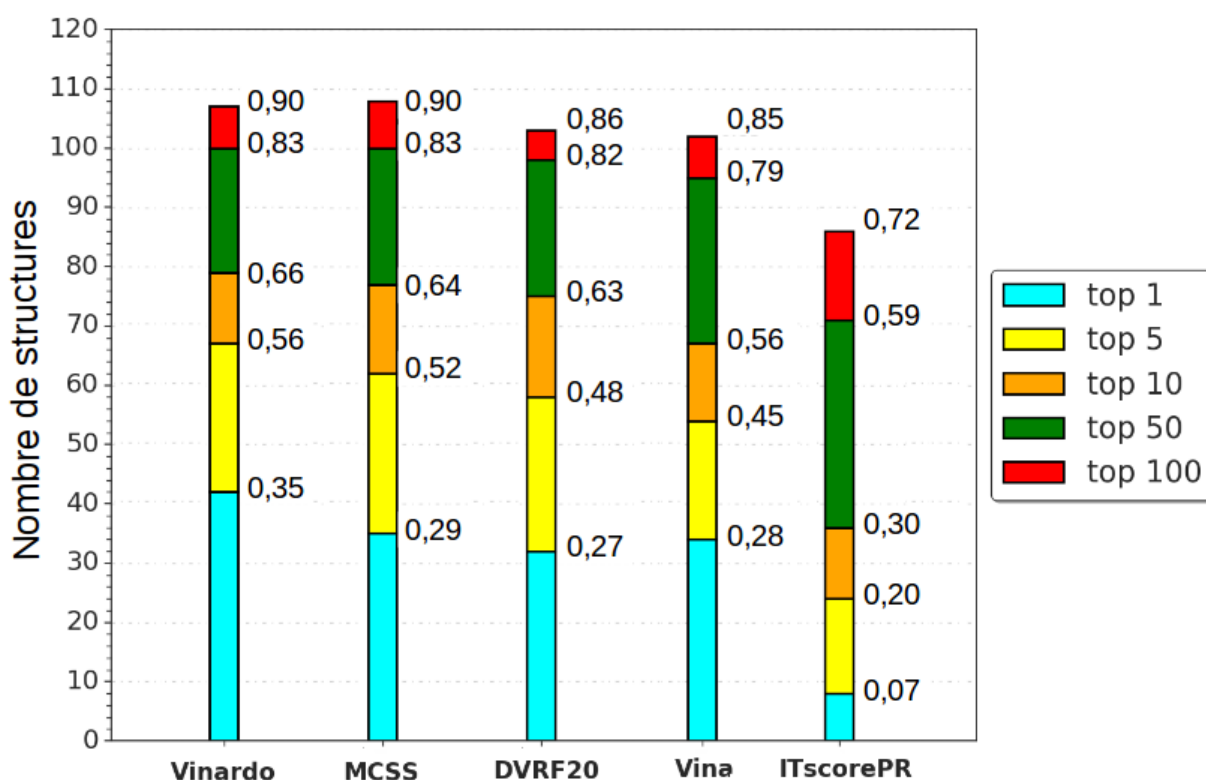


Figure 51: Comparaison des fonctions de score dans leur capacité à discriminer une pose native. Chaque histogramme représente le nombre de structures pour lesquelles une pose native est classée dans les tops 1, 5, 10, 50 ou 100. Le taux de succès rapporté sur 120 protéines est indiqué. Les poses évaluées sont les poses représentatives obtenues après un clustering à 2 Å de l'ensemble des poses issues de calculs de docking réalisés avec la structure 310.

## 4 Discussion

L'application de l'approche FBDRNA impose de retenir un nombre de poses réduit pour limiter la complexité combinatoire inhérente à une recherche de chaînes ARNsb. Une stratégie mise au point précédemment pour réduire l'espace des solutions repose sur une procédure de sélection qui est en partie dépendante des capacités de la fonction de score MCSS à discriminer avec suffisamment de précision les poses natives d'intérêt. La fonction de score MCSS a montré des performances intéressantes mais les observations portent sur un jeu de données limité à seulement neuf sites d'interaction nucléotidiques. Les travaux présentés dans ce chapitre ont été réalisés avec l'objectif d'avoir une vision plus représentative des performances discriminatives de la fonction de score MCSS, d'une part par l'utilisation d'un jeu de données plus important, et d'autre part en la comparant à quatre autres fonctions de score : Vina, Vinardo, DVRF20 et ITscorePR. Les résultats observés sur 120 complexes protéine-nucléotide montrent que la fonction de score Vinardo est celle conduisant aux meilleures performances pour identifier le mode d'interaction natif. Ces performances sont identiques à celles de MCSS pour les tops 50 et 100, et légèrement supérieures pour les tops 1, 5 et 10. MCSS montre par ailleurs des performances comparables à celles de DVRF20, et légèrement supérieures à celles de Vina. ITscorePR se distingue quant à elle assez nettement des autres fonctions de score par ses moindres performances affichées sur ce jeu de données.

L'ensemble des résultats provient de la ré-évaluation de l'énergie d'interaction de poses générées à partir de la structure 310. Cette structure a été choisie sur la base qu'elle conduit à générer une plus grande proportion de poses natives et que nous pensions, au moment de la réalisation de ces travaux, que cette caractéristique pourrait être avantageuse, notamment en augmentant les chances de retenir comme pose représentative une pose native suite à la procédure de clustering à 2 Å. Nous avons cependant montré qu'elle n'apporte aucun avantage particulier dans cette approche de traitement post-docking. Les résultats du chapitre précédent visant à comparer différentes structures de ligand nucléotidique montrent qu'aucune différence nette n'apparaît pour la discrimination des poses natives par la fonction de score MCSS lorsque les poses générées par les structures de ligand 010, 210, 310 et 410. En conséquence, nous supposons que les performances observées pour chacune des fonctions de score testées pour la ré-évaluation de l'énergie d'interaction des poses 310 sont extrapolables à des poses générées à partir de ces autres structures de ligand.

La fonction de score MCSS est une fonction de score relativement simple basée sur un modèle physique. Elle comprend une composante de van der Waals et une composante coulombique pour les interactions intermoléculaires auxquelles s'ajoutent des composantes intramoléculaires

permettant de considérer les contraintes imposées sur la conformation liée du ligand. Les fonctions de score Vinardo et Vina sont elles aussi relativement simples ; comparées à MCSS, elles ne comprennent pas de composantes intramoléculaires mais considèrent spécifiquement les interactions hydrophobe et hydrogène en fonction du type d'atomes considérés et de leur distance. DVRF20 est en revanche plus complexe puisqu'elle considère 20 descripteurs pour caractériser les interactions et estimer leur énergie par une approche de forêts d'arbres décisionnels. Compte tenu de cette complexité, les performances observées pour DVRF20 sur ce jeu de données sont en dessous des attentes espérées. Par ailleurs, le calibrage des fonctions de score Vinardo, Vina et DVRF20 est basé sur des données d'affinités de liaison tirées de complexes protéine-ligand qui incluent quelques complexes protéine-nucléotide (Liu et al., 2017). Notons qu'une fonction de score montrant de bonnes capacités à prédire des affinités de liaison ne montre pas nécessairement une bonne capacité à prédire le mode d'interaction natif (Su et al., 2019). Les résultats affichés par ces fonctions de score peuvent cependant présenter un certain biais en faveur de complexes protéine-nucléotide qui pourraient être similaires ou homologues à certains complexes utilisés pour leur apprentissage. Des vérifications doivent être effectuées pour évaluer cette redondance potentielle entre données apprises et données testées. Néanmoins, compte tenu de cet aspect, les performances affichées par la fonction de score MCSS sont assez remarquables puisque son calibrage est complètement indépendant de données d'affinité de liaison. Cela illustre l'avantage des fonctions de score basées sur un modèle physique qui visent à être transférables à différents systèmes.

Les faibles performances affichées par la fonction ITscorePR sont surprenantes. Cette fonction de score à potentiel statistique a été calibrée itérativement sur 110 complexes protéine-ARN jusqu'à ce que le potentiel parviennent à discriminer précisément des chaînes ARN natives de chaînes leurs générées artificiellement. On pourrait penser que les 110 complexes protéine-ARN sont majoritairement composés de chaînes ARN structurées faisant intervenir des interactions différentes de celles impliquées pour la reconnaissance d'ARNsb non structurées dans leur forme liée. Dans ce dernier cas, la reconnaissance implique des contacts spécifiques entre la base et la protéine, comme c'est le cas pour les complexes protéine-nucléotide utilisés ici comme jeu de données (voir en "Annexes", Fig. 56-A2). Cependant, cette hypothèse ne semble pas expliquer les faibles performances observées pour ITscorePR puisque cette dernière a montré sur la forme liée de protéines liant des ARN de très bonnes capacités pour discriminer le mode d'interaction natif des chaînes ARNsb non structurées dans leur forme liée (S.-Y. Huang & Zou, 2014b). Cette capacité pourrait être évaluée sur les chaînes générées par l'approche FBDRNA. Dans le cas où de bonnes prédictions du mode d'interaction seraient observées, on pourrait supposer que la procédure suivie

pour calibrer ITscorePR a conduit à un surentraînement en faveur des interactions protéine-ARN et que cette fonction de score n'est simplement pas adaptée pour estimer l'énergie d'interaction entre protéines et mono-nucléotides.

L'ensemble des résultats montrent que la fonction de score MCSS présentent des performances intéressantes comparées aux autres fonctions de score testées : sur la base de leur capacité à identifier une pose native au moins dans le top 5, la fonction de score MCSS serait la deuxième fonction de score la plus performante derrière Vinardo. Les meilleures performances observées pour Vinardo suggèrent que cette fonction de score pourrait être bénéfique à l'approche FBDRNA en permettant de réduire le nombre de poses à sélectionner pour conserver des poses natives, et donc réduire l'espace de solutions des chaînes identifiées à partir de ces poses. Cette hypothèse sera testée en ré-évaluant l'énergie d'interaction des poses retenues après clustering à 2 Å dans l'étape "diviser pour mieux régner". Il est néanmoins important de souligner qu'une des difficultés rencontrée dans l'approche FBDRNA est liée à la discrimination des poses natives associées aux nucléotides expérimentaux établissant peu de contacts avec la protéine. Ces poses natives ont une énergie d'interaction peu favorable qui rend leur sélection délicate. Les résultats présentés dans ce chapitre découlent d'observations faites sur 120 complexes protéine-nucléotide. Ils donnent une vision relativement générale des performances discriminatives des fonctions de score comparées mais ne permettent pas de dire qu'une fonction de score donnée favorise l'identification du mode d'interaction natif lorsque ce dernier implique peu de contacts protéine-nucléotide. Malgré de meilleures performances dans l'identification de poses natives dans les tops 1, 5 et 10 sur les 120 protéines du jeu données, il n'est donc pas garanti que Vinardo permette de faciliter l'identification du mode d'interaction natif de nucléotides liés à des RBDs.



## VII Conclusions générales et perspectives

Les interactions protéine-ARN interviennent dans de nombreux processus cellulaires fondamentaux. Ces interactions font intervenir des protéines de liaison à l'ARN (RBPs) généralement composées de un à plusieurs domaines de liaison à l'ARN (RBDs). Ces RBDs représentent les modules primaires de liaison des RBPs qui reconnaissent de courtes séquences ARN non structurées dans leur forme liée (ARNsb), et généralement de manière spécifique. La compréhension détaillée des mécanismes de reconnaissance passe généralement par l'obtention de structures de complexes protéine-ARN à haute résolution. Ces structures sont typiquement obtenues par la cristallographie aux rayons X et la RMN. La résolution structurale de complexes protéine-ARN par ces méthodes expérimentales est malheureusement difficile et laborieuse. Des approches de modélisation comme les méthodes de docking peuvent être utilisées en alternative pour obtenir de tels complexes. Modéliser les interactions protéine-ARN représente cependant un défi délicat à surmonter, notamment en raison de la variabilité conformationnelle de l'ARN. Le challenge est encore davantage accru pour les approches de docking classiques lorsque l'ARN est non structurée dans sa forme liée. Dans ce cas, ces approches classiques ne peuvent se reposer sur une structure initiale de l'ARN en raison du large espace conformationnel qu'il est nécessaire d'explorer.

Au commencement de cette thèse en octobre 2014, aucune approche répondant à cette problématique n'était encore parue. Les travaux réalisés au cours de cette thèse ont donc porté sur le développement d'une approche de docking (FBDRNA) visant à prédire le mode d'interaction de courtes chaînes ARN non structurées dans leur forme liée à des RBDs. Pour traiter le problème de l'échantillonnage conformationnel, l'approche FBDRNA repose sur une segmentation de l'ARN en fragments mono-nucléotidiques (1-nt). Cette fragmentation permet de réduire la combinatoire en restreignant l'exploration de l'espace conformationnel de l'ARN localement à une échelle nucléotidique. Des chaînes ARN sont ensuite être construites à partir d'un sous-ensemble de poses nucléotidiques en connectant celles satisfaisant différentes contraintes (distances, séquence). La méthode FBDRNA a été développée pour répondre à deux objectifs distincts. A partir de la structure d'un RBD et de la connaissance de son site d'interaction, (i) prédire le mode d'interaction d'une séquence ARN connue ou (ii) prédire la séquence préférentiellement reconnue par le RBD simultanément à son mode d'interaction. La méthode pré-suppose que l'ARN est non structurée dans sa forme liée.

Développée et appliquée sur la forme liée de trois RBDs, l'approche a permis, à partir de la connaissance du site d'interaction des RBDs et de la séquence ARN, de prédire correctement (top 1) et avec une grande précision ( $\sim 1,5 \text{ \AA}$ ) le mode d'interaction natif de chaînes 3-nt sur deux des trois domaines. Pour le troisième domaine, la chaîne 3-nt de plus basse énergie reproduit correctement ( $2 \text{ \AA}$ ) le mode d'interaction de deux des trois nucléotides de la chaîne expérimentale.

Lorsque l'approche a été appliquée sans *a priori* sur la séquence ARN, la séquence native préférentiellement reconnue a pu être discriminée des autres séquences pour un seul domaine, et son mode d'interaction natif a pu être classé au deuxième rang parmi plusieurs dizaines de milliers de chaînes prédites. Cependant, pour les deux autres domaines, malgré que des chaînes reproduisant le mode d'interaction natif ont été discriminées efficacement (top 35) parmi les dizaines de milliers de modèles générés, la séquence native attendue n'a pu être distinguée des autres séquences.

Bien que ces résultats n'aient été obtenus que sur un faible jeu de données et à partir de la forme liée de RBDs, ils sont très encourageants dans l'optique de pouvoir prédire le mode d'interaction d'une courte chaîne ARNsb à partir de sa séquence. Comparée aux approches existantes parues au cours de cette thèse (RNA-LIM, l'approche basée sur ATTRACT et Rosetta *RNP-denovo*), l'approche FBDRNA peut être vue comme une version plus résolutive de la méthode RNA-LIM (chapitre IV de l'introduction, section 3.1). Dans l'optique de modéliser le mode d'interaction de chaînes ARNsb plus longues, on peut envisager, après validation sur un jeu de données plus important, que l'approche FBDRNA pourrait être utilisée en première étape de la méthode Rosetta *RNP-denovo* (chapitre IV de l'introduction, section 3.3) ou de l'approche basée sur ATTRACT (chapitre IV de l'introduction, section 3.2) pour prédire les courtes chaînes d'ancrage que ces approches utilisent comme contraintes pour la modélisation des chaînes.

L'approche FBDRNA développée ne permet en revanche pas, dans l'état actuel, de répondre clairement au problème d'identification d'une séquence préférentiellement reconnue par un RBD. Les travaux effectués pour répondre à cette problématique restent néanmoins intéressants. Ils ont en effet permis d'identifier une stratégie permettant de limiter la combinatoire inhérente à une reconstruction de chaînes à partir de poses sans contrainte de séquence imposée. Ils ont par ailleurs permis d'identifier des voies pouvant potentiellement améliorer les prédictions. Par ailleurs, l'excellente capacité de la fonction de score utilisée pour discriminer le mode d'interaction natif de chaînes indépendamment de leur séquence permet d'envisager la sélection d'un nombre réduit de candidats pour une ré-évaluation de leur énergie d'interaction par des fonctions de score plus précises (*e.g.* MM-PBSA ou MM-GBSA). Soulignons enfin l'aspect novateur de ces travaux puisque l'approche FBDRNA est à notre connaissance la seule méthode développée pour prédire

une séquence ARN préférentiellement reconnue par un RBD donné simultanément à son mode d'interaction.

### **Projets futurs d'amélioration de l'approche**

Plusieurs voies ont été identifiées pour compléter et améliorer l'approche FBDRNA. Ces améliorations demandent toutefois en premier lieu d'élargir le jeu de données des RBDs, quantitativement et qualitativement. Idéalement, ce jeu devra inclure la forme liée et non liée d'une même structure. Des modèles par homologie pourront être utilisés dans les cas où la forme non liée ne serait pas disponible. Ce jeu de données sera utile pour valider la robustesse de l'approche face à aux changements conformationnels des RBDs et envisager son application dans des conditions où la structure du RBD n'aurait pas été résolue expérimentalement. Par ailleurs, les banques de données comme RBPDB (Cook, Kazan, Zuberi, Morris, & Hughes, 2010), CISBP-RNA (Ray et al., 2013), ou AtTRACT (Giudice et al., 2016) pourront être utilisées pour obtenir des motifs ARN préférentiellement reconnus par des RBDs du jeu de données.

Pour parvenir à retenir l'ensemble des poses natives d'intérêt, l'approche FBDRNA passe par une procédure de sélection reposant entre autres sur des étapes de clustering. Les seuils utilisés pourront être optimisés de manière à réduire le nombre de poses retenues tout en conservant les poses natives d'intérêt. On peut également imaginer définir un ensemble de paramètres de sélection adaptés pour chaque famille de RBDs disponible. Le protocole de sélection mis au point repose également sur la capacité de la fonction de score à discriminer les poses natives. Des travaux préliminaires comparant différentes fonctions de score (partie VI) ont montré que la fonction de score Vinardo présentait sur un jeu de données protéine-nucléotide de meilleures performances discriminatives par rapport à la fonction de score MCSS utilisée dans l'approche FBDRNA. Cette fonction de score pourra être testée directement dans l'approche en ré-évaluant l'énergie d'interaction des poses générées par MCSS.

Les fonctions de score testées sur le jeu de données protéine-nucléotide (partie VI) ont aussi été choisies parce qu'elles ont montré sur des jeux de données standards des performances intéressantes pour classer des ligands par ordre croissant de leur affinité de liaison envers un récepteur donné. Cette capacité pourrait bénéficier à l'approche FBDRNA lorsque les chaînes sont recherchées sans contrainte imposée sur la séquence. Lorsque l'approche FBDRNA est appliquée sans *a priori* sur la séquence ARN, une des étapes de sélection des poses passe par la sélection des deux nucléotides de plus basse énergie qui convergent à une même position. Une fonction de score performante pourrait permettre de ne retenir qu'un seul nucléotide réduisant ainsi le nombre total de poses à sélectionner

et en conséquence l'espace de solutions des chaînes recherchées. Les fonctions de score testées sur le jeu données protéine-nucléotide pourront donc aussi être testées sur les RBDs pour la prédiction d'une séquence ARN préférentiellement reconnue. Rappelons à ce sujet que la fonction de score MCSS présente un biais favorisant l'énergie des purines par rapport aux pyrimidines. Ce biais est pénalisant pour l'identification d'un nucléotide spécifiquement reconnu à un site d'interaction donné. Il peut être corrigé par un Z-score ; il serait donc aussi intéressant d'évaluer l'effet de cette correction sur la capacité de la fonction de score à identifier un nucléotide préférentiellement reconnu.

Que la séquence ARN soit connue ou non, le protocole de sélection des poses pour la recherche de chaînes conduit à générer de une à plusieurs dizaines de milliers de solutions environ. L'approche FBDRNA impose de minimiser l'ensemble des solutions avant d'estimer leur énergie d'interaction. Afin de réduire les temps de calculs, le nombre de chaînes à minimiser pourrait être réduit en écartant au préalable les chaînes composées de nucléotides de plus faible énergie. Par exemple, si la somme de l'énergie des nucléotides de la chaîne ne dépasse pas un certain seuil, alors la chaîne est éliminée. Cette hypothèse pourra aussi être évaluée sur un jeu de données plus important.

### **Perspectives d'application à long terme**

Les travaux d'amélioration nécessaires au développement de l'approche FBDRNA pourraient permettre à plus long terme d'envisager son application à partir uniquement de la séquence d'un RBD. De nombreuses séquences sont annotées comme des RBDs classiques (RRM, KH, Zn-CCCH), et ce dans différents règnes du vivant. Bénéficier d'une méthode permettant d'affecter à chacune des séquences annotées comme RBD une ou plusieurs séquences ARN préférentiellement reconnues seraient d'un intérêt fondamental pour identifier leurs cibles et révéler le réseau d'interaction dans lequel interviennent ces RBDs. Par ailleurs, si cette méthode permet en plus de prédire leur mode d'interaction, l'accès aux mécanismes de reconnaissance à l'échelle atomique permettrait de mieux comprendre l'interaction et de potentiellement identifier un code de reconnaissance qui serait lui-même d'un intérêt majeur, à la fois d'un point de vue fondamental et thérapeutique : un tel code permettrait par exemple d'envisager l'ingénierie de protéines pour aller cibler des ARN d'intérêt. Le développement de l'approche FBDRNA mérite donc d'être poursuivi pour pouvoir un jour répondre à ces objectifs.

## VIII Annexes

### 1 Définition d'un seuil de clustering adapté pour la stratégie de sélection "diviser pour mieux régner"

Afin de définir un seuil de clustering offrant le meilleur équilibre entre les valeurs  $N_c$  et  $N_p$ , quatre seuils RMSD différents ont été testés : 4 Å, 5 Å, 6 Å et 7 Å. Pour chaque ensemble de poses issues des calculs de docking réalisés sur les trois systèmes, une matrice est d'abord calculée de manière à recenser les valeurs RMSD entre toutes les paires de poses. A partir de cette matrice, les poses sont regroupées entre elles d'après un seuil de clustering. Les clusters issus de ce regroupement sont ordonnés du cluster le plus peuplé au moins peuplé. Pour chacun des clusters résultants, une deuxième étape de clustering est effectuée avec un seuil de 2 Å de manière à éliminer la redondance entre les poses.

*Tableau 20: Détails des valeurs  $N_c$  et  $N_p$  obtenues à partir du protocole "diviser pour mieux régner" pour différents seuils de clustering. Pour chaque nucléotide expérimental constituant les chaînes ARN du jeu de données, les valeurs  $N_c$ ,  $N_p$  et  $N_f$  sont indiquées. La valeur  $N_c$  correspond au rang du cluster le plus peuplé contenant au moins une pose native ; la valeur  $N_p$  correspond au rang de la pose native de plus basse énergie trouvée dans le cluster de rang  $N_c$ . La valeur  $N_f$  indique le nombre minimal de poses totales à sélectionner pour retenir au moins une pose native ;  $N_f$  est le produit de  $N_c$  par  $N_p$ . La valeur  $N_f$  en gras met en évidence la plus haute valeur observée sur les trois systèmes pour chaque seuil de clustering testé.*

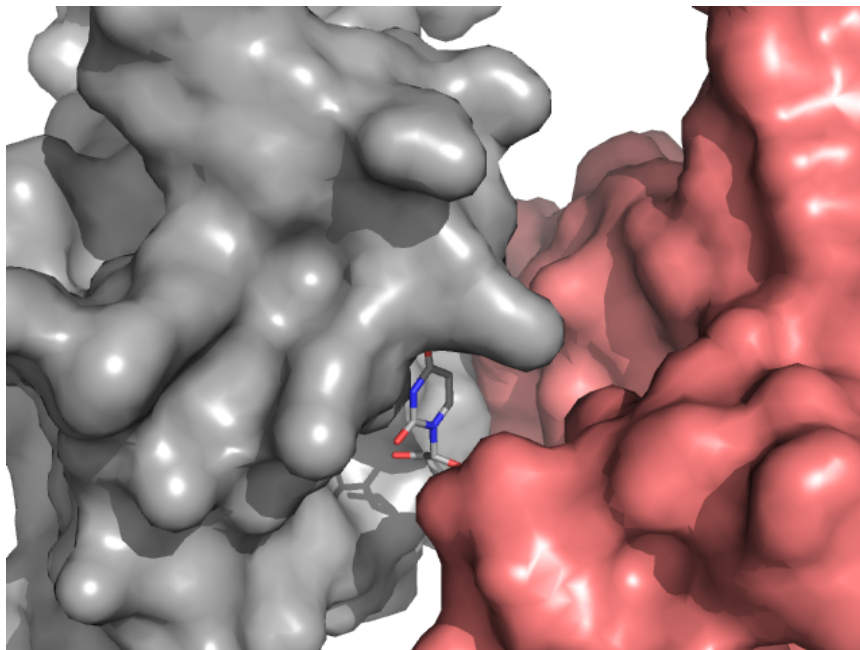
Code PDB	Nucléotide expérimental	Seuil RMSD utilisé pour le clustering											
		4 Å			5 Å			6 Å			7 Å		
		$N_c$	$N_p$	$N_f$	$N_c$	$N_p$	$N_f$	$N_c$	$N_p$	$N_f$	$N_c$	$N_p$	$N_f$
2XNR	U1	11	1	<i>11</i>	5	1	5	3	1	3	3	12	36
	C2	4	1	4	1	1	<i>1</i>	1	1	<i>1</i>	1	1	<i>1</i>
	U3	6	1	6	8	1	8	2	3	6	2	4	8
5ELH	U1	21	9	<i>189</i>	11	21	<b>231</b>	2	157	<i>314</i>	1	173	<i>173</i>
	U2	25	2	<i>50</i>	6	1	6	1	23	23	1	3	3
	A3	41	1	<i>41</i>	10	1	<i>10</i>	5	1	5	2	1	2
5WWX	A1	21	16	<b>336</b>	9	1	9	4	93	<b>372</b>	4	84	<b>336</b>
	G2	3	1	3	2	1	2	1	6	6	1	6	6
	A3	12	15	<i>180</i>	2	7	<i>14</i>	1	17	<i>17</i>	1	37	<i>37</i>

Le tableau 20 recense, pour chaque nucléotide expérimental et pour les quatre seuils de clustering testés, les valeurs  $N_c$  et  $N_p$  minimales pour retenir dans la sélection leur pose native. La valeur  $N_c$  correspond au rang du cluster le plus peuplé contenant au moins une pose native ; la valeur  $N_p$  correspond au rang de la pose native de plus basse énergie trouvée dans le cluster de rang

Nc. La valeur Nf indique enfin le nombre minimal de poses totales à sélectionner pour retenir au moins une pose native ; ce nombre est défini par le produit de Nc par Np. La valeur Nf est utilisée comme indice pour définir le seuil de clustering permettant de sélectionner le plus petit nombre de poses tout en retenant des poses natives pour l'ensemble des neuf nucléotides expérimentaux. Le seuil de clustering conduisant à la plus petite sélection de poses pour les trois systèmes et le seuil à 5 Å.

## 2 Contacts cristallins du nucléotide U1 de 5ELH

Le cristal de la structure 5ELH a été reconstruit par opérations de symétrie. La figure 52 montre que le nucléotide U1 est pris en sandwich entre deux unités asymétriques (Fig. 52). Cette observation sème le doute sur la pertinence biologique du mode d'interaction du nucléotide U1 de 5ELH et pourrait être une raison expliquant les difficultés rencontrées à bien classer les poses natives associées à ce site.



*Figure 52: Contacts cristallins du nucléotide U1 de 5ELH. L'unité asymétrique projetée par opérations de symétrie est représentée en surface grise ; l'unité asymétrique utilisée pour les calculs de docking est en surface rouge. Le nucléotide U1, représenté en bâtonnet, est pris en sandwich entre les deux unités asymétriques.*

### 3 Comparaison de l'énergie d'interaction des poses puriques et pyrimidiques

Les quatre types de nucléotides ont été dockés au niveau du site d'interaction des RBDs 2XNR, 5ELH et 5WWX. Les différents ensembles de poses obtenues ont alors été regroupés par type de nucléotides. La figure 53 montre la distribution des énergies d'interaction obtenues pour les poses puriques (A et G) et pyrimidiques (C et U).

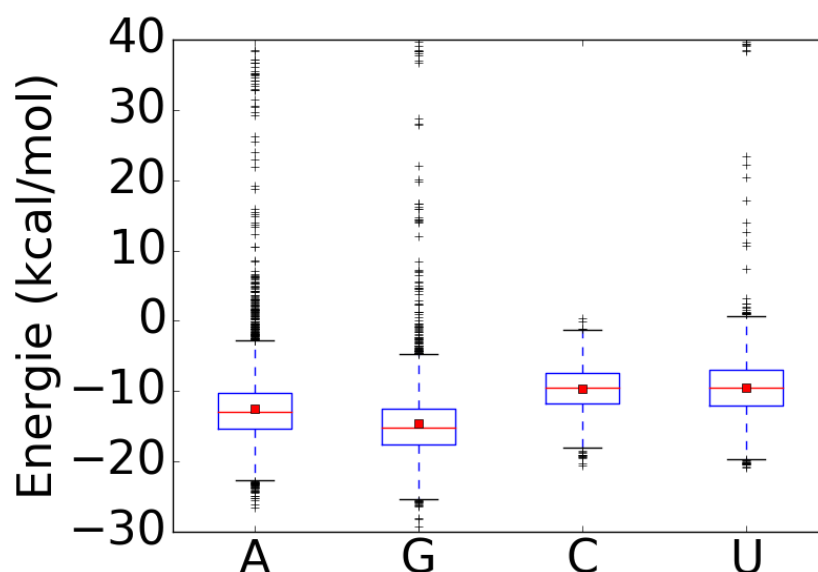
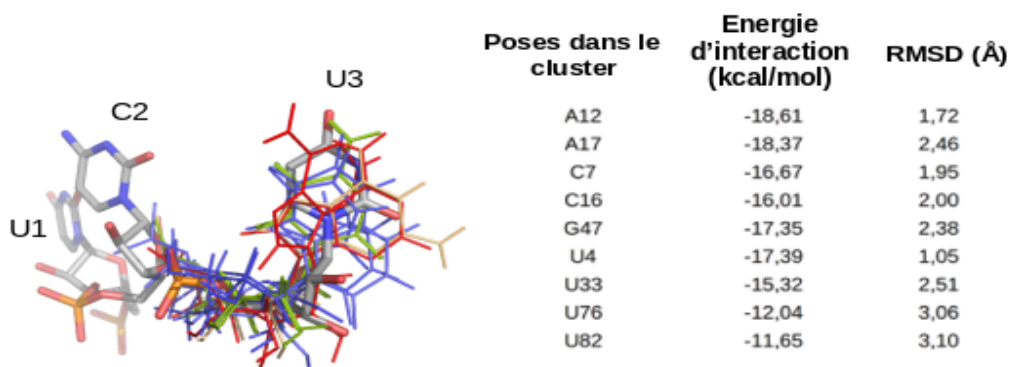


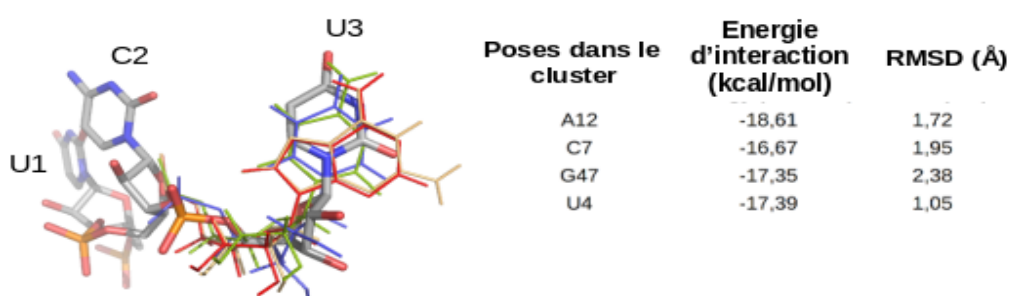
Figure 53: Distribution de l'énergie d'interaction de l'ensemble des poses générées par type de nucléotides dockés sur les trois RBDs du jeu de données. Les distributions sont représentées sous forme de boîtes à moustaches ; la boîte donne les limites du premier et dernier quartile, la barre du milieu en rouge représente la médiane ; les moustaches s'étendent jusqu'aux valeurs les plus extrêmes dans la limite de 1.5 fois la hauteur de la boîte, et les points au-delà des moustaches sont représentés par des points isolés ; la moyenne est indiquée par un carré rouge.

### 4 Sélection des deux nucléotides de plus basse énergie dans les clusters à 2 Å

La figure 54 ci-dessous illustre l'étape 3 de sélection représentée à la figure 41 du chapitre V.



### 1. Sélection de la pose de plus basse énergie pour chaque type de nucléotide



### 2. Sélection des 2 poses de plus basse énergie

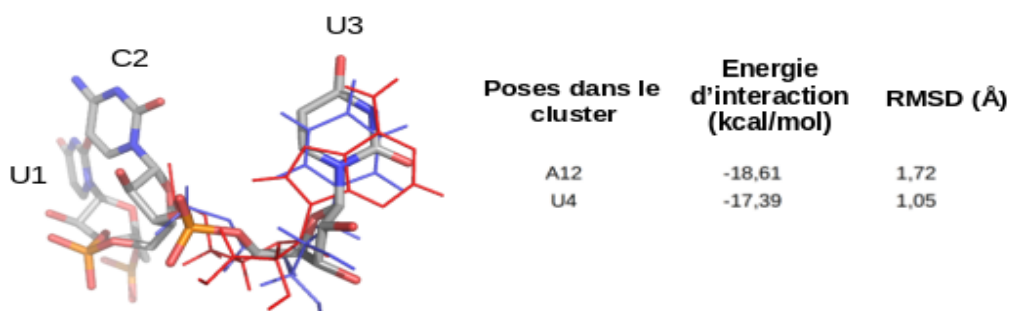


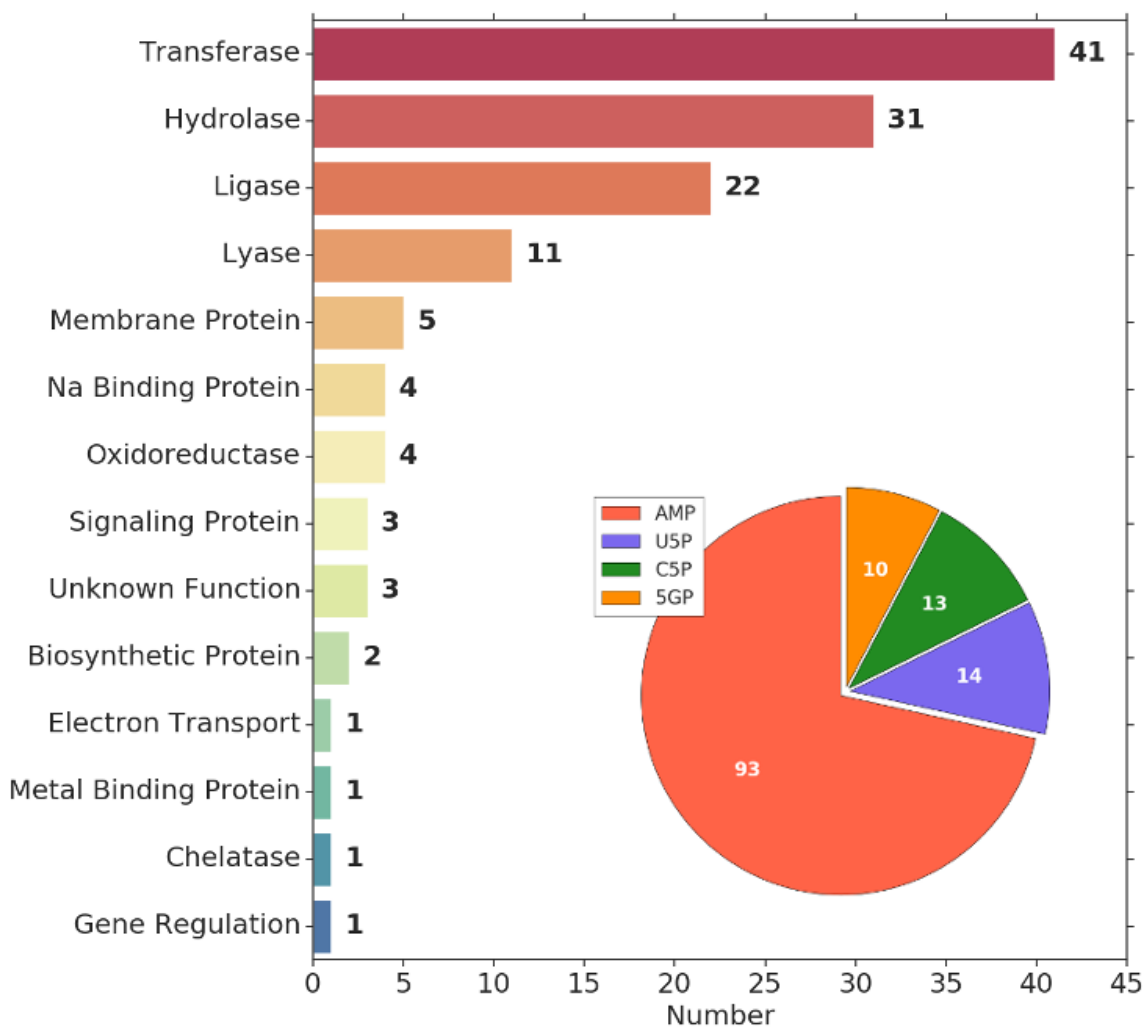
Figure 54: Illustration de la sélection des deux nucléotides de plus basse énergie dans un groupe de poses issu d'un clustering à 2 Å. La figure montre un groupe de poses obtenu sur 2XNR et dont les nucléotides se superposent au nucléotide expérimental U3 de la chaîne ARN  $U_1C_2U_3$ . Ce groupe est initialement composé de plusieurs type de nucléotides. La 1ère étape consiste à retenir pour chaque type de nucléotides la pose de plus basse énergie. La seconde étape retient les deux poses de plus basse énergie. Les poses dans le cluster sont représentées à gauche des tables en mode ligne ; les adénines sont en rouge, les uraciles en bleu, les guanines en orange et les cytosines en vert. La chaîne ARN UCU est représentée en bâtonnet avec les atomes de carbone en gris, d'oxygène en rouge, d'azote en bleu et de phosphore en orange.



## 5 Description du jeu de données protéine-nucléotide non-redondant

130 complexes protéine-ribonucléotide-5'-P (nucléotide) non-redondants et de haute résolution ont été sélectionnés (voir matériels et méthodes) pour bénéficier d'un jeu de données spécifiques des interactions entre protéines et acides nucléiques. Voici la liste de leur code PDB : 1EX7, 1HDI, 1HXP, 1IYB, 1JP4, 1KTG, 1NH8, 1QF9, 1QGX, 1RAO, 1S68, 1UA4, 1UCD, 1UJ2, 1UUY, 1WXI, 1XTT, 1Y1P, 1Z4M, 2A7X, 2CFM, 2CNQ, 2EQA, 2FFC, 2FJB, 2G1U, 2GXQ, 2II6, 2J91, 2JB7, 2JBH, 2OUN, 2Q4H, 2QRK, 2R85, 2UV4, 2VFK, 2WNB, 2XBU, 2XWM, 2YAB, 2YRX, 2YVO, 3AKE, 3C85, 3CJ9, 3CLS, 3DDJ, 3DJX, 3DLZ, 3EWY, 3FEG, 3FWZ, 3G1Z, 3GLV, 3GRU, 3IB8, 3KD6, 3KGD, 3L9W, 3LFR, 3LKM, 3M84, 3N1S, 3NUA, 3NYQ, 3O0M, 3OMF, 3PLN, 3REX, 3RL4, 3RPZ, 3SF0, 3TTF, 3UQ8, 3UWQ, 3W07, 4BLW, 4BRQ, 4CO4, 4CS3, 4D05, 4D7A, 4EEI, 4EMD, 4EQL, 4EUM, 4FBC, 4FE3, 4G0P, 4H2W, 4HE2, 4IG1, 4IJN, 4IKE, 4JEM, 4KBF, 4M0K, 4M9D, 4MA0, 4MPO, 4MX2, 4NDF, 4O6M, 4OKE, 4OZL, 4P86, 4PNO, 4R78, 4UUW, 4WW7, 4X9D, 4XBA, 4ZCP, 4ZFN, 5B6D, 5B8F, 5BPH, 5COT, 5D4N, 5ED3, 5ERS, 5GMD, 5JDA, 5K1D, 5M45, 5T8S, 5V0I, 5V1M, 5X0J.

L'ensemble des 130 complexes protéine-nucléotide à haute résolution est constitué de différentes catégories fonctionnelles de protéines dont l'essentielle est associée à des activités enzymatiques, le reste se répartissant en protéines de diverses fonctions. La figure 55 montre leur répartition ainsi que le type de ligand nucléotidique complexé à chacune des structures. Notons un enrichissement écrasant en protéines liées à un AMP. Cette disproportion est certainement une conséquence de l'importance biologique de l'adénine qui est retrouvée dans un grand nombre de processus biochimiques essentiels.



**Figure 55 : Répartition des catégories fonctionnelles des protéines composant le jeu de données et du type de nucléotides qui leur est associé. L'étiquette "Na binding protein" correspond à des protéines liant des acides nucléiques (ADN et/ou ARN).**

La figure 56 recense quelques propriétés des sites de liaison de l'ensemble des complexes. La plupart des nucléotides établit un nombre de contacts avec la protéine compris entre 100 et 200 (Fig. 56, A-1), un contact étant défini par tout atome du ligand séparé par moins de 4 Å de tout atome de la protéine. La majorité des contacts nucléotide-protéine implique des atomes de la base, puis du ribose (Fig. 56, A-2). Le nombre de liaisons hydrogène varie lui de 0 à 15 lorsque l'ensemble du nucléotide est considéré (Fig. 56, B-1). Il est plus important au niveau de la base et du groupement phosphate (Fig. 56, B-2). Pour ce dernier, cette observation peut s'expliquer par le fait qu'une majorité des protéines correspondent à des transférases où des liaisons hydrogène au niveau du groupement phosphate doivent participer à catalyser son transfert. Le nombre de liaisons hydrogènes et le nombre de contacts trouvés en plus grand nombre au niveau de la base reflètent

son importance dans le mode d'interaction et sa reconnaissance sélective. La fraction de la surface enfouie du nucléotide (*i.e.* non accessible au solvant) se distribue pour la majorité des structures entre 75 et 100 % (Fig. 56, C). Notons finalement la gamme étendue des énergies d'interactions qui va de -45 à -10 kcal/mol (Fig. 56, D). L'énergie est principalement dominée par la composante de van der Waals (vdW); seuls 11 complexes - non montrés ici - ont une composante électrostatique plus favorable à celle de vdW (ils correspondent à l'épaule gauche de la distribution électrostatique sur la Fig. 56, D). Globalement, l'analyse des sites de liaison semble refléter des modes d'interaction suffisamment variés pour valider l'intérêt de ce jeu de données pour l'évaluation de diverses performances de docking.

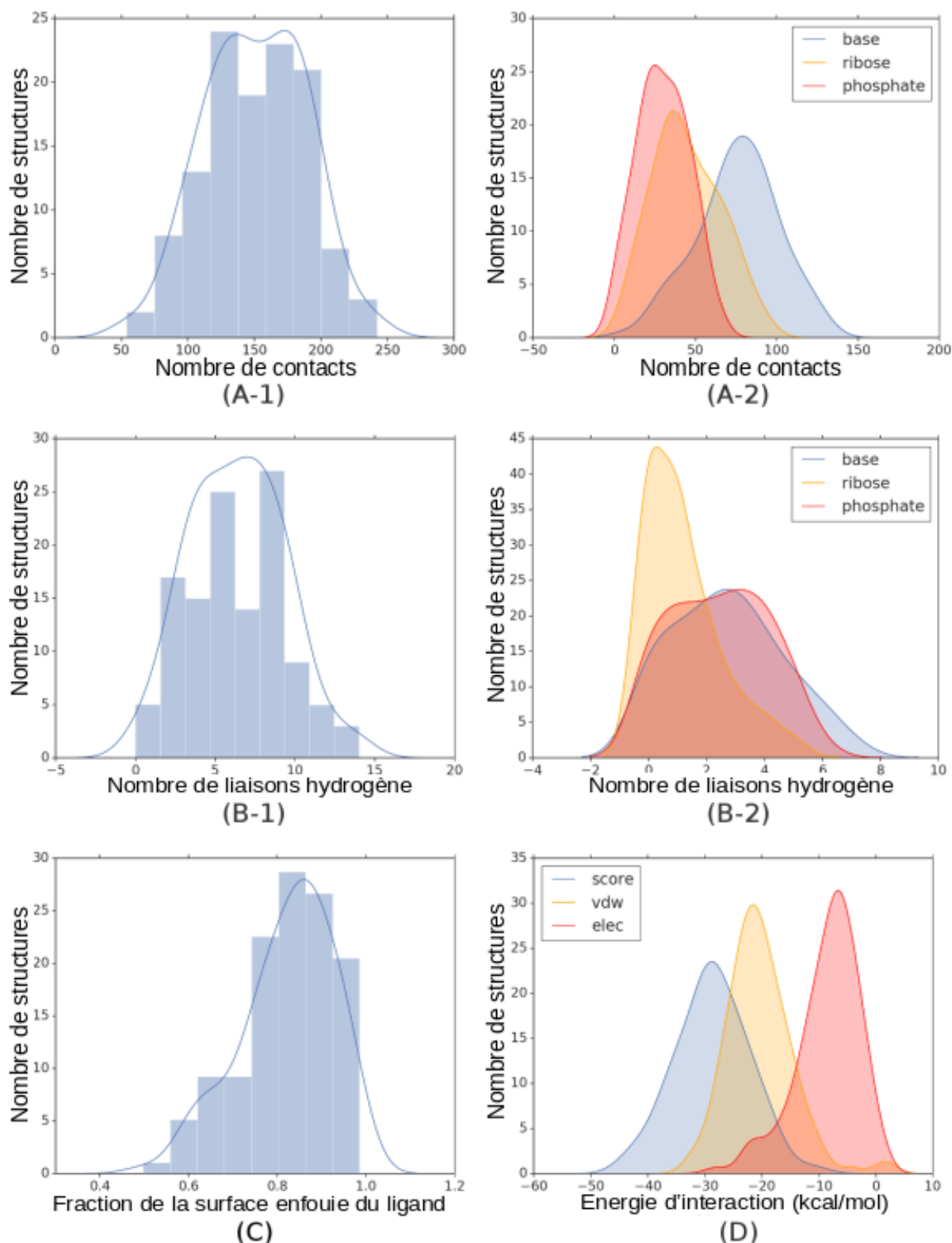


Figure 56: Distributions de quelques propriétés du jeu de données composé de 130 complexes protéine-nucléotide. (A) Nombre de contacts entre le ligand et la protéine ; (B) Nombre de liaisons hydrogène établies entre le ligand et la protéine - pour ces deux propriétés, le ligand est considéré dans sa globalité (A-1, B-1) ou est décomposé en ses groupements base, ribose et phosphate (A-2, B-2) ; (C) Fraction de la surface enfouie du ligand ; (D) Energies d'interaction (score) et ses composantes de van der Waals (vdw) et électrostatique (elec).

### **5.1 Analyse des contacts et liaisons hydrogènes protéine-ligand**

L'analyse des contacts et des liaisons hydrogène a été faite avec le programme BINANA (Zhang & Skolnick, 2005). Un contact est défini entre un atome du ligand et un atome de la protéine si la distance qui les sépare est inférieure à 4 Å. Pour les liaisons hydrogène, les groupes hydroxyle et amine agissent comme donneurs, et les atomes d'azote et d'oxygène comme accepteurs. Une liaison hydrogène est identifiée si la distance séparant le donneur de liaison hydrogène de l'accepteur est inférieure à 4 Å et si l'angle formé entre l'atome donneur, l'hydrogène et l'atome accepteur est inférieur à 40 degrés.

### **5.2 Evaluation de l'énergie d'interaction entre protéine et ligand cristallisé**

Pour évaluer l'énergie d'interaction entre le nucléotide expérimental et la protéine, les 130 complexe ont été minimisés dans les mêmes conditions que celles utilisées pour les calculs de docking. La protéine a d'abord été préparée tel qu'indiqué dans la partie "Méthodes générales – section 1.2". CHARMM associe un patch à chacune des structures de ligand nucléotidique disponible. Il est donc possible de modifier simplement le ligand expérimental en conséquence. La distribution des énergies présentée à la figure 56 est issue de la structure 210. Elle est celle qui se rapproche le plus des ligands cristallisés avec un atome d'oxygène du groupement phosphate déprotoné, et un autre atome non protoné. La procédure de minimisation du ligand est comme suit : 500 étapes initiales de SD suivies de 300 autres étapes de SD, puis 20 répétitions de 500 étapes de gradient conjugué et enfin 200 étapes de ABNR. L'énergie est estimée suivant le schéma de scoring MCSS (voir "Méthodes générales", section 1.5).

### **5.3 Fraction de la surface enfouie du ligand**

Le programme NACCESS (Durrant & McCammon, 2011) a été utilisé pour mesurer l'accessibilité au solvant du ligand d'une part ( $SAS_{ligand}$ ), et du complexe protéine-ligand d'autre part ( $SAS_{complexe}$ ). La fraction de la surface du ligand enfouie ( $FBSA_{ligand}$ ) est alors définie par :

$$FBSA_{ligand} = 1 - (SAS_{complexe} / SAS_{ligand}) \quad (24)$$

## **6 Ajustement du jeu de données protéine-nucléotide pour les calculs de docking**

Plusieurs complexes se sont révélés problématiques suite aux calculs de docking. Tout d'abord, deux complexes ont été retirés en raison de l'impossibilité de positionner des poses au niveau de leur site d'interaction :

- 5M45, qui présente un site d'interaction trop enfoui au sein de la protéine ; l'accessibilité au solvant de son ligand est nulle
- 5DJH, qui présente un clash stérique entre la base de l'adénine et le groupe carbonyle de G207 après minimisation de la protéine

Huit autres structures se sont aussi été révélées problématiques *a posteriori* des calculs de docking réalisés avec les cinq différentes structures de ligand testées.

Deux cas de figures peuvent être distingués.

Le premier correspond à des structures pour lesquelles aucune pose native ( $\text{RMSD} \leq 2 \text{ \AA}$ ) n'a pu être générée. Une pose native est définie par un seuil RMSD qui doit être inférieur ou égal à 2 Å, le calcul RMSD utilisant les coordonnées du ligand expérimental comme référence (indice X-ray dans les 3 tableaux qui suivent). La minimisation de ce dernier, réalisée dans les mêmes conditions que celles du docking, montre pour ces structures qu'il dévie fortement par rapport à sa position/conformation initiale (tableau 21 – colonne Ligand X-ray min); par ailleurs, lorsque le ligand expérimental minimisé est utilisé comme référence (indice X-ray min), des poses natives sont retrouvées (tableau 17 – colonne Native pose<sub>X-ray min</sub>). Cela indique que le ligand expérimental (non minimisé) n'est pas/plus dans un minimum énergétique dans les conditions de docking utilisées (les causes peuvent provenir de la préparation de la protéine et/ou des paramètres du champ de force).

Tableau 21: Complexes pour lesquels aucune pose native n'a pu être générés en raison d'une déviation importante du ligand expérimental minimisé.

PDBID	R	Native pose <sub>X-ray</sub>		Ligand <sub>X-ray min</sub>		Native pose <sub>X-ray min</sub>	
		E <sub>int</sub>	RMSD <sub>exp</sub>	E <sub>int</sub>	RMSD <sub>X-ray</sub>	E <sub>int</sub>	RMSD <sub>X-ray min</sub>
1HXP	010	NA	NA	-20,43	2,22	-20,27	0,87
1HXP	110	NA	NA	-17,06	2,41	-16,68	1,10
1HXP	210	NA	NA	-20,82	2,29	-22,60	1,19
1HXP	410	NA	NA	-19,91	2,12	-22,06	0,77
2CFM	010	NA	NA	-36,70	3,41	-35,45	0,57
2CFM	210	NA	NA	-45,10	3,50	-39,11	0,83
2CFM	310	NA	NA	-39,15	3,02	-39,87	1,14
4OKE	010	NA	NA	-23,36	2,01	-24,88	1,48
4OKE	410	NA	NA	-18,69	2,95	-18,57	0,52
4XBA	210	NA	NA	-23,18	3,46	-26,88	0,27
5ERS	010	NA	NA	-21,83	2,87	-18,20	0,04
5ERS	210	NA	NA	-22,49	2,89	-22,23	1,98
5ERS	310	NA	NA	-24,67	2,85	-23,90	1,89
5ERS	410	NA	NA	-24,31	2,83	-21,70	1,91

Dans le deuxième cas de figure, des poses natives sont générées pour les structures concernées mais leur énergie d'interaction ( $E_{int}$ ) est fortement défavorable (tableau 22 – colonne Native pose<sub>X-ray</sub>). Deux situations peuvent être distinguées au regard du comportement du ligand expérimental minimisé qui, soit dévie relativement peu par rapport à sa position/conformation initiale, soit dévie d'une manière plus importante. Pour la première, les énergies d'interaction du ligand expérimental minimisé ainsi que celles des poses natives définies par rapport à ce dernier sont également défavorables (tableau 22 – colonne Native pose<sub>X-ray min</sub>), témoignant de la présence de clashes stériques interdisant au ligand d'atteindre une énergie d'interaction favorable.

Tableau 22: Complexes pour lesquels des poses natives ont été générées mais leur énergie d'interaction est défavorable en raison de clashes stériques

PDBID	R	Native pose <sub>X-ray</sub>		Ligand <sub>X-ray min</sub>		Native pose <sub>X-ray min</sub>	
		$E_{int}$	RMSD <sub>exp</sub>	$E_{int}$	RMSD <sub>X-ray</sub>	$E_{int}$	RMSD <sub>X-ray min</sub>
2CFM	110	8,37	1,49	-3,35	0,65	8,37	1,20
2Q4H	010	10,45	1,89	86,93	0,92	10,45	1,42
2Q4H	210	21,05	1,89	134,09	1,11	21,05	1,85
2Q4H	310	33,13	1,79	41,11	1,63	33,13	1,27
2Q4H	410	20,31	1,89	131,70	1,11	20,31	1,85
3L9W	010	19,45	1,19	3,40	1,05	19,45	0,81
3L9W	110	15,82	1,20	8,02	1,06	15,82	0,67
3L9W	210	14,36	1,17	-0,47	1,04	14,36	0,85
3L9W	310	13,64	1,17	0,13	1,04	13,64	0,83
3L9W	410	12,86	1,18	-1,46	1,04	12,86	0,86
3REX	010	20,75	1,13	-2,53	2,16	-2,07	1,96
3REX	110	14,60	1,87	0,51	2,26	13,68	1,38
3REX	210	19,14	1,98	-6,44	2,32	-4,99	2,00
3REX	310	18,43	1,09	-5,79	2,28	-5,10	1,97

La seconde situation s'apparente au premier cas de figure du tableau 22 où le ligand expérimental (non minimisé) ne correspond pas à un minimum énergétique dans les conditions de docking utilisées. Des poses natives définies par rapport à ce dernier sont cependant générées mais leur énergie est positive (tableau 23). Les énergies d'interaction des poses natives définies par rapport au ligand expérimental minimisé sont en revanche très favorables (tableau 23 – colonnes Native pose<sub>X-ray min</sub> et Ligand X-ray min, respectivement).

Tableau 23: Complexes pour lesquels des poses natives ont été générées mais leur énergie d'interaction est défavorable en raison d'un mauvais référentiel du minimum énergétique

PDBID	R	Native pose <sub>X-ray</sub>		Ligand <sub>X-ray</sub> min		Native pose <sub>X-ray</sub> min	
		E <sub>int</sub>	RMSD <sub>exp</sub>	E <sub>int</sub>	RMSD <sub>X-ray</sub>	E <sub>int</sub>	RMSD <sub>X-ray</sub> min
2CFM	410	15,11	1,20	-46,26	3,41	-43,24	0,70
4XBA	310	7,39	1,92	-31,64	3,83	-31,71	1,30
4XBA	410	8,49	1,84	-26,87	3,36	-25,79	0,24
5ERS	110	9,86	1,70	-17,71	2,89	-19,13	1,97

Parmi les huit structures recensées, certaines sont problématiques quelle que soit la configuration du fragment utilisée (3L9W, 3REX et 5ERS), d'autres ne le sont que pour quelques-unes d'entre elles (1HXP, 2Q4H, 4OKE et 4XBA). Par souci d'homogénéité et de simplicité d'analyses, ces huit structures ont été retirées du jeu de données pour l'ensemble des travaux réalisés et présentés ci-dessous. Ce sont donc 120 complexes qui sont au final considérés.



## Bibliographie

- Adam, S. A., Nakagawa, T., Swanson, M. S., Woodruff, T. K., & Dreyfuss, G. (1986). mRNA polyadenylate-binding protein: gene isolation and sequencing and identification of a ribonucleoprotein consensus sequence. *Molecular and Cellular Biology*, 6(8), 2932–2943. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/3537727>
- Afroz, T., Cienikova, Z., Cléry, A., & Allain, F. H. T. (2015). One, two, three, four! How multiple RRM reads the genome sequence. *Methods in Enzymology*, 558(1), 235–278. <https://doi.org/10.1016/bs.mie.2015.01.015>
- Agostini, F., Zanzoni, A., Klus, P., Marchese, D., Cirillo, D., & Tartaglia, G. G. (2013). catRAPID omics: a web server for large-scale prediction of protein-RNA interactions. *Bioinformatics*, 29(22), 2928–2930. <https://doi.org/10.1093/bioinformatics/btt495>
- Ainger, K., Avossa, D., Diana, A. S., Barry, C., Barbarese, E., & Carson, J. H. (1997). Transport and localization elements in myelin basic protein mRNA. *The Journal of Cell Biology*, 138(5), 1077–1087. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/9281585>
- Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8), 831–838. <https://doi.org/10.1038/nbt.3300>
- Ashtawy, H. M., & Mahapatra, N. R. (2012). A Comparative Assessment of Ranking Accuracies of Conventional and Machine-Learning-Based Scoring Functions for Protein-Ligand Binding Affinity Prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(5), 1301–1313. <https://doi.org/10.1109/TCBB.2012.36>
- Auweter, S. D., Oberstrass, F. C., & Allain, F. H. T. (2006). Sequence-specific binding of single-stranded RNA: Is there a code for recognition? *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gkl620>
- Baker, N. A. (2005). *Biomolecular Applications of Poisson-Boltzmann Methods*. <https://doi.org/10.1002/0471720895.ch5>
- Bakheet, T., Frevel, M., Williams, B. R., Greer, W., & Khabar, K. S. (2001). ARED: human AU-rich element-containing mRNA database reveals an unexpectedly diverse functional repertoire of encoded proteins. *Nucleic Acids Research*, 29(1), 246–254. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/11125104>
- Ball, P. (2008). Water as an active constituent in cell biology. *Chemical Reviews*, 108(1), 74–108. <https://doi.org/10.1021/cr068037a>

- Bandziulis, R. J., Swanson, M. S., & Dreyfuss, G. (1989). RNA-binding proteins as developmental regulators. *Genes & Development*, 3(4), 431–437. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/2470643>
- Bao, X., Guo, X., Yin, M., Tariq, M., Lai, Y., Kanwal, S., ... Esteban, M. A. (2018). Capturing the interactome of newly transcribed RNA. *Nature Methods*, 15(3), 213–220. <https://doi.org/10.1038/nmeth.4595>
- Barabino, S. M. L., & Keller, W. (1999). Last but Not Least: Regulated Poly(A) Tail Formation. *Cell*, 99(1), 9–11. [https://doi.org/10.1016/S0092-8674\(00\)80057-4](https://doi.org/10.1016/S0092-8674(00)80057-4)
- Barra, J., & Leucci, E. (2017). Probing Long Non-coding RNA-Protein Interactions. *Frontiers in Molecular Biosciences*, 4. <https://doi.org/10.3389/fmolb.2017.00045>
- Barreau, C., Paillard, L., & Osborne, H. B. (2005). AU-rich elements and associated factors: are there unifying principles? *Nucleic Acids Research*, 33(22), 7138–7150. <https://doi.org/10.1093/nar/gki1012>
- Bass, B. L. (2002). RNA Editing by Adenosine Deaminases That Act on RNA. *Annual Review of Biochemistry*, 71(1), 817–846. <https://doi.org/10.1146/annurev.biochem.71.110601.135501>
- Beckmann, B. M., Horos, R., Fischer, B., Castello, A., Eichelbaum, K., Alleaume, A.-M., ... Hentze, M. W. (2015). The RNA-binding proteomes from yeast to man harbour conserved enigmRBPs. *Nature Communications*, 6, 10127. <https://doi.org/10.1038/ncomms10127>
- Benos, P. V., Lapedes, A. S., & Stormo, G. D. (2002). Probabilistic code for DNA recognition by proteins of the EGR family. *Journal of Molecular Biology*, 323(4), 701–727. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/12419259>
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., ... Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1), 235–242. <https://doi.org/10.1093/nar/28.1.235>
- Beyer, A. L., Christensen, M. E., Walker, B. W., & LeSturgeon, W. M. (1977). Identification and characterization of the packaging proteins of core 40S hnRNP particles. *Cell*, 11(1), 127–138. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/872217>
- Bissantz, C., Folkers, G., & Rognan, D. (2000). Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *Journal of Medicinal Chemistry*, 43(25), 4759–4767. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/11123984>
- Bitetti-putzer, R., Joseph-mccarthy, D., Hogle, J. M., & Karplus, M. (2001). Functional group placement in protein binding sites : a comparison of GRID and MCSS. *Biological Chemistry*, 935–960.
- Böhm, H. J. (1992). The computer program LUDI: a new method for the de novo design of enzyme inhibitors. *Journal of Computer-Aided Molecular Design*, 6(1), 61–78. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/1583540>

- Braun, I. C., Herold, A., Rode, M., & Izaurralde, E. (2002). Nuclear export of mRNA by TAP/NXF1 requires two nucleoporin-binding sites but not p15. *Molecular and Cellular Biology*, 22(15), 5405–5418. <https://doi.org/10.1128/MCB.22.15.5405>
- Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., & Karplus, M. (1983). CHARMM A program for macromolecular energy, minimization, and dynamics calculations, *Journal of Computational Chemistry* Volume 4, Issue 2. *J Comput Chem*, 4(2), 187–217. Retrieved from [http://onlinelibrary.wiley.com/doi/10.1002/jcc.540040211/abstract%5Cnhttp://onlinelibrary.wiley.com.ezlibproxy1.ntu.edu.sg/store/10.1002/jcc.540040211/asset/540040211\\_ftp.pdf?v=1&t=hx5y0x9g&s=1e40e6d074660054d725a0cd04404a6a27230a9d](http://onlinelibrary.wiley.com/doi/10.1002/jcc.540040211/abstract%5Cnhttp://onlinelibrary.wiley.com.ezlibproxy1.ntu.edu.sg/store/10.1002/jcc.540040211/asset/540040211_ftp.pdf?v=1&t=hx5y0x9g&s=1e40e6d074660054d725a0cd04404a6a27230a9d)
- Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., SWAMINATHAN, S., & Karplus, M. (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4(2), 187–217. <https://doi.org/10.1002/jcc.540040211>
- Caflisch, a. (1996). Computational combinatorial ligand design: application to human alpha-thrombin. *Journal of Computer-Aided Molecular Design*, 10(5), 372–396. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/8951649>
- Campbell, Z. T., Bhimsaria, D., Valley, C. T., Rodriguez-Martinez, J. A., Menichelli, E., Williamson, J. R., ... Wickens, M. (2012). Cooperativity in RNA-protein interactions: global analysis of RNA binding specificity. *Cell Reports*, 1(5), 570–581. <https://doi.org/10.1016/j.celrep.2012.04.003>
- Chaput, L., & Mouawad, L. (2017). Efficient conformational sampling and weak scoring in docking programs? Strategy of the wisdom of crowds. *Journal of Cheminformatics*, 9(1), 37. <https://doi.org/10.1186/s13321-017-0227-x>
- Charifson, P. S., Corkery, J. J., Murcko, M. A., & Walters, W. P. (1999). Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *Journal of Medicinal Chemistry*, 42(25), 5100–5109. <https://doi.org/10.1021/jm990352k>
- Chaudhury, A., Chander, P., & Howe, P. H. (2010). Heterogeneous nuclear ribonucleoproteins (hnRNPs) in cellular processes: Focus on hnRNP E1's multifunctional regulatory roles. *RNA (New York, N.Y.)*, 16(8), 1449–1462. <https://doi.org/10.1261/rna.2254110>
- Chauvot de Beauchene, I., de Vries, S. J., & Zacharias, M. (2016a). Binding Site Identification and Flexible Docking of Single Stranded RNA to Proteins Using a Fragment-Based Approach. *PLoS Computational Biology*, 12(1), e1004697. <https://doi.org/10.1371/journal.pcbi.1004697>
- Chauvot de Beauchene, I., de Vries, S. J., & Zacharias, M. (2016b). Binding Site Identification and Flexible Docking of Single Stranded RNA to Proteins Using a Fragment-Based Approach. *PLoS Computational Biology*, 12(1), 1–21. <https://doi.org/10.1371/journal.pcbi.1004697>

- Chen, C. Y., Gherzi, R., Ong, S. E., Chan, E. L., Raijmakers, R., Pruijn, G. J., ... Karin, M. (2001). AU binding proteins recruit the exosome to degrade ARE-containing mRNAs. *Cell*, *107*(4), 451–464. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/11719186>
- Chen, Z. G., Stauffacher, C., Li, Y., Schmidt, T., Bomu, W., Kamer, G., ... Johnson, J. E. (1989). Protein-RNA interactions in an icosahedral virus at 3.0 Å resolution. *Science (New York, N.Y.)*, *245*(4914), 154–159. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/2749253>
- Chojnowski, G., Walen, T., & Bujnicki, J. M. (2014). RNA Bricks--a database of RNA 3D motifs and their interactions. *Nucleic Acids Research*, *42*(Database issue), D123-31. <https://doi.org/10.1093/nar/gkt1084>
- Cirillo, D., Agostini, F., & Tartaglia, G. G. (2013). Predictions of protein-RNA interactions. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *3*(2), 161–175. <https://doi.org/10.1002/wcms.1119>
- Claußen, H., Buning, C., Rarey, M., & Lengauer, T. (2001). FLEXE: Efficient molecular docking considering protein structure variations. *Journal of Molecular Biology*, *308*(2), 377–395. <https://doi.org/10.1006/jmbi.2001.4551>
- Cléry, A., & Allain, F. H.-T. (2013). *FROM STRUCTURE TO FUNCTION OF RNA BINDING DOMAINS*. Retrieved from <https://www.ncbi.nlm.nih.gov/books/NBK63528/>
- Cléry, A., Blatter, M., & Allain, F. H. T. (2008). RNA recognition motifs: boring? Not quite. *Current Opinion in Structural Biology*, *18*(3), 290–298. <https://doi.org/10.1016/j.sbi.2008.04.002>
- Cole, C. N., & Scarcelli, J. J. (2006). Transport of messenger RNA from the nucleus to the cytoplasm. *Current Opinion in Cell Biology*, *18*(3), 299–306. <https://doi.org/10.1016/J.CEB.2006.04.006>
- Colman, D. R., Kreibich, G., Frey, A. B., & Sabatini, D. D. (1982). Synthesis and incorporation of myelin polypeptides into CNS myelin. *The Journal of Cell Biology*, *95*(2 Pt 1), 598–608. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/6183276>
- Comeau, S. R., Gatchell, D. W., Vajda, S., & Camacho, C. J. (2004). ClusPro: An automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics*, *20*(1), 45–50. <https://doi.org/10.1093/bioinformatics/btg371>
- Cook, K. B., Hughes, T. R., & Morris, Q. D. (2015). High-throughput characterization of protein-RNA interactions. *Briefings in Functional Genomics*, *14*(1), 74–89. <https://doi.org/10.1093/bfgp/elu047>
- Cook, K. B., Kazan, H., Zuberi, K., Morris, Q., & Hughes, T. R. (2010). RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gkq1069>
- Corcoran, D. L., Georgiev, S., Mukherjee, N., Gottwein, E., Skalsky, R. L., Keene, J. D., & Ohler, U. (2011). PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biology*, *12*(8), R79. <https://doi.org/10.1186/gb-2011-12-8-r79>

- Curinha, A., Oliveira Braz, S., Pereira-Castro, I., Cruz, A., & Moreira, A. (2014). Implications of polyadenylation in health and disease. *Nucleus (Austin, Tex.)*, 5(6), 508–519. <https://doi.org/10.4161/nucl.36360>
- Davis, A. M., & Teague, S. J. (1999). Hydrogen Bonding, Hydrophobic Interactions, and Failure of the Rigid Receptor Hypothesis. *Angewandte Chemie International Edition*, 38(6), 736–749. [https://doi.org/10.1002/\(SICI\)1521-3773\(19990315\)38:6<736::AID-ANIE736>3.0.CO;2-R](https://doi.org/10.1002/(SICI)1521-3773(19990315)38:6<736::AID-ANIE736>3.0.CO;2-R)
- Davis, M. E., & McCammon, J. A. (1989). Solving the finite difference linearized Poisson-Boltzmann equation: A comparison of relaxation and conjugate gradient methods. *Journal of Computational Chemistry*. <https://doi.org/10.1002/jcc.540100313>
- De Beauchene, I. C., De Vries, S. J., & Zacharias, M. (2016). Fragment-based modelling of single stranded RNA bound to RNA recognition motif containing proteins. *Nucleic Acids Research*, 44(10), 4565–4580. <https://doi.org/10.1093/nar/gkw328>
- DesJarlais, R. L., Kuntz, I. D., Sheridan, R. P., Venkataraghavan, R., & Dixon, J. S. (1986). Docking Flexible Ligands to Macromolecular Receptors by Molecular Shape. *Journal of Medicinal Chemistry*, 29(11), 2149–2153. <https://doi.org/10.1021/jm00161a004>
- Dill, K. A. (1997). Additivity principles in biochemistry. *Journal of Biological Chemistry*, 272(2), 701–704. <https://doi.org/10.1074/jbc.272.2.701>
- Ding, Y., Tang, Y., Kwok, C. K., Zhang, Y., Bevilacqua, P. C., & Assmann, S. M. (2014). In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature*, 505(7485), 696–700. <https://doi.org/10.1038/nature12756>
- Dominguez, D., Freese, P., Alexis, M. S., Su, A., Hochman, M., Palden, T., ... Burge, C. B. (2018). Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. *Molecular Cell*, 70(5), 854-867.e9. <https://doi.org/10.1016/j.molcel.2018.05.001>
- Dreyfuss, G., Kim, V. N., & Kataoka, N. (2002). Messenger-RNA-binding proteins and the messages they carry. *Nature Reviews Molecular Cell Biology*, 3(3), 195–205. <https://doi.org/10.1038/nrm760>
- Dreyfuss, G., Matunis, M. J., Pinol-Roma, S., & Burd, C. G. (1993). hnRNP Proteins and the Biogenesis of mRNA. *Annual Review of Biochemistry*, 62(1), 289–321. <https://doi.org/10.1146/annurev.bi.62.070193.001445>
- Drosphila, R. M., Lu, D., Searles, M. A., & Klug, A. (2003). *Crystal structure of a zinc-finger-.pdf*. 426(November), 471–475.
- Dunbar, J. B., Smith, R. D., Yang, C.-Y., Ung, P. M.-U., Lexa, K. W., Khazanov, N. A., ... Carlson, H. A. (2011). CSAR Benchmark Exercise of 2010: Selection of the Protein–Ligand Complexes. *Journal of Chemical Information and Modeling*, 51(9), 2036–2046. <https://doi.org/10.1021/ci200082t>

- Dunbrack, R. L., & Karplus, M. (1993). Backbone-dependent rotamer library for proteins: Application to side-chain prediction. *Journal of Molecular Biology*, Vol. 230, pp. 543–574. <https://doi.org/10.1006/jmbi.1993.1170>
- Durrant, J. D., & McCammon, J. A. (2011). BINANA: a novel algorithm for ligand-binding characterization. *Journal of Molecular Graphics & Modelling*, 29(6), 888–893. <https://doi.org/10.1016/j.jmgm.2011.01.004>
- Eaton, H. L., & Wyss, D. F. (2011). Effective Progression of Nuclear Magnetic Resonance-Detected Fragment Hits. In *Methods in Enzymology*. <https://doi.org/10.1016/B978-0-12-381274-2.00017-0>
- Eisen, M. B., Wiley, D. C., Karplus, M., & Hubbard, R. E. (1994). HOOK: A program for finding novel molecular architectures that satisfy the chemical and steric requirements of a macromolecule binding site. *Proteins: Structure, Function, and Bioinformatics*, 19(3), 199–221. <https://doi.org/10.1002/prot.340190305>
- Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V., & Mee, R. P. (1997). Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *Journal of Computer-Aided Molecular Design*, 11(5), 425–445. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/9385547>
- Ellington, A. D., & Szostak, J. W. (1990). In vitro selection of RNA molecules that bind specific ligands. *Nature*, 346(6287), 818–822. <https://doi.org/10.1038/346818a0>
- Ericksen, S. S., Wu, H., Zhang, H., Michael, L. A., Newton, M. A., Hoffmann, F. M., & Wildman, S. A. (2017). Machine Learning Consensus Scoring Improves Performance Across Targets in Structure-Based Virtual Screening. *Journal of Chemical Information and Modeling*, 57(7), 1579–1590. <https://doi.org/10.1021/acs.jcim.7b00153>
- Erlanson, D. A. (2006). Fragment-based lead discovery: a chemical update. *Current Opinion in Biotechnology*. <https://doi.org/10.1016/j.copbio.2006.10.007>
- Evensen, E., Joseph-McCarthy, D., Weiss, G. A., Schreiber, S. L., & Karplus, M. (2007). Ligand design by a combinatorial approach based on modeling and experiment: application to HLA-DR4. *Journal of Computer-Aided Molecular Design*, 21(7), 395–418. <https://doi.org/10.1007/s10822-007-9119-x>
- Fernández-Recio, J., Totrov, M., & Abagyan, R. (2004). Identification of Protein–Protein Interaction Sites from Docking Energy Landscapes. *Journal of Molecular Biology*, 335(3), 843–865. <https://doi.org/10.1016/j.jmb.2003.10.069>
- Fleishman, S. J., Corn, J. E., Strauch, E. M., Whitehead, T. A., Andre, I., Thompson, J., ... Baker, D. (2010). Rosetta in CAPRI rounds 13–19. *Proteins: Structure, Function, and Bioinformatics*, 78(15), 3212–3218. <https://doi.org/10.1002/prot.22784>
- Fornes, O., Garcia-Garcia, J., Bonet, J., & Oliva, B. (2014). On the Use of Knowledge-Based Potentials for the Evaluation of Models of Protein–Protein, Protein–DNA, and Protein–RNA

Interactions. *Advances in Protein Chemistry and Structural Biology*, 94, 77–120.

<https://doi.org/10.1016/B978-0-12-800168-4.00004-4>

Friedersdorf, M. B., & Keene, J. D. (2014). Advancing the functional utility of PAR-CLIP by quantifying background binding to mRNAs and lncRNAs. *Genome Biology*, 15(1), R2.

<https://doi.org/10.1186/gb-2014-15-1-r2>

Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., ... Shenkin, P. S. (2004). Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *Journal of Medicinal Chemistry*, 47(7), 1739–1749.

<https://doi.org/10.1021/jm0306430>

Friesner, R. A., Murphy, R. B., Repasky, M. P., Frye, L. L., Greenwood, J. R., Halgren, T. A., ... Mainz, D. T. (2006). Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein–Ligand Complexes. *Journal of Medicinal Chemistry*, 49(21), 6177–6196.

<https://doi.org/10.1021/jm051256o>

Furukawa, Y., Suzuki, Y., Fukuoka, M., Nagasawa, K., Nakagome, K., Shimizu, H., ... Akiyama, S. (2016). A molecular mechanism realizing sequence-specific recognition of nucleic acids by TDP-43. *Scientific Reports*, 6(1), 20576.

<https://doi.org/10.1038/srep20576>

Gabel, J., Desaphy, J., & Rognan, D. (2014). Beware of machine learning-based scoring functions—on the danger of developing black boxes. *Journal of Chemical Information and Modeling*, 54(10), 2807–2815.

<https://doi.org/10.1021/ci500406k>

Gao, Q. Q., Putzbach, W. E., Murmann, A. E., Chen, S., Sarshad, A. A., Peter, J. M., ... Peter, M. E. (2018). 6mer seed toxicity in tumor suppressive microRNAs. *Nature Communications*, 9(1), 4504.

<https://doi.org/10.1038/s41467-018-06526-1>

Garneau, N. L., Wilusz, J., & Wilusz, C. J. (2007). The highways and byways of mRNA decay.

*Nature Reviews Molecular Cell Biology*, 8(2), 113–126. <https://doi.org/10.1038/nrm2104>

Gerstberger, S., Hafner, M., Ascano, M., & Tuschl, T. (2014). Evolutionary conservation and expression of human RNA-binding proteins and their role in human genetic disease. *Advances in Experimental Medicine and Biology*, 825, 1–55.

[https://doi.org/10.1007/978-1-4939-1221-6\\_1](https://doi.org/10.1007/978-1-4939-1221-6_1)

Gerstberger, S., Hafner, M., & Tuschl, T. (2014). A census of human RNA-binding proteins. *Nature Publishing Group*, (November).

<https://doi.org/10.1038/nrg3813>

Geuens, T., Bouhy, D., & Timmerman, V. (2016). The hnRNP family: insights into their role in health and disease. *Human Genetics*, 135(8), 851–867.

<https://doi.org/10.1007/s00439-016-1683-5>

Ghosh, A., Rapp, C. S., & Friesner, R. A. (2002). Generalized Born Model Based on a Surface Integral Formulation. *The Journal of Physical Chemistry B*.

<https://doi.org/10.1021/jp982533o>

- Giudice, G., Sánchez-Cabo, F., Torroja, C., & Lara-Pezzi, E. (2016). ATtRACT—a database of RNA-binding proteins and associated motifs. *Database*, 2016, baw035. <https://doi.org/10.1093/database/baw035>
- Goodford, P. J. (1985). A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *Journal of Medicinal Chemistry*, 28(7), 849–857. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/3892003>
- Goodsell, D. S., & Olson, A. J. (1990). Automated docking of substrates to proteins by simulated annealing. *Proteins: Structure, Function, and Bioinformatics*, 8(3), 195–202. <https://doi.org/10.1002/prot.340080302>
- Grauffel, C., Stote, R. H., & Dejaegere, A. (2010). Force field parameters for the simulation of modified histone tails. *Journal of Computational Chemistry*, 31(13), 2434–2451. <https://doi.org/10.1002/jcc.21536>
- Grishin, N. V. (2001). KH domain: one motif, two folds. *Nucleic Acids Research*, 29(3), 638–643. <https://doi.org/10.1093/nar/29.3.638>
- Haider, M. K., Bertrand, H. O., & Hubbard, R. E. (2011). Predicting fragment binding poses using a combined MCSS MM-GBSA approach. *Journal of Chemical Information and Modeling*, 51(5), 1092–1105. <https://doi.org/10.1021/ci100469n>
- Hajduk, P. J., & Greer, J. (2007). A decade of fragment-based drug design: Strategic advances and lessons learned. *Nature Reviews Drug Discovery*. <https://doi.org/10.1038/nrd2220>
- Hajduk, P. J., & Sauer, D. R. (2008). Statistical Analysis of the Effects of Common Chemical Substituents on Ligand Potency. *Journal of Medicinal Chemistry*, 51(3), 553–564. <https://doi.org/10.1021/jm070838y>
- Hall, D., Li, S., Yamashita, K., Azuma, R., Carver, J. A., & Standley, D. M. (2014). A novel protein distance matrix based on the minimum arc-length between two amino-acid residues on the surface of a globular protein. *Biophysical Chemistry*, 190–191, 50–55. <https://doi.org/10.1016/J.BPC.2014.01.005>
- Hall, D., Li, S., Yamashita, K., Azuma, R., Carver, J. A., & Standley, D. M. (2015). RNA-LIM: a novel procedure for analyzing protein/single-stranded RNA propensity data with concomitant estimation of interface structure. *Analytical Biochemistry*, 472, 52–61. <https://doi.org/10.1016/j.ab.2014.11.004>
- Hartshorn, M. J., Verdonk, M. L., Chessari, G., Brewerton, S. C., Mooij, W. T. M., Mortenson, P. N., & Murray, C. W. (2007). Diverse, high-quality test set for the validation of protein-ligand docking performance. *Journal of Medicinal Chemistry*, 50(4), 726–741. <https://doi.org/10.1021/jm061277y>
- Helder, S., Blythe, A. J., Bond, C. S., & Mackay, J. P. (2016). Determinants of affinity and specificity in RNA-binding proteins. *Current Opinion in Structural Biology*, 38, 83–91. <https://doi.org/10.1016/j.sbi.2016.05.005>



- Hennig, J., & Sattler, M. (2015). Deciphering the protein-RNA recognition code: Combining large-scale quantitative methods with structural biology. *BioEssays*.  
<https://doi.org/10.1002/bies.201500033>
- Hentze, M. W., Castello, A., Schwarzl, T., & Preiss, T. (2018). A brave new world of RNA-binding proteins. *Nature Reviews Molecular Cell Biology*, 19(5), 327–341.  
<https://doi.org/10.1038/nrm.2017.130>
- Hollingworth, D., Candel, A. M., Nicastrò, G., Martin, S. R., Briata, P., Gherzi, R., & Ramos, A. (2012). KH domains with impaired nucleic acid binding as a tool for functional analysis. *Nucleic Acids Research*, 40(14), 6873–6886. <https://doi.org/10.1093/nar/gks368>
- Hoque, M., Ji, Z., Zheng, D., Luo, W., Li, W., You, B., ... Tian, B. (2013). Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing. *Nature Methods*, 10(2), 133–139. <https://doi.org/10.1038/nmeth.2288>
- Houk, K. N., Leach, A. G., Kim, S. P., & Zhang, X. (2003). Binding affinities of host-guest, protein-ligand, and protein-transition-state complexes. *Angewandte Chemie (International Ed. in English)*, 42(40), 4872–4897. <https://doi.org/10.1002/anie.200200565>
- Hu, B., Yang, Y.-C. T., Huang, Y., Zhu, Y., & Lu, Z. J. (2017). POSTAR: a platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. *Nucleic Acids Research*, 45(D1), D104–D114. <https://doi.org/10.1093/nar/gkw888>
- Huang, R., Han, M., Meng, L., & Chen, X. (2018). Transcriptome-wide discovery of coding and noncoding RNA-binding proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 115(17), E3879. <https://doi.org/10.1073/PNAS.1718406115>
- Huang, S.-Y., & Zou, X. (2014a). A knowledge-based scoring function for protein-RNA interactions derived from a statistical mechanics-based iterative method. *Nucleic Acids Research*, 42(7), e55. <https://doi.org/10.1093/nar/gku077>
- Huang, S.-Y., & Zou, X. (2014b). A knowledge-based scoring function for protein-RNA interactions derived from a statistical mechanics-based iterative method. *Nucleic Acids Research*, 42(7), e55. <https://doi.org/10.1093/nar/gku077>
- Huang, S. Y., & Zou, X. (2007). Ensemble docking of multiple protein structures: Considering protein structural variations in molecular docking. *Proteins: Structure, Function and Genetics*, 66(2), 399–421. <https://doi.org/10.1002/prot.21214>
- Huang, S. Y., & Zou, X. (2008). An iterative knowledge-based scoring function for protein-protein recognition. *Proteins: Structure, Function and Genetics*, 72(2), 557–579.  
<https://doi.org/10.1002/prot.21949>
- Huang, S. Y., & Zou, X. (2013). A nonredundant structure dataset for benchmarking protein-RNA computational docking. *Journal of Computational Chemistry*, 34(4), 311–318.  
<https://doi.org/10.1002/jcc.23149>

- Ji, N., Ostroverkhov, V., Tian, C. S., & Shen, Y. R. (2008). Characterization of Vibrational Resonances of Water-Vapor Interfaces by Phase-Sensitive Sum-Frequency Spectroscopy. *Physical Review Letters*, *100*(9), 096102. <https://doi.org/10.1103/PhysRevLett.100.096102>
- Jiang, F., & Kim, S. H. (1991). "Soft docking": Matching of molecular surface cubes. *Journal of Molecular Biology*, *219*(1), 79–102. [https://doi.org/10.1016/0022-2836\(91\)90859-5](https://doi.org/10.1016/0022-2836(91)90859-5)
- Jo, S., Kim, T., Iyer, V. G., & Im, W. (2008). CHARMM-GUI: A web-based graphical user interface for CHARMM. *Journal of Computational Chemistry*, *29*(11), 1859–1865. <https://doi.org/10.1002/jcc.20945>
- Jones, G., Willett, P., Glen, R. C., Leach, A. R., & Taylor, R. (1997). Development and validation of a genetic algorithm for flexible docking 1 Edited by F. E. Cohen. *Journal of Molecular Biology*, *267*(3), 727–748. <https://doi.org/10.1006/jmbi.1996.0897>
- Jones, J. E. and S. J. (2008). *Evaluating conformational changes in protein structures binding RNA*. (March), 498–508. <https://doi.org/10.1002/prot>
- Jones, S. (2016). Protein–RNA interactions: structural biology and computational modeling techniques. *Biophysical Reviews*, *8*(4), 359–367. <https://doi.org/10.1007/s12551-016-0223-9>
- Jonikas, M. A., Radmer, R. J., Laederach, A., Das, R., Pearlman, S., Herschlag, D., & Altman, R. B. (2009). Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA (New York, N.Y.)*, *15*(2), 189–199. <https://doi.org/10.1261/rna.1270809>
- Joseph-McCarthy, D. (1999). Computational approaches to structure-based ligand design. *Pharmacology & Therapeutics*, *84*(2), 179–191. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/10596905>
- Kahvejian, A., Roy, G., & Sonenberg, N. (2001). The mRNA closed-loop model: The function of PABP and PABP-interacting proteins in mRNA translation. *Cold Spring Harbor Symposia on Quantitative Biology*, *66*, 293–300. <https://doi.org/10.1101/sqb.2001.66.293>
- Källblad, P., & Dean, P. M. (2003). Efficient conformational sampling of local side-chain flexibility. *Journal of Molecular Biology*, *326*(5), 1651–1665. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/12595271>
- Kappel, K., & Das, R. (2019). Sampling Native-like Structures of RNA-Protein Complexes through Rosetta Folding and Docking. *Structure*, *27*(1), 140-151.e5. <https://doi.org/10.1016/j.str.2018.10.001>
- Kaserer, T., Temml, V., Kutil, Z., Vanek, T., Landa, P., & Schuster, D. (2015). Prospective performance evaluation of selected common virtual screening tools. Case study: Cyclooxygenase (COX) 1 and 2. *European Journal of Medicinal Chemistry*, *96*, 445–457. <https://doi.org/10.1016/j.ejmech.2015.04.017>
- Keegan, L. P., Gallo, A., & O'Connell, M. A. (2001). The many roles of an RNA editor. *Nature Reviews Genetics*, *2*(11), 869–878. <https://doi.org/10.1038/35098584>

- Knegtel, R. M. A., Kuntz, I. D., & Oshiro, C. M. (1997). Molecular docking to ensembles of protein structures. *Journal of Molecular Biology*, 266(2), 424–440. <https://doi.org/10.1006/jmbi.1996.0776>
- Kobren, S. N., & Singh, M. (2018). Systematic domain-based aggregation of protein structures highlights DNA-, RNA-and other ligand-binding positions. *Nucleic Acids Research*, (1). <https://doi.org/10.1093/nar/gky1224>
- Koes, D. R., Baumgartner, M. P., & Camacho, C. J. (2013). Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *Journal of Chemical Information and Modeling*, 53(8), 1893–1904. <https://doi.org/10.1021/ci300604z>
- Kozakov, D., Clodfelter, K. H., Vajda, S., & Camacho, C. J. (2005). Optimal Clustering for Detecting Near-Native Conformations in Protein Docking. *Biophysical Journal*, 89(2), 867–875. <https://doi.org/10.1529/biophysj.104.058768>
- Kramer, K., Sachsenberg, T., Beckmann, B. M., Qamar, S., Boon, K.-L., Hentze, M. W., ... Urlaub, H. (2014). Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins. *Nature Methods*, 11(10), 1064–1070. <https://doi.org/10.1038/nmeth.3092>
- Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R., & Ferrin, T. E. (1982). A geometric approach to macromolecule-ligand interactions. *Journal of Molecular Biology*, 161(2), 269–288. [https://doi.org/10.1016/0022-2836\(82\)90153-x](https://doi.org/10.1016/0022-2836(82)90153-x)
- Lal, A., Mazan-Mamczarz, K., Kawai, T., Yang, X., Martindale, J. L., & Gorospe, M. (2004). Concurrent versus individual binding of HuR and AUF1 to common labile target mRNAs. *The EMBO Journal*, 23(15), 3092–3102. <https://doi.org/10.1038/sj.emboj.7600305>
- Lambert, N., Robertson, A., Jangi, M., McGeary, S., Sharp, P. A., & Burge, C. B. (2014). RNA Bind-n-Seq: Quantitative Assessment of the Sequence and Structural Binding Specificity of RNA Binding Proteins. *Molecular Cell*, 54(5), 887–900. <https://doi.org/10.1016/j.molcel.2014.04.016>
- Lang, B., Armaos, A., & Tartaglia, G. G. (2019). RNAct: Protein-RNA interaction predictions for model organisms with supporting experimental data. *Nucleic Acids Research*, 47(D1), D601–D606. <https://doi.org/10.1093/nar/gky967>
- Leach, A. R. (1994). Ligand docking to proteins with discrete side-chain flexibility. *Journal of Molecular Biology*, 235(1), 345–356. [https://doi.org/10.1016/S0022-2836\(05\)80038-5](https://doi.org/10.1016/S0022-2836(05)80038-5)
- Leach, A. R., & Kuntz, I. D. (1992). Conformational analysis of flexible ligands in macromolecular receptor sites. *Journal of Computational Chemistry*, 13(6), 730–748. <https://doi.org/10.1002/jcc.540130608>
- Leclerc, F., & Karplus, M. (1999). MCSS-based predictions of RNA binding sites. *Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta)*, 101(1–3), 131–137. <https://doi.org/10.1007/s002140050419>

- Lensink, M. F., Velankar, S., Baek, M., Heo, L., Seok, C., & Wodak, S. J. (2018). The challenge of modeling protein assemblies: the CASP12-CAPRI experiment. *Proteins: Structure, Function, and Bioinformatics*, 86, 257–273. <https://doi.org/10.1002/prot.25419>
- Lensink, M. F., & Wodak, S. J. (2010). Docking and scoring protein interactions: CAPRI 2009. *Proteins*, 78(15), 3073–3084. <https://doi.org/10.1002/prot.22818>
- Leung, C. S., Leung, S. S. F., Tirado-Rives, J., & Jorgensen, W. L. (2012). Methyl effects on protein-ligand binding. *Journal of Medicinal Chemistry*, 55(9), 4489–4500. <https://doi.org/10.1021/jm3003697>
- Li, C. H., Cao, L. Bin, Su, J. G., Yang, Y. X., & Wang, C. X. (2012). A new residue-nucleotide propensity potential with structural information considered for discriminating protein-RNA docking decoys. *Proteins: Structure, Function, and Bioinformatics*, 80(1), 14–24. <https://doi.org/10.1002/prot.23117>
- Li, H., Leung, K. S., Wong, M. H., & Ballester, P. J. (2015). Improving autodock vina using random forest: The growing accuracy of binding affinity prediction by the effective exploitation of larger data sets. *Molecular Informatics*, 34(2–3), 115–126. <https://doi.org/10.1002/minf.201400132>
- Li, S., Yamashita, K., Amada, K. M., & Standley, D. M. (2014). Quantifying sequence and structural features of protein–RNA interactions. *Nucleic Acids Research*, 42(15), 10086. <https://doi.org/10.1093/NAR/GKU681>
- Li, X., Kazan, H., Lipshitz, H. D., & Morris, Q. D. (2014). Finding the target sites of RNA-binding proteins. *Wiley Interdisciplinary Reviews: RNA*, 5(1), 111–130. <https://doi.org/10.1002/wrna.1201>
- Li, Y., Han, L., Liu, Z., & Wang, R. (2014). Comparative Assessment of Scoring Functions on an Updated Benchmark: 2. Evaluation Methods and General Results. *Journal of Chemical Information and Modeling*, 54(6), 1717–1736. <https://doi.org/10.1021/ci500081m>
- Licatalosi, D. D., & Darnell, R. B. (2006). Splicing Regulation in Neurologic Disease. *Neuron*, 52(1), 93–101. <https://doi.org/10.1016/J.NEURON.2006.09.017>
- Liu, Z., Su, M., Han, L., Liu, J., Yang, Q., Li, Y., & Wang, R. (2017). Forging the Basis for Developing Protein-Ligand Interaction Scoring Functions. *Accounts of Chemical Research*, 50(2), 302–309. <https://doi.org/10.1021/acs.accounts.6b00491>
- Lu, X.-J., & Olson, W. K. (2008). 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nature Protocols* 2008 3:7, 3(7), 1213–1227. <https://doi.org/10.1038/nprot.2008.104>
- Lunde, B. M., Hörner, M., & Meinhart, A. (2011). Structural insights into cis element recognition of non-polyadenylated RNAs by the Nab3-RRM. *Nucleic Acids Research*, 39(1), 337–346. <https://doi.org/10.1093/nar/gkq751>

- Lunde, B. M., Moore, C., & Varani, G. (2007). RNA-binding proteins: modular design for efficient function. *Nat Rev Mol Cell Biol*, 8(6), 479–490. <https://doi.org/10.1038/nrm2178>
- Lykke-Andersen, J., & Wagner, E. (2005). Recruitment and activation of mRNA decay enzymes by two ARE-mediated decay activation domains in the proteins TTP and BRF-1. *Genes & Development*, 19(3), 351–361. <https://doi.org/10.1101/gad.1282305>
- Maniatis, T., & Reed, R. (2002). An extensive network of coupling among gene expression machines. *Nature*, 416(6880), 499–506. <https://doi.org/10.1038/416499a>
- Marchese, D., de Groot, N. S., Lorenzo Gotor, N., Livi, C. M., & Tartaglia, G. G. (2016a). Advances in the characterization of RNA-binding proteins. *Wiley Interdisciplinary Reviews: RNA*, 7(6), 793–810. <https://doi.org/10.1002/wrna.1378>
- Marchese, D., de Groot, N. S., Lorenzo Gotor, N., Livi, C. M., & Tartaglia, G. G. (2016b). Advances in the characterization of RNA-binding proteins. *Wiley Interdisciplinary Reviews: RNA*. <https://doi.org/10.1002/wrna.1378>
- Maris, C., Dominguez, C., & Allain, F. H.-T. (2005). The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *FEBS Journal*, 272(9), 2118–2131. <https://doi.org/10.1111/j.1742-4658.2005.04653.x>
- Matunis, E. L., Matunis, M. J., & Dreyfuss, G. (1992). Characterization of the major hnRNP proteins from *Drosophila melanogaster*. *Journal of Cell Biology*, 116(2), 257–269. <https://doi.org/10.1083/jcb.116.2.257>
- McHugh, C. A., Russell, P., & Guttman, M. (2014). Methods for comprehensive experimental identification of RNA-protein interactions. *Genome Biology*, 15(1), 203. <https://doi.org/10.1186/GB4152>
- Miao, Z., & Westhof, E. (2015). Prediction of nucleic acid binding probability in proteins: a neighboring residue network based score. *Nucleic Acids Research*, 43(11), 5340–5351. <https://doi.org/10.1093/nar/gkv446>
- Miranker, A., & Karplus, M. (1991). Functionality maps of binding sites: A multiple copy simultaneous search method. *Proteins: Structure, Function, and Bioinformatics*, 11(1), 29–34. <https://doi.org/10.1002/prot.340110104>
- Miyamura, Y., Suzuki, T., Kono, M., Inagaki, K., Ito, S., Suzuki, N., & Tomita, Y. (2003). Mutations of the RNA-specific adenosine deaminase gene (DSRAD) are involved in dyschromatosis symmetrica hereditaria. *American Journal of Human Genetics*, 73(3), 693–699. <https://doi.org/10.1086/378209>
- Moore, M. J. (2005). From birth to death: The complex lives of eukaryotic mRNAs. *Science*, 309(5740), 1514–1518. <https://doi.org/10.1126/science.1111443>
- Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., ... Al, M. E. T. (1998). <Using AutoDock.pdf>. 19(14), 1639–1662. <https://doi.org/10.1002/jcc.20634>

- Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., & Olson, A. J. (2009). AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *Journal of Computational Chemistry*, 30(16), 2785. <https://doi.org/10.1002/JCC.21256>
- Muegge, I., & Martin, Y. C. (1999). A General and Fast Scoring Function for Protein–Ligand Interactions: A Simplified Potential Approach. *Journal of Medicinal Chemistry*, 42(5), 791–804. <https://doi.org/10.1021/jm980536j>
- Munro, T. P., Magee, R. J., Kidd, G. J., Carson, J. H., Barbarese, E., Smith, L. M., & Smith, R. (1999). Mutational analysis of a heterogeneous nuclear ribonucleoprotein A2 response element for RNA trafficking. *The Journal of Biological Chemistry*, 274(48), 34389–34395. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/10567417>
- Murn, J., Teplova, M., Zarnack, K., Shi, Y., & Patel, D. J. (2016). Recognition of distinct RNA motifs by the clustered CCCH zinc fingers of neuronal protein Unkempt. *Nature Structural & Molecular Biology*, 23(1), 16. <https://doi.org/10.1038/NSMB.3140>
- Murray, C. W., & Rees, D. C. (2009). The rise of fragment-based drug discovery. *Nature Chemistry*. <https://doi.org/10.1038/nchem.217>
- Nilsen, T. W., & Graveley, B. R. (2010). Expansion of the eukaryotic proteome by alternative splicing. *Nature*, 463(7280), 457–463. <https://doi.org/10.1038/nature08909>
- Noel T. Southall, †, Ken A. Dill, †,‡ and, & Haymet\*, A. D. J. (2001). *A View of the Hydrophobic Effect*. <https://doi.org/10.1021/JP015514E>
- Oberstrass, F. C., Auweter, S. D., Erat, M., Hargous, Y., Henning, A., Wenter, P., ... Allain, F. H. T. (2005). Structural biology - Structure of PTB bound to RNA: Specific binding and implications for splicing regulation. *Science*, 309(5743), 2054–2057. <https://doi.org/10.1126/science.1114066>
- Pan, Q., Shai, O., Lee, L. J., Frey, B. J., & Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, 40(12), 1413–1415. <https://doi.org/10.1038/ng.259>
- Pan, X., Rijnbeek, P., Yan, J., & Shen, H.-B. (2018). Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. *BMC Genomics*, 19(1), 511. <https://doi.org/10.1186/s12864-018-4889-1>
- Pedro, L., & Quinn, R. J. (2016). Native mass spectrometry in fragment-based drug discovery. *Molecules*. <https://doi.org/10.3390/molecules21080984>
- Pelham, H. R. B. (1980). *BPPR- Inc 2-2011.pdf*. 77(7), 4170–4174.
- Pemberton, L. F., & Paschal, B. M. (2005). Mechanisms of receptor-mediated nuclear import and nuclear export. *Traffic*, 6(3), 187–198. <https://doi.org/10.1111/j.1600-0854.2005.00270.x>

- Pérez-Cano, L., & Fernández-Recio, J. (2010). Optimal protein-RNA area, OPRA: A propensity-based method to identify RNA-binding sites on proteins. *Proteins: Structure, Function, and Bioinformatics*, 78(1), 25–35. <https://doi.org/10.1002/prot.22527>
- Pons, C., Grosdidier, S., Solernou, A., Pérez-Cano, L., & Fernández-Recio, J. (2010). Present and future challenges and limitations in protein-protein docking. *Proteins: Structure, Function, and Bioinformatics*, 78(1), 95–108. <https://doi.org/10.1002/prot.22564>
- Qamar, S., Kramer, K., & Urlaub, H. (2015). Studying RNA–Protein Interactions of Pre-mRNA Complexes by Mass Spectrometry. In *Methods in enzymology* (Vol. 558, pp. 417–463). <https://doi.org/10.1016/bs.mie.2015.02.010>
- Quiroga, R., & Villarreal, M. A. (2016). Vinardo: A Scoring Function Based on Autodock Vina Improves Scoring, Docking, and Virtual Screening. *PloS One*, 11(5), e0155183. <https://doi.org/10.1371/journal.pone.0155183>
- Rabani, M., Kertesz, M., & Segal, E. (2008). Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes. *Proceedings of the National Academy of Sciences of the United States of America*, 105(39), 14885–14890. <https://doi.org/10.1073/pnas.0803169105>
- Ramanathan, M., Porter, D. F., & Khavari, P. A. (2019). Methods to study RNA–protein interactions. *Nature Methods*, 16(3), 225–234. <https://doi.org/10.1038/s41592-019-0330-1>
- Ramos, A., Gru, S., Adams, J., Micklem, D. R., Proctor, M. R., Freund, S., ... Varani, G. (2000). RNA recognition by a Staufen double-stranded RNA-binding domain | *The EMBO Journal*. 19(5), 997–1009. Retrieved from <http://emboj.embopress.org/content/19/5/997.long>
- Rastelli, G., Degliesposti, G., Del Rio, A., & Sgobba, M. (2009). Binding Estimation after Refinement, a New Automated Procedure for the Refinement and Rescoring of Docked Ligands in Virtual Screening. *Chemical Biology & Drug Design*, 73(3), 283–286. <https://doi.org/10.1111/j.1747-0285.2009.00780.x>
- Ray, D., Kazan, H., Chan, E. T., Castillo, L. P., Chaudhry, S., Talukder, S., ... Hughes, T. R. (2009). Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nature Biotechnology*, 27(7), 667–670. <https://doi.org/10.1038/nbt.1550>
- Ray, D., Kazan, H., Cook, K. B., Weirauch, M. T., Najafabadi, H. S., Li, X., ... Hughes, T. R. (2013). A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499(7457), 172–177. <https://doi.org/10.1038/nature12311>
- Rentzsch, R., & Renard, B. Y. (2015). Docking small peptides remains a great challenge: an assessment using AutoDock Vina. *Briefings in Bioinformatics*, 16(6), 1045–1056. <https://doi.org/10.1093/bib/bbv008>
- Richard, P., & Manley, J. L. (2009). Transcription termination by nuclear RNA polymerases. *Genes & Development*, 23(11), 1247–1269. <https://doi.org/10.1101/gad.1792809>

- Richmond, G. (2001). S TRUCTURE AND B ONDING OF M OLECULES AT A QUEOUS S URFACES. *Annual Review of Physical Chemistry*, 52(1), 357–389. <https://doi.org/10.1146/annurev.physchem.52.1.357>
- Ripin, N., Boudet, J., Duszczuk, M. M., Hinniger, A., Faller, M., Krepl, M., ... Allain, F. H.-T. (2019). Molecular basis for AU-rich element recognition and dimerization by the HuR C-terminal RRM. *Proceedings of the National Academy of Sciences*, 116(8), 2935–2944. <https://doi.org/10.1073/pnas.1808696116>
- Rodriguez, M. S., Dargemont, C., & Stutz, F. (2004). Nuclear export of RNA. *Biology of the Cell*, 96(8), 639–655. <https://doi.org/10.1016/j.biolcel.2004.04.014>
- Rotstein, S. H., & Murcko, M. A. (1993). GroupBuild: a fragment-based method for de novo drug design. *Journal of Medicinal Chemistry*, 36(12), 1700–1710. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/8510098>
- Rouskin, S., Zubradt, M., Washietl, S., Kellis, M., & Weissman, J. S. (2014). Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature*, 505(7485), 701–705. <https://doi.org/10.1038/nature12894>
- Schmidt, M. F., & Rademann, J. (2009). Dynamic template-assisted strategies in fragment-based drug discovery. *Trends in Biotechnology*. <https://doi.org/10.1016/j.tibtech.2009.06.001>
- Schneider, G., & Fechner, U. (2005). Computer-based de novo design of drug-like molecules. *Nature Reviews Drug Discovery*. <https://doi.org/10.1038/nrd1799>
- Shi, Y., Di Giammartino, D. C., Taylor, D., Sarkeshik, A., Rice, W. J., Yates, J. R., ... Manley, J. L. (2009). Molecular Architecture of the Human Pre-mRNA 3' Processing Complex. *Molecular Cell*, 33(3), 365–376. <https://doi.org/10.1016/j.molcel.2008.12.028>
- Shuker, S. B., Hajduk, P. J., Meadows, R. P., & Fesik, S. W. (1996). Discovering high-affinity ligands for proteins: SAR by NMR. *Science*, 274(5292), 1531–1534. Retrieved from <http://www.sciencemag.org/content/274/5292/1531.long>
- Si, J., Cui, J., Cheng, J., & Wu, R. (2015a). Computational Prediction of RNA-Binding Proteins and Binding Sites. *International Journal of Molecular Sciences*, 16(11), 26303–26317. <https://doi.org/10.3390/ijms161125952>
- Si, J., Cui, J., Cheng, J., & Wu, R. (2015b). Computational Prediction of RNA-Binding Proteins and Binding Sites. *International Journal of Molecular Sciences*, 16(11), 26303–26317. <https://doi.org/10.3390/ijms161125952>
- Singh, R., & Valcárcel, J. (2005). Building specificity with nonspecific RNA-binding proteins. *Nature Structural & Molecular Biology*, 12(8), 645–653. <https://doi.org/10.1038/nsmb961>
- Skrisovska, L., Bourgeois, C. F., Stefl, R., Grellscheid, S. N., Kister, L., Wenter, P., ... Allain, F. H. T. (2007). The testis-specific human protein RBMY recognizes RNA through a novel mode of interaction. *EMBO Reports*, 8(4), 372–379. <https://doi.org/10.1038/sj.embor.7400910>



- Sotriffer, C., & Matter, H. (2011). The Challenge of Affinity Prediction: Scoring Functions for Structure-Based Virtual Screening. In *Virtual Screening: Principles, Challenges, and Practical Guidelines*. <https://doi.org/10.1002/9783527633326.ch7>
- Stefl, R., Oberstrass, F. C., Hood, J. L., Jourdan, M., Zimmermann, M., Skrisovska, L., ... Allain, F. H. T. (2010). The Solution Structure of the ADAR2 dsRBM-RNA Complex Reveals a Sequence-Specific Readout of the Minor Groove. *Cell*, *143*(2), 225–237. <https://doi.org/10.1016/j.cell.2010.09.026>
- Stultz, C. M., & Karplus, M. (2000). *Dynamic Ligand Design and Combinatorial Optimization : Designing Inhibitors to Endothiapepsin*. 289(July 1999), 258–289.
- Su, M., Yang, Q., Du, Y., Feng, G., Liu, Z., Li, Y., & Wang, R. (2019). Comparative Assessment of Scoring Functions: The CASF-2016 Update. *Journal of Chemical Information and Modeling*, *59*(2), 895–913. <https://doi.org/10.1021/acs.jcim.8b00545>
- Sugimoto, Y., König, J., Hussain, S., Zupan, B., Curk, T., Frye, M., & Ule, J. (2012). Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome Biology*, *13*(8), R67. <https://doi.org/10.1186/gb-2012-13-8-r67>
- Sutherland, J. M., Siddall, N. A., Hime, G. R., & McLaughlin, E. A. (2015). RNA binding proteins in spermatogenesis: an in depth focus on the Musashi family. *Asian Journal of Andrology*, *17*(4), 529–536. <https://doi.org/10.4103/1008-682X.151397>
- Teplova, M., Hafner, M., Teplov, D., Essig, K., Tuschl, T., & Patel, D. J. (2013). Structure-function studies of STAR family quaking proteins bound to their in vivo RNA target sites. *Genes and Development*, *27*(8), 928–940. <https://doi.org/10.1101/gad.216531.113>
- Terp, G. E., Johansen, B. N., Christensen, I. T., & Jørgensen, F. S. (2001). A new concept for multidimensional selection of ligand conformations (MultiSelect) and multidimensional scoring (MultiScore) of protein-ligand binding affinities. *Journal of Medicinal Chemistry*, *44*(14), 2333–2343. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/11428927>
- Thomas, P. D., & Dill, K. a. (1996). Structures : How Accurate Are They ? g g. *Journal of Molecular Biology*, *257*, 457–469.
- Treiber, T., Treiber, N., Plessmann, U., Harlander, S., Daiß, J.-L., Eichner, N., ... Meister, G. (2017). A Compendium of RNA-Binding Proteins that Regulate MicroRNA Biogenesis. *Molecular Cell*, *66*(2), 270-284.e13. <https://doi.org/10.1016/j.molcel.2017.03.014>
- Trott, O., & Olson, A. J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, *31*(2), 455–461. <https://doi.org/10.1002/jcc.21334>
- Tsuda, K., Kuwasako, K., Takahashi, M., Someya, T., Inoue, M., Terada, T., ... Yokoyama, S. (2009). Structural basis for the sequence-specific RNA-recognition mechanism of human CUG-BP1 RRM3. *Nucleic Acids Research*, *37*(15), 5151–5166. <https://doi.org/10.1093/nar/gkp546>

- Tuffery, P., Etchebest, C., Hazout, S., & Lavery, R. (1991). A new approach to the rapid determination of protein side chain conformations. *Journal of Biomolecular Structure and Dynamics*, 8(6), 1267–1289. <https://doi.org/10.1080/07391102.1991.10507882>
- Tuszynska, I., Magnus, M., Jonak, K., Dawson, W., Bujnicki, J. M., P., L., ... E, W. (2015). NPDock: a web server for protein–nucleic acid docking. *Nucleic Acids Research*, 43(W1), W425–W430. <https://doi.org/10.1093/nar/gkv493>
- Valverde, R., Edwards, L., & Regan, L. (2008). Structure and function of KH domains. *FEBS Journal*, 275(11), 2712–2726. <https://doi.org/10.1111/j.1742-4658.2008.06411.x>
- van Hoof, A., & Wagner, E. J. (2011). A brief survey of mRNA surveillance. *Trends in Biochemical Sciences*, 36(11), 585–592. <https://doi.org/10.1016/j.tibs.2011.07.005>
- Vanommeslaeghe, K., & MacKerell, A. D. (2012). Automation of the CHARMM general force field (CGenFF) I: Bond perception and atom typing. *Journal of Chemical Information and Modeling*. <https://doi.org/10.1021/ci300363c>
- Velec, H. F. G., Gohlke, H., & Klebe, G. (2005). DrugScore<sup>CSD</sup> Knowledge-Based Scoring Function Derived from Small Molecule Crystal Data with Superior Recognition Rate of Near-Native Ligand Poses and Better Affinity Prediction. *Journal of Medicinal Chemistry*, 48(20), 6296–6303. <https://doi.org/10.1021/jm050436v>
- Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W., & Taylor, R. D. (2003). Improved protein-ligand docking using GOLD. *Proteins: Structure, Function, and Bioinformatics*, 52(4), 609–623. <https://doi.org/10.1002/prot.10465>
- Vieth, M., Hirst, J. D., Kolinski, A., & Brooks, C. L. (1998). Assessing energy functions for flexible docking. *Journal of Computational Chemistry*, 19(14), 1612–1622. [https://doi.org/10.1002/\(SICI\)1096-987X\(19981115\)19:14<1612::AID-JCC7>3.0.CO;2-M](https://doi.org/10.1002/(SICI)1096-987X(19981115)19:14<1612::AID-JCC7>3.0.CO;2-M)
- Walia, R. R., Caragea, C., Lewis, B. A., Towfic, F., Terribilini, M., El-Manzalawy, Y., ... Honavar, V. (2012). Protein-RNA interface residue prediction using machine learning: an assessment of the state of the art. *BMC Bioinformatics*, 13(1), 89. <https://doi.org/10.1186/1471-2105-13-89>
- Walia, R. R., EL-Manzalawy, Y., Honavar, V. G., & Dobbs, D. (2017). *Sequence-Based Prediction of RNA-Binding Residues in Proteins*. [https://doi.org/10.1007/978-1-4939-6406-2\\_15](https://doi.org/10.1007/978-1-4939-6406-2_15)
- Wang, C., & Zhang, Y. (2017). Improving scoring-docking-screening powers of protein-ligand scoring functions using random forest. *Journal of Computational Chemistry*, 38(3), 169–177. <https://doi.org/10.1002/jcc.24667>
- Wang, E. T., Sandberg, R., Luo, S., Khrebtkova, I., Zhang, L., Mayr, C., ... Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221), 470–476. <https://doi.org/10.1038/nature07509>
- Wang, H., Zeng, F., Liu, Q., Liu, H., Liu, Z., Niu, L., ... Li, X. (2013). The structure of the ARE-binding domains of Hu antigen R (HuR) undergoes conformational changes during RNA

binding. *Acta Crystallographica Section D Biological Crystallography*, 69(3), 373–380.  
<https://doi.org/10.1107/S0907444912047828>

- Wang, R., Fang, X., Lu, Y., & Wang, S. (2004). The PDBbind Database: Collection of Binding Affinities for Protein–Ligand Complexes with Known Three-Dimensional Structures. *Journal of Medicinal Chemistry*, 47(12), 2977–2980. <https://doi.org/10.1021/jm030580l>
- Wang, R., Lai, L., & Wang, S. (2002). Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *Journal of Computer-Aided Molecular Design*, 16(1), 11–26. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/12197663>
- Weiner, S. J., Kollman, P. A., Nguyen, D. T., & Case, D. A. (1986). An all atom force field for simulations of proteins and nucleic acids. *Journal of Computational Chemistry*. <https://doi.org/10.1002/jcc.540070216>
- Welch, W., Ruppert, J., & Jain, A. N. (1996). Hammerhead: Fast, fully automated docking of flexible ligands to protein binding sites. *Chemistry and Biology*, 3(6), 449–462. [https://doi.org/10.1016/S1074-5521\(96\)90093-9](https://doi.org/10.1016/S1074-5521(96)90093-9)
- Wheeler, E. C., Van Nostrand, E. L., & Yeo, G. W. (2018). Advances and challenges in the detection of transcriptome-wide protein–RNA interactions. *Wiley Interdisciplinary Reviews: RNA*. <https://doi.org/10.1002/wrna.1436>
- Wójcikowski, M., Ballester, P. J., & Siedlecki, P. (2017). Performance of machine-learning scoring functions in structure-based virtual screening. *Scientific Reports*, 7(1), 46710. <https://doi.org/10.1038/srep46710>
- Xu, Y., Vanommeslaeghe, K., Aleksandrov, A., MacKerell, A. D., Nilsson, L., & Nilsson, L. (2016). Additive CHARMM force field for naturally occurring modified ribonucleotides. *Journal of Computational Chemistry*, 37(10), 896–912. <https://doi.org/10.1002/jcc.24307>
- Yang, L., Wang, C., Li, F., Zhang, J., Nayab, A., Wu, J., ... Gong, Q. (2017). The human RNA-binding protein and E3 ligase MEX-3C binds the MEX-3-recognition element (MRE) motif with high affinity. *The Journal of Biological Chemistry*, 292(39), 16221–16234. <https://doi.org/10.1074/jbc.M117.797746>
- Yu, J., Ciancetta, A., Dudas, S., Duca, S., Lottermoser, J., & Jacobson, K. A. (2018). Structure-Guided Modification of Heterocyclic Antagonists of the P2Y<sub>14</sub> Receptor. *Journal of Medicinal Chemistry*, 61(11), 4860–4882. <https://doi.org/10.1021/acs.jmedchem.8b00168>
- Zhan Deng, Claudio Chuaqui, and, & Singh\*, J. (2003). *Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein–Ligand Binding Interactions*. <https://doi.org/10.1021/JM030331X>
- Zhang, Y., & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 33(7), 2302–2309. <https://doi.org/10.1093/nar/gki524>

- Zhao, H., Yang, Y., Janga, S. C., Kao, C. C., & Zhou, Y. (2014). Prediction and validation of the unexplored RNA-binding protein atlas of the human proteome. *Proteins*, 82(4), 640. <https://doi.org/10.1002/PROT.24441>
- Zheng, C., Zhou, Y., Zhu, J., Ji, H., Chen, J., & Li, Y. (2007). Construction of a three-dimensional pharmacophore for Bcl-2 inhibitors by flexible docking and the multiple copy simultaneous search method. *Bioorganic & Medicinal Chemistry*, 15, 6407–6417. <https://doi.org/10.1016/j.bmc.2007.06.052>
- Zhu, Y., Xu, G., Yang, Y. T., Xu, Z., Chen, X., Shi, B., ... Wang, P. (2019). POSTAR2: deciphering the post-transcriptional regulatory logics. *Nucleic Acids Research*, 47(D1), D203–D211. <https://doi.org/10.1093/nar/gky830>
- Zilian, D., & Sotriffer, C. A. (2013). SFCscore<sup>RF</sup> : A Random Forest-Based Scoring Function for Improved Affinity Prediction of Protein–Ligand Complexes. *Journal of Chemical Information and Modeling*, 53(8), 1923–1933. <https://doi.org/10.1021/ci400120b>
- Zipeto, M. A., Jiang, Q., Melese, E., & Jamieson, C. H. M. (2015). RNA rewriting, recoding, and rewiring in human disease. *Trends in Molecular Medicine*, 21(9), 549–559. <https://doi.org/10.1016/J.MOLMED.2015.07.001>
- Zoete, V., Grosdidier, A., & Michielin, O. (2009). Docking, virtual high throughput screening and in silico fragment-based drug design. *Journal of Cellular and Molecular Medicine*, 13(2), 238–248. <https://doi.org/10.1111/j.1582-4934.2008.00665.x>

**Titre :** Développement et application d'une approche de docking par fragments pour modéliser les interactions entre protéines et ARN simple-brin

**Mots clés :** ARN simple-brin - protéine - interaction - amarrage - sélectivité

**Résumé :** Les interactions ARN-protéine interviennent dans de nombreux processus cellulaires fondamentaux. L'obtention de détails à l'échelle atomique de ces interactions nous éclaire sur leurs fonctions, mais permet également d'envisager la conception rationnelle de ligands pouvant les moduler. Lorsque les deux techniques majeures que sont la RMN et la cristallographie aux rayons X ne permettent pas d'obtenir une structure 3D entre les deux partenaires, des approches de docking peuvent être utilisées pour apporter des modèles. L'application de ces approches aux complexes ARN-protéine se heurtent cependant à une difficulté. Ces complexes résultent en effet souvent de la liaison spécifique d'une courte séquence d'ARN simple-brin (ARNsb) à sa protéine cible.

Hors, la flexibilité inhérente aux segments simples-brins impose dans une approche classique de docking d'explorer un large ensemble de leur espace conformationnel. L'objectif du projet est de contourner cette difficulté par le développement d'une approche de docking dite "par fragments". Ce dernier s'est fait à partir de domaines de liaison à l'ARN très représentés dans le monde du vivant. Les résultats ont montré une excellente capacité prédictive de l'approche à partir de la séquence de l'ARN. Ils ont de plus montré un potentiel intéressant dans la prédiction de séquences d'ARN simple-brin préférentiellement reconnues par des domaines de liaisons à l'ARN.

**Title :** Development and application of a fragment-based docking approach to model protein-ssRNA interactions

**Keywords :** single-stranded RNA - protein - interaction - docking - selectivity

**Abstract :** RNA-protein interactions mediate numerous fundamental cellular processes. Atomic scale details of these interactions shed light on their functions but can also allow the rational design of ligands that could modulate them. NMR and X-ray crystallography are the 2 main techniques used to resolve 3D high-resolution structures between two interacting molecules. Docking approaches can also be utilized to give models as an alternative. However, the application of these approaches to RNA-protein complexes is hampered by an issue. RNA-protein interactions often relies on the specific recognition of a short single-stranded RNA (ssRNA) sequence by the protein.

The inherent flexibility of the ssRNA segment would impose, in a classical docking approach, to explore their resulting large conformation space which is not computationally reliable. The goal of this project is to overcome this barrier by using a fragment-based docking approach. This approach developed from some of the most represented RNA-binding domains showed excellent results in the prediction of the ssRNA-protein binding mode from the RNA sequence and also a great potential to predict preferential RNA binding sequences.

